Typical Sequences for Polish Alphabets

Patrick Mitran, Member, IEEE,

Abstract

The notion of typical sequences plays a key role in the theory of information. Central to the idea of typicality is that a sequence $x_1, x_2, \ldots, x_n$ that is $P_X$-typical should, loosely speaking, have an empirical distribution that is in some sense close to the distribution $P_X$. The two most common notions of typicality are that of strong (letter) typicality and weak (entropy) typicality. While weak typicality allows one to apply many arguments that can be made with strongly typical arguments, some arguments for strong typicality cannot be generalized to weak typicality.

In this paper, we consider an alternate definition of typicality, namely one based on the weak* topology and that is applicable to Polish alphabets (which includes $\mathbb{R}^n$). This notion is a generalization of strong typicality in the sense that it degenerates to strong typicality in the finite alphabet case, and can also be applied to mixed and continuous distributions. Furthermore, it is strong enough to prove a Markov lemma, and thus can be used to directly prove a more general class of results than weak typicality. As an example of this technique, we directly prove achievability for Gel’fand-Pinsker channels with input constraints for a large class of alphabets and channels without first proving a finite alphabet result and then resorting to delicate quantization arguments.

While this large class does not include Gaussian distributions with power constraints, it is shown to be straightforward to recover this case by considering a sequence of truncated Gaussian distributions.

Index Terms

Typical sequences, weak* topology, capacity, Gel’fand-Pinsker.

I. INTRODUCTION

Perhaps the most intuitive method of deriving achievable rates in information theory for stationnary memoryless problems is with the concept of typical sequences. Roughly speaking, a sequence is typical if its empirical distribution is close, in some sense, to some ideal distribution.

The author is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario N2L 3G1 (email: pmitran@ecemail.uwaterloo.ca).
Ignoring minor variations in definitions, there are essentially two broad notions of typical sequences. The first notion, called weakly typical sequences, measures the closeness of a sequence \( x = (x_1, \ldots, x_n) \) to a distribution \( P_X \) by quantifying the probability of the sequence \( x \). Specifically, the length \( n \) sequence \( x \) is \((P_X, \epsilon)\)-weakly typical if
\[
\left| \frac{1}{n} \sum_{\ell} \log P_X(x_\ell) - H(X) \right| < \epsilon,
\]
where \( P_X \) is a probability mass function (pmf) and \( H(X) \) the entropy of \( X \) if \( X \) is discrete while \( P_X \) is a probability density function (pdf) and \( H(X) \) the differential entropy of \( X \) if \( X \) is continuous. For this reason, weakly typical sequences are often referred to as entropy typical. The notion of weak typicality does not appear to generalize well to mixed distributions.

By contrast, strong typicality characterizes a sequence \( x \) by the relative frequency of the occurrence of each letter of the alphabet of \( X \). Specifically, if \( N(a|x) \) denotes the number of occurrences of the letter \( a \) in the length \( n \) sequence \( x \), then \( x \) is \((P_X, \epsilon)\)-strongly typical if
\[
|N(a|x)/n - P_X(a)| < \epsilon,
\]
for all \( a \). Strong typicality is sometimes referred to by the more descriptive name of letter typicality. Evidently, strong typicality implies at most a countable alphabet.

Strong typicality has at least two key consistency properties not shared with weak typicality. First, strong typicality is sufficient for proving a Markov lemma, which is a key technique in many network information theory proofs. The Markov lemma is essentially a corollary of the following broad statement, which one would intuitively expect to be true for any reasonable definition of typicality: *if a typical sequence, generated in some arbitrary fashion, is input to a stationnary memoryless channel, then the input and output sequences should be jointly typical in some sense.* Unfortunately, this statement is not possible in general with weak typicality. We call this desirable property the channel consistency property.

A second desirable property of a typical sequence involves cost functions. Specifically, let \( g : \mathcal{X} \to \mathbb{R} \) be a mapping from the alphabet of \( X \) to the reals. If the length \( n \) sequence \( x \) is \( P_X \)-typical, one would expect that the weighted sum \( \frac{1}{n} \sum_{\ell} g(x_\ell) \) would be reasonably approximated by \( E_{P_X}[g(X)] \). While such a statement can be formalized for strong typicality, entropy typicality by itself is not sufficient to imply this property. We call this desirable property the cost consistency property.

1In some variations, the additional condition that \( N(a|x) = 0 \) if \( P_X(a) = 0 \) is imposed.
2It should be noted that for continuous/discrete distributions, a variation of so-called distortion typical sequences can resolve this issue for a specific cost function \( g(x) \).
Finally, typical sequences should satisfy certain large deviations results. For example, if $P_{XY}$ is a joint distribution with marginals $P_X$ and $P_Y$, $y$ is $P_Y$-typical and $X$ is generated independently of $y$ with each letter i.i.d. according to $P_X$, then the probability of $X$ and $y$ being jointly $P_{XY}$-typical should be on the order of $\approx 2^{-nI(X;Y)}$. This result is usually shown for strong typical sequences.

The contribution of this paper is to introduce a notion of typicality based on any metric that induces the weak* convergence of probability measures. We call this notion weak* typicality. The notion is sufficiently general to apply to any distribution (discrete, continuous or mixed) where the alphabet is a Polish space, and reduces to strong typicality for finite alphabets. This includes, for example, mixed distributions in $\mathbb{R}^n$.

We show that this notion of typicality allows for both of the above consistency properties in addition to the usual rules that one expects a typical sequence to follow, e.g., if a pair of sequences $(x, y)$ are jointly typical, one expects that each of $x$ and $y$ are typical. As an example of this weak* typicality, we directly prove an achievability result for Gel’fand-Pinsker channels with alphabets in Polish spaces and cost constraint at the transmitter without having to resort to delicate quantization arguments which are typically handwaved. Indeed, a key contribution of this work is that by employing the notion of weak* typicality, one does not need to directly invoke quantization arguments.

Two important remarks are in order. While, the notion of weak* typicality avoids the technical difficulties of first proving a result in the discrete case and then employing delicate quantization arguments, some of the large deviation results for weak* typical sequences are proved by using quantization arguments. However, these arguments are not necessary for the application of weak* typical sequences.

Second, for the cost consistency property to apply, the cost function must be bounded (and continuous). This initially precludes weak* typical sequences from being directly applicable to Gaussian input distributions with power constraints. However, the result in the Gaussian case can be recovered by considering a sequence of truncated Gaussians.

The techniques proved here do not result in more general expressions for channel capacity. The most general expression for channel capacity is given by information spectrum methods [11], [19] which apply not only to channels with memory but even non-ergodic channels. The information spectrum approach though, looks at quantities such as

$$I(X;Y) := \varliminf_{n \to \infty} \frac{1}{n} \log \frac{W^n(Y^n|X^n)}{P_{Y^n}(Y^n)}. \quad (3)$$

This characterization is based on ratios of probabilities and, similar to weak typicality, this characterization does not appear to allow for a Markov lemma. Nevertheless, the results presented here allow one
to generalize many results derived using strong typical sequences to a larger class of alphabets in a straightforward manner.

Weak* convergence of probability measures has found some uses in the information theory literature. Perhaps the most relevant application to this work was by Csiszár to compute the capacity of arbitrary varying channels with general alphabets and states [5]. In that work, channels must satisfy a weak* continuity property and we adopt the same requirement for channels here. In [5], the capacity is first computed for the special case of finite input alphabets and this result is then used to derive the general case. It should be noted that while weak* convergence of measures plays a key role in [5], no new notion of typical sequences is introduced there.

A key difficulty with continuous alphabets is the analytical characterization of mutual information. In [17] it was shown that channel capacity is a lower semi-continuous function in the weak* topology. In [18], sufficient conditions are found which ensure that channel capacity can be approached by discrete input distributions or uniform input distribution with finite support for general alphabet channels. In [8], necessary and sufficient conditions for weak* continuity and strict concavity of mutual information were found in addition to conditions that characterize the capacity value and capacity achieving measure for channels with side-information at the receiver. In [12] and [13], Keiffer proves coding theorems for stationnary (but not necessarily memorryless) channels that are weak* continuous. While a notion of typicality appears in [12] and [13], it is a variant of strong typicality and defined only on discrete alphabets.

Weak* convergence has also found applications in analyzing the stability of recursive algorithms such as variants of the LMS adaptive filtering algorithm [2].

The rest of this paper is structured as follows. In Section II we define channel inputs, input constraints, channels and information measures such as divergence, as well as some measure theoretic considerations. In Section IV, we define weak* typicality and prove our key results. In Section V, we provide two examples of weak* typical sequences by proving achievability results for point-to-point and Gel’fand-Pinsker channels. In Section VI we discuss the Gaussian case. In Section VII we conclude this work. The Appendix contains two technical proofs for weak* typical sequences.

II. Preliminaries

A. Alphabets and weak* convergence

In this paper all measureable spaces are Polish (i.e., complete separable metric spaces) and endowed with the Borel $\sigma$-field generated by the open sets. We denote a measureable space by $E$ and the $\sigma$-field
by \( \mathcal{E} \). \( \mathcal{M}_1(E) \) denotes the set of probability measures on \( E \). When we need to distinguish two spaces, we employ subscripts such as \( (E_X, \mathcal{E}_X) \) and \( (E_Y, \mathcal{E}_Y) \). Unless clear from context, a random variable \( X \) will take values in the alphabet \( E_X \) and have distribution \( \mathcal{L}(X) \) which is usually denoted by \( P_X \).

\( P(A) \) and \( E[X] \) denote the probability of event \( A \) and mean of random variable \( X \) where the underlying measure is always clear from context, or explicitly stated by subscript.

If \( E_X \) and \( E_Y \) are Polish, then so is \( E_X \times E_Y \). The corresponding \( \sigma \)-field \( \mathcal{E}_{XY} := \mathcal{E}_X \otimes \mathcal{E}_Y \) is the smallest \( \sigma \)-field containing all rectangles \( A \times B, A \in \mathcal{E}_X, B \in \mathcal{E}_Y \).

A sequence \( (x_1, \ldots, x_n) \in E^n_X \) will be denoted by \( x^n \). When the length is clear from context or not relevant, we simply write \( x \). An i.i.d. random sequence \( \mathbf{X} = (X_1, \ldots, X_n) \) consists of a sequence of independent random variables \( X_1, X_2, \ldots, X_n \), each taking values in \( E_X \) and for which the laws \( \mathcal{L}(X_i) = P_X \) are identical.

In this paper, the notion of weak* convergence of probability measures plays a key role. We denote the weak* convergence of a sequence of measures \( P_n \) to a limiting measure \( P \) by \( P = \text{w-lim}_{n \to \infty} P_n \). The Portemanteau theorem [15, Theorem 13.16] provides the following equivalent conditions which will be used in the sequel.

**Theorem 1:** Let \( E \) be a Polish space and \( P, P_1, P_2, \ldots \in \mathcal{M}_1(E) \). Then the following are equivalent

1) \( P = \text{w-lim}_{n \to \infty} P_n \).

2) \( \lim_{n \to \infty} \int f \, dP_n = \int f \, dP \) for all bounded continuous \( f \).

3) \( \lim_{n \to \infty} P_n(A) = P(A) \) for all \( A \in \mathcal{E} \) with \( P(\partial A) = 0 \), where \( \partial A \) denotes the boundary of \( A \).

We note that condition 2 is usually taken to be the definition of weak* convergence.

**B. Channels and Channel Inputs**

A memoryless channel from an input \( X \) to an output \( Y \) is described by a transition kernel \( W_{Y|X}(B|x) \) for \( x \in E_X, B \in \mathcal{E}_Y \) which must satisfy the usual measurability conditions of a kernel. Furthermore, as in [5], we make the following additional continuity assumption on \( W_{Y|X}(\cdot|x) \):

**Definition 2:** A transition kernel \( W_{Y|X}(\cdot|x) \) is said to be a channel if \( W_{Y|X}(\cdot|x) \) depends continuously (in the weak* sense) on \( x \), i.e.,

\[
W_{Y|X}(\cdot|x) = \text{w-lim}_{n \to \infty} W_{Y|X}(\cdot|x_n),
\]

whenever \( x = \lim_n x_n \).
Given a measure $P_X$ and a kernel $W_{Y|X}$ we denote by $P_X \otimes W_{Y|X}$ and $P_X W_{Y|X}$ the measures on $E_X \times E_Y$ and $E_Y$ uniquely defined by

\[
(P_X \otimes W_{Y|X})(A \times B) := \int_A W_{Y|X}(B|x)P_X(dx) \tag{5}
\]

\[
(P_X W_{Y|X})(B) := \int_{E_X} W_{Y|X}(B|x)P_X(dx), \tag{6}
\]

for Borel sets $A \in \mathcal{E}_X$ and $B \in \mathcal{E}_Y$. When clear from context, we will denote the marginal $P_X W_{Y|X}$ of $P_X \otimes W_{Y|X}$ by $P_Y$.

In practice, sequences are input to communication channels. In this paper, all sequences belong to product spaces, i.e., an input sequence $x = (x_1, \ldots, x_n)$ belongs to $\times_{i=1}^n E_X := E_X^n$ while an output sequence, say $Y^n$, belongs to $\times_{i=1}^n E_Y := E_Y^n$.

In general, an input sequence $x$ results in a random output sequence $Y$ described by a transition kernel $W_{Y|X}(B^n|x)$ where $B^n \in \mathcal{E}_Y^n$. In this paper, all channels are stationary and memoryless. Thus if a sequence $x = (x_1, \ldots, x_n)$ is input into a kernel $W_{Y|X}$, then the probability that the output $Y$ lies in a product set $\times_{\ell=1}^n B_\ell, B_\ell \in \mathcal{E}_Y$, is $\prod_{\ell=1}^n W_{Y|X}(B_\ell|x_\ell)$, where $W_{Y|X}(\cdot|x)$ is assumed to satisfy the constraint of Definition 2.

It is common for channel inputs to be required to satisfy some constraints. For example, the Gaussian channel typically has a power constraint

\[
\frac{1}{n} \sum_{\ell=1}^n |x_\ell|^2 < P. \tag{7}
\]

We now establish the equivalent concept in this paper.

**Definition 3:** We say that a function $g(x)$ and a threshold $\Gamma$ form an input constraint provided $g(\cdot)$ is continuous and bounded. We say that the input vector $x = (x_1, \ldots, x_n)$ satisfies the input constraint provided

\[
\frac{1}{n} \sum_{\ell=1}^n g(x_\ell) < \Gamma, \tag{8}
\]

and with abuse of notation, we define $g(x) := \frac{1}{n} \sum_{\ell=1}^n g(x_\ell)$.

**Remark 4:** A bounded constraint or cost $g(x)$ is sometimes assumed in the literature [16]. If the input alphabet $E_X$ is compact, then any continuous $g(x)$ is always bounded. While the bounded assumption on $g(x)$ initially precludes a power constraint of the form (7) when the alphabet is $\mathbb{R}$, we will see that the classic result in the additive Gaussian noise case with a power constraint can be recovered by considering a sequence (in $L$) of compact input alphabet $E_{X_L}$. Finally, if the sequence of empirical input distributions
on $X$ converges weakly* to $P_X$, then the continuity assumption can be relaxed to $P_X(D_g) = 0$ where $D_g$ is the set of discontinuities of $g(\cdot)$ (see part (iii) of [15, Theorem 13.16]).

C. Two Examples

We now give two common examples of alphabets that satisfy the above constraints.

First, consider two random variables $X$ and $Y$ whose alphabets $E_X$ and $E_Y$ are finite. Then these trivially satisfy the assumptions of Section II-A if we choose as metric the trivial metric $d_X(x_1, x_2) = 1$ if $x_1 \neq x_2$ and 0 otherwise, and likewise for $d_Y(\cdot, \cdot)$. Furthermore, in this case, any channel from $X$ to $Y$ satisfies the weak* continuity assumption of Section II-B since if a sequence $x_n \to x$, then in fact $x_n = x$ for all $n$ greater than some $N$, i.e. $W(\cdot|x_n) = W(\cdot|x)$ for all $n > N$.

As a second example, we consider two random variables $X$ and $Y$ whose alphabets are $E_X = E_Y = \mathbb{R}^N$. With the usual metric on the real line, these are Polish spaces. Furthermore, suppose that $Y = X + Z$, where $Z$ is independent of $X$ and has a density $f(z)$, i.e.,

$$W_{Y|X}(B|x) = \int_B f(y - x) \, dy. \tag{9}$$

Then as shown in [17, Lemma 2], the channel $W_{Y|X}$ satisfies the weak* continuity assumption of Section II-B.

III. INFORMATION MEASURES

A. Definitions

In this section, we provide some basic background on information measures and introduce a key result.

Let $P$ and $M$ be two probability measures defined on $E$ and let $Q = \{Q_1, Q_2, \ldots, Q_{|Q|}\}$ be a finite (measurable) partition of $E$. Then, we define (see [10, Section 2.3])

$$H_P||M(Q) := \sum_{i=1}^{\frac{|Q|}{P(Q_i)}} P(Q_i) \log \frac{P(Q_i)}{M(Q_i)}. \tag{10}$$

The divergence of $P$ with respect to $M$ is defined as (see [10, Section 5.2])

$$D(P||M) = \sup_Q H_P||M(Q), \tag{11}$$

where the supremum is over all finite measurable partitions $Q$ of $E$.

Recall that a field $\mathcal{F}$ has the properties that i) $E \in \mathcal{F}$, ii) if $F \in \mathcal{F}$ then $F^c \in \mathcal{F}$ and iii) $\mathcal{F}$ is closed under finite unions. In the next subsection, we will construct a field that, in addition to generating the
σ-field, i.e., $\mathcal{E} = \sigma(\mathcal{F})$, has a desirable property. Fields that generate the σ-field are of particular interest due to the following two lemmas.

**Lemma 5:** [10, Lemma 5.2.2] Let $(E, \mathcal{E})$ be a measurable space, $\mathcal{F}$ a field that generates $\mathcal{E}$ and $P$ and $M$ two measures defined on this space. Then

$$D(P||M) = \sup_{Q \subset \mathcal{F}} H_P||M(Q).$$

The above lemma states that it is sufficient to restrict the finite partitions to subsets of a generating field.

Suppose $(E_X, \mathcal{E}_X)$ and $(E_Y, \mathcal{E}_Y)$ are measure spaces, generated by the fields $\mathcal{F}_X$ and $\mathcal{F}_Y$ respectively. Then the product σ-field $\mathcal{E}_{XY}$ is generated by the field of rectangles $\mathcal{F}_{XY} = \mathcal{F}_X \times \mathcal{F}_Y$. Thus, we have the following result (see [10, Lemma 5.5.1]).

**Lemma 6:** Let $P_{XY}$ be a measure on $E_{XY}$ and $P_X$ and $P_Y$ the respective marginals on $E_X$ and $E_Y$. Then

$$I(X;Y) = D(P_{XY}||P_X \times P_Y) = \sup_{Q_X \subset \mathcal{F}_X, Q_Y \subset \mathcal{F}_Y} H_{P_{XY}}||P_X \times P_Y(Q_X \times Q_Y).$$

**B. A Special Field**

In this subsection, given a measure $P$, we will construct a generating field $\mathcal{F}_P$ with the key property that the $P$-measure of the boundary of any set in the field is zero. This last property is desirable as it will ensure that if a sequence of measures $P_n$ converges weakly* to $P$, then $P_n(A) \to P(A)$ for all sets $A \in \mathcal{F}_P$.

While there is no lack of standard constructions for fields that generate the σ-field, given any such field $\mathcal{F}$, there is no guarantee that $P(\partial A) = 0$ for all $A \in \mathcal{F}$ and thus, it is necessary to construct such a generating field $\mathcal{F}_P$ specifically for each limiting measure $P$.

**Lemma 7:** Let $P$ by a probability measure defined on a Polish measure space $(E, \mathcal{E})$. Then there exists a countable family $\mathcal{A} \subset \mathcal{E}$ of open sets that i) generates the Borel σ-field $\mathcal{E}$, i.e., $\sigma(\mathcal{A}) = \mathcal{E}$, and ii) $P(\partial A) = 0$ for all $A \in \mathcal{A}$.

**Proof:** Since $E$ is Polish, there is a countably dense subset of $E$, say $E'$. For $\mathcal{A}$ to generate $\mathcal{E}$, it is sufficient that for each $x \in E'$ there is a countably dense subset $R_x$ of $\mathbb{R}^+$ with each ball $B(x, r) \in \mathcal{A}$ for all $r \in R_x$. This is because then each open set of $E$ is the countable union of balls in $\mathcal{A}$.

It thus remains to be shown that the sets $R_x$ can be chosen such that each ball $B(x, r)$ has $P(\partial B(x, r)) = 0$. For $r > 0$, let $F_x(r) = P(B(x, r))$. Then $F_x(r)$ is a non-decreasing, bounded below by 0 and above $^{3}\sigma(\mathcal{F})$ is the smallest σ-field that contains the family $\mathcal{F}$ of sets.
by 1. Thus it has left and right limits and at most a countable number of jump discontinuities and thus,
there is a countably dense subset $R'_x$ of $\mathbb{R}^+$ for which $F_x(r)$ is continuous. We claim that choosing
$R_x = R'_x$ will do. Since $\partial B(x, r) \subset \{y : d(x, y) = r\}$, then $P(\partial B(x, r)) \leq F_x(r+) - F_x(r)$
where $F_x(r+)$ is the right limit of $F_x(r)$. However, for $r \in R'_x$, $F_x(r+) = F_x(r)$ by continuity.

Corollary 8: Let $P$ be a probability measure defined on a Polish measure space $(E, \mathcal{E})$. Then there
exists a countable generating field $\mathcal{F}_P$ with the property that $P(\partial A) = 0$ for all $A \in \mathcal{F}_P$.

Proof: Since $\partial (A \cap B) \subset \partial A \cup \partial B$, $\partial (A \cup B) \subset \partial A \cup \partial B$ and $\partial A = \partial (A^c)$, one can extend the
countable family $\mathcal{A}$ in Lemma 7 to include all finite intersections/unions of balls in $\mathcal{A}$ and complements
of balls in $\mathcal{A}$, i.e., extend $\mathcal{A}$ to a field $\mathcal{F}_P$.

IV. WEAK* TYPICAL SEQUENCES

A. Definitions

Given a sequence $x = (x_1, \ldots, x_n) \in E^m_X$, one can associate an empirical distribution $P_x$ defined by

$$P_x(A) := \frac{1}{n} \sum_{\ell=1}^{n} 1_{\{x_{\ell} \in A\}}. \quad (14)$$

When clear from context, we denote the empirical distribution by $P^n_X$. Likewise, given two sequences
$x = (x_1, \ldots, x_n) \in E^n_X$ and $y = (y_1, \ldots, y_n) \in E^n_Y$, the joint empirical distribution $P_{x,y}$ is defined by

$$P_{x,y}(A \times B) := \frac{1}{n} \sum_{\ell=1}^{n} 1_{\{x_{\ell} \in A\}} 1_{\{y_{\ell} \in B\}}, \quad (15)$$

which is denoted by $P^n_{X,Y}$ when clear from context.

A sequence $x$ should be typical if its empirical distribution is in some sense close to some probability
measure. In this paper, closeness is measured with respect to the weak* topology.

Specifically, let $d(\cdot, \cdot)$ be any metric on the space of probability measures $\mathcal{M}_1(E)$ that induces the
weak* topology and fix this metric for the rest of the paper. The Prohorov metric is an example of such
a metric and the exact choice of the metric is irrelevant.

We denote by $B(M, \epsilon)$ the ball of distributions $\{P \in \mathcal{M}_1(E) : d(P, M) < \epsilon\}$. We will say that
an empirical distribution $P^n$ is $P$-typical when its distance from $P$ is sufficiently small. We make the
following definitions.

Definition 9: Let $d(\cdot, \cdot)$ be a metric on the space of probability measures $\mathcal{M}_1(E_X)$ that induces the
weak* topology. A sequence $x = (x_1, \ldots, x_n)$ is said to be weakly* $(P_X, \epsilon)$-typical if $d(P_X, P_X) < \epsilon$.

Similarly, we say that an empirical distribution $P^n_X$ is weakly* $(P_X, \epsilon)$-typical if $d(P^n_X, P_X) < \epsilon$. We
denote the set of such length $n$ typical sequences by $A^n(\epsilon)$. 

DRAFT
Remark 10: For finite alphabets, it is interesting to compare the definition of weak* typicality to that of strong typicality. Specifically, if $|E_X|$ is finite and $P_x \in A^n_\epsilon(P_X)$, then $|P_x(a) - P_X(a)| < \delta$ for some $\delta > 0$ and all $a \in E_X$, and $\delta \to 0$ as $\epsilon \to 0$. This coincides with the definition of strong typicality (except for the occasional requirement that $P_x(a) = 0$ if $P_X(a) = 0$). Thus weak* typical sequences can be viewed as a generalization of strong typical sequences.

Unless stated otherwise, in the sequel all typical sequences are weak* typical sequences.

We also find it convenient to introduce a notion of asymptotically typical sequences. Given a set of sequences $x^{n_1}, x^{n_2}, \ldots$ of lengths $n_1 < n_2 < \ldots$, there is a corresponding sequence of empirical distributions $P^n_{X_1}, P^n_{X_2}, \ldots$.

Definition 11: We say that the sequence of sequences $\{x^{n_k}\}$ is asymptotically $P_X$-typical provided that the corresponding sequence of empirical measures satisfies

$$P_X = \lim_{k \to \infty} P^n_{X_k}. \quad (16)$$

Remark 12: If a sequence of sequences $\{x^{n_k}\}$ is asymptotically $P_X$-typical, then for any $\epsilon > 0$ there exists a $K$ such that for all $k > K$, $x^{n_k}$ is weak* $(P_X, \epsilon)$-typical.

It some cases it will be more convenient to first prove certain results for asymptotically typical sequences of sequences, and then as a corollary infer behavior of typical sequences for large length $n$.

Jointly weak* typical sequences are defined analogously. Specifically,

Definition 13: Two sequences $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ are said to be weak* $(P_{XY}, \epsilon)$-typical if $d(P_{X,Y}, P_{XY}) < \epsilon$. Similarly, we say that the empirical distribution $P^n_{X(Y)}$ is weak* $(P_{XY}, \epsilon)$-typical if $d(P^n_{X(Y)}, P_{XY}) < \epsilon$. We denote the set of such pairs of length $n$ weak* typical sequences by $A^n_{\epsilon}(P_{XY})$.

Likewise, one can also define a sequence of a pair of sequences $\{(x^{n_k}, y^{n_k})\}$ to be asymptotically $P_{XY}$-typical in the obvious way.

B. Consistency Properties

There are several desirable properties that typical and jointly typical sequences should possess.

First, a random i.i.d. sequence should be typical with high probability. Second, a $(P_X, \epsilon)$-typical sequence $x$ should have a cost $\frac{1}{n} \sum_i g(x_i)$ close to $E_{P_X}[g(X)]$. Third, if two sequences are jointly typical, then one would expect each sequence to be typical in its own right.

The following lemma shows that the first is indeed true for asymptotically typical sequences.
Lemma 14: Let $X_1, X_2, \ldots$ be a sequence of independent random variables with values in $E_X$ with identical distribution $P_X$ and $\{X^{n_k}\}$ a corresponding sequence of sequences. Then almost surely $\{X^{n_k}\}$ is asymptotically $P_X$-typical.

Proof: This is a direct restatement of Varadarajan’s Theorem [7, Theorem 11.4.1].

We now show that all three statements are true for weak* typical sequences, where the first is a consequence of the result for asymptotically typical sequences.

Theorem 15: The following hold.

1) Let $X_1, X_2, \ldots, X_n$ be independent random variables with values in $E_X$ with identical distribution $P_X$. Then for any $\epsilon > 0$,

$$\lim_{n \to \infty} P(X^n \in A^n_\epsilon(P_X)) = 1.$$  (17)

2) For every $\delta > 0$, there is an $\bar{\epsilon}(\delta) > 0$ such that if $x \in A^{n}_{\epsilon(\delta)}(P_X)$ then

$$|E_{P_X}[g(X)] - E_{P_X}[g(X)]| < \delta.$$  (18)

3) For any $\epsilon > 0$, there is an $\bar{\epsilon}(\epsilon) > 0$ such that if $(x, y) \in A^{n}_{\epsilon(\delta)}(P_{XY})$ then $x \in A^{n}_{\epsilon}(P_X)$.

Proof: i) Otherwise one could find a sequence of sequences $\{X^{n_k}\}$ in Lemma 14 that would not be asymptotically typical almost surely.

ii) Let $M_X$ be any (not necessarily empirical) measure on $E_X$. By part 2 of Theorem 1, there is an $\bar{\epsilon}(\delta)$ such that $M_X \in B(P_X, \bar{\epsilon}(\delta))$ implies

$$|E_{M_X}[g(X)] - E_{P_X}[g(X)]| < \delta.$$  (19)

The result follows since $x \in A^{n}_{\epsilon(\delta)}(P_X)$ iff $P_X \in B(P_X, \bar{\epsilon}(\delta))$.

iii) By part 3 of Theorem 1 if $M_X^k$ is a sequence of (not necessarily empirical) measures in $E_{XY}$ with marginals $M_X^k$ and such that $\lim_k d(M_{XY}^k, P_{XY}) = 0$, then $\lim_k d(M_X^k, P_X) = 0$. Thus for each $\epsilon$, there is an $\bar{\epsilon}(\epsilon)$ such that $d(M_{XY}, P_{XY}) < \bar{\epsilon}(\epsilon)$ implies $d(M_X, P_X) < \epsilon$. The result again follows since $(x, y) \in A^{n}_{\epsilon(\delta)}(P_{XY})$ iff $P_{X,Y} \in B(P_{XY}, \bar{\epsilon}(\epsilon))$.

An important desirable property of typical sequences is that if a typical sequence is input to a channel then the input and output should be jointly typical in some sense. We have the following theorem.

Lemma 16: Let the sequence of sequences $\{x^{n_k}\}$ be asymptotically $P_X$-typical and consider the output sequences $Y^{n_k}$ generated by the stationary memoryless channel $W_{Y|X}(\cdot|x)$. Then the sequence of sequences $\{(x^{n_k}, Y^{n_k})\}$ is asymptotically $P_X \otimes W_{Y|X}$-typical almost surely.
Remark 17: Consider the Markov chain \( X - Y - Z \). Then if a sequence of sequences \( \{(x^n, y^n)\} \) is asymptotically \( P_{XY} \)-typical and is used to generate output sequences \( Z^n \) according to the channel \( W_{Z|XY}(|x, y) = W_{Z|Y}(\cdot|y) \), then the sequences \( \{(x^n, y^n, Z^n)\} \) are almost surely asymptotically \( P_{XY} \otimes W_{Z|Y} \)-typical, i.e., the Markov Lemma holds for asymptotically typical sequences.

Proof: For ease of notation, we consider the case \( n_k = k \). Let \( P^k_X \) denote the empirical distribution of the \( k \)-length sequence \( x^k \). Let \( P_Y \) denote the marginal \( P_X W_{Y|X} \) and let \( P^k_{XY} \) denote the empirical distribution of the pair of sequences \( x^k \) and \( Y^k \).

The outline of the proof is as follows. First, for each of \( P_X \) and \( P_Y \), consider two fields \( \mathcal{F}_X \) and \( \mathcal{F}_Y \) as described in Corollary \[6\]. We will first show that almost surely for all \( A \in \mathcal{F}_X \), \( B \in \mathcal{F}_Y \),

\[
\lim_{k \to \infty} P^k_{XY}(A \times B) = P_X \otimes W_{Y|X}(A \times B).
\]

Thus, by [1, Chapter 1, Theorem 2.2], \( P_X \otimes W_{Y|X} = w\lim \) \( P^k_{XY} \) since each open set of \( E_X \times E_Y \) is a countable union of rectangular sets \( A \times B \), \( A \in \mathcal{F}_X \), \( B \in \mathcal{F}_Y \).

Now, consider a set \( A \in \mathcal{F}_X \) and \( B \in \mathcal{F}_Y \) and observe that

\[
\left| P_{XY}(A \times B) - P^k_{XY}(A \times B) \right| 
\leq \left| P_{XY}(A \times B) - P^k_X \otimes W_{Y|X}(A \times B) \right| + \left| P^k_X \otimes W_{Y|X}(A \times B) - P^k_{XY}(A \times B) \right| , \tag{20}
\]

where \( P_{XY} = P_X \otimes W_{Y|X} \). From Lemma 2 of [5], since \( P_X = w\lim \) \( P^k_X \), then \( P_X \otimes W_{Y|X} = w\lim \) \( P^k_X \otimes W_{Y|X} \). Since \( P_X \otimes W_{Y|X}(\partial(A \times B)) \leq P_X(\partial A) + P_Y(\partial B) = 0 \), it follows that \( \lim_k P^k_X \otimes W_{Y|X}(A \times B) = P_X \otimes W_{Y|X}(A \times B) \) and thus the first term on the right side of (20) is 0 in the limit.

As for the second term on the right side of (20), we note that

\[
P^k_X \otimes W_{Y|X}(A \times B) - P^k_{XY}(A \times B) = \frac{1}{k} \sum_{i=1}^{k} 1_{\{x_i \in A\}} \left[ W_{Y|X}(B|x_i) - 1_{\{Y_i \in B\}} \right] \tag{21}
\]

Let \( Z_i = 1_{\{x_i \in A\}} \left[ W_{Y|X}(B|x_i) - 1_{\{Y_i \in B\}} \right] \). Then, the second term on the right of (20) is

\[
\left| \frac{1}{k} \sum_{i=1}^{k} Z_i \right| . \tag{22}
\]

Note that given the non-random sequence \( x^k \), i) the random variables \( Z_i \) are independent, ii) \( E[Z_i] = 0 \), and iii) \( \sup_i \text{var}[Z_i] \leq 1 \) since \( -1 \leq Z_i \leq 1 \). Then by [15, Theorem 5.29], the sum in (22) converges to 0 almost surely, i.e.,

\[
\lim_{k \to \infty} \left| P^k_X \otimes W_{Y|X}(A \times B) - P^k_{XY}(A \times B) \right| = 0 \tag{23}
\]

almost surely. Since the family of sets \( \mathcal{F}_X \) and \( \mathcal{F}_Y \) are countable, it follows that almost surely the right side of (20) vanishes as \( k \to \infty \) for all \( A \in \mathcal{F}_X \) and \( B \in \mathcal{F}_Y \).
Theorem 18: Let \( x^n \) be an input sequence to a stationary memoryless channel \( W_{Y|X}(\cdot|x) \) and let \( Y^n \) be the corresponding output sequence. For every \( \epsilon > 0 \) and \( \delta > 0 \), there exists an \( \bar{\epsilon}(\epsilon, \delta) \) such that if \( x^n \in A^n_{\epsilon(\epsilon, \delta)}(P_X) \) for all \( n \) greater than some \( N \), then
\[
\liminf_{n \to \infty} P \left( (x^n, Y^n) \in A^n_{\epsilon}(P_X \otimes W_{Y|X}) \right) > 1 - \delta. \tag{24}
\]

Proof: Suppose that for a given \( \epsilon \) and \( \delta \), no such \( \bar{\epsilon}(\epsilon, \delta) \) can be found. Then, we can find a sequence of sequences \( \{x^n_k\} \) with corresponding empirical measures \( P^n_k \) for increasing \( n_k \) such that \( d(P^n_k, P_X) < 1/k \), i.e., \( P_X = \text{w-lim}_k P^n_k \) and
\[
\liminf_{k \to \infty} P \left[ (x^n_k, Y^n_k) \in A^n_{\epsilon}(P_X \otimes W_{Y|X}) \right] \leq 1 - \delta. \tag{25}
\]
But this contradicts the almost sure asymptotic \( (P_X \otimes W_{Y|X}) \)-typicality of \( (x^n_k, Y^n_k) \) in Lemma 16. \( \blacksquare \)

C. Large Deviations

The next theorems provide some large deviations results for weak* typical sequences. The first theorem looks at the probability of a random i.i.d. sequence drawn according to a law \( P \) to be weak* \( M \)-typical. This is shown to be \( \approx 2^{-nD(M||P)} \) which is the same result as for weak and strong typical sequences. For example, let \( P_{XY} \) be the joint law for \( E_X \times E_Y \) with marginals \( P_X \) and \( P_Y \). Then the probability that a pair of sequences \( X \) and \( Y \) drawn according to \( P_X \otimes P_Y \) is \( P_{XY} \)-typical is \( \approx 2^{-nI(X;Y)} \).

The next two theorems then consider the more special case of when the sequence \( Y \) is non-random and known to be weak* typical but \( X \) is random. There, we again show that the probability that the pair of sequences is weak* typical is \( \approx 2^{-nI(X;Y)} \). This result is normally proved for strong typical sequences using the notion of conditional strongly typical sequences and no analog exists in general for weak typical sequences.

Theorem 19: Let \( P \) and \( M \) be measures on a common probability space \( (E, \mathcal{E}) \) and let a random sequence \( X^n \) be chosen i.i.d. according to the law \( \mathcal{L}(X_i) = P \), i.e., draw the sequence \( X^n \) according to the measure \( \mu^n = \otimes_{i=1}^n P \). Define the sequence of probabilities \( a_n = \mu^n(X^n \in A^n_{\epsilon}(M)) \), i.e., the probability that the drawn sequence is weak* \( (M, \epsilon) \)-typical. If \( D(M||P) \) is finite, then there is an \( \epsilon(\delta) > 0 \) such that for all \( \epsilon < \epsilon(\delta) \)
\[
-D(M||P) \leq \liminf \frac{1}{n} \log a_n \leq \limsup \frac{1}{n} \log a_n \leq -D(M||P) + \delta. \tag{26}
\]
If \( D(M||P) = \infty \), then for each \( L > 0 \) there is an \( \bar{\epsilon}(L) \) such that for \( \epsilon < \bar{\epsilon}(L) \), the right side of (26) is \(-L\).
Remark 20: Under these assumptions, it follows that for any $\bar{\delta} > 0$, and $\delta > 0$, there is a sufficiently large $N$ such that for all $n > N$,

$$
\mu^n(X^n \in A^n(M)) \leq 2^n(-D(M||P) + \delta + \bar{\delta}).
$$

(27)

Since both $\delta$ and $\bar{\delta}$ are arbitrary, they can be absorbed into a single "$\delta$" term.

Proof: Let $P^n_X$ be the empirical measure of the drawn sequence. We first show the lower bound. By the definition of weak* typicality, we recognize that

$$
a_n = \mu^n(X^n \in A^n(M)) = \mu^n(P^n_X \in B(M, \epsilon)),
$$

(28)

where $B(M, \epsilon)$ is an open ball in the space $M_1(E)$ in the weak* topology. Then, by Sanov’s Theorem (Corollary 6.2.3 of [6]) we have the large deviations principle

$$
-\inf_{\nu \in B(M, \epsilon)} \Lambda^*(\nu) \leq \liminf \frac{1}{n} \log \mu^n(P^n_X \in B(M, \epsilon)).
$$

(29)

The lower bound then follows since in the weak* topology $\Lambda^*(\nu) = D(\nu||P)$ (Lemma 6.2.13 of [6]) and the inf is bounded by selecting any choice of $\nu \in B(M, \epsilon)$, say $\nu = M$.

To prove the upper bound, we use a quantization argument. Consider a field $F_M$ as described in Corollary 8 for $M$.

If $D(M||P)$ is finite, then $M \ll P$, and for any $\delta > 0$, choose a sufficiently fine finite partition $Q \subset F_M$ of $E$ such that the induced discrete probabilities $Q_P$ and $Q_M$ of $P$ and $M$ on the atoms of $Q$ satisfy

$$
D(Q_M||Q_P) \geq D(M||P) - \delta/2.
$$

(30)

Otherwise, $D(M||P) = \infty$ and for each $L > 0$, we can find a partition $Q$ such that $D(Q_M||Q_P) > 2L$.

In either case, denote these atoms by $A_1, \ldots, A_K$ for some integer $K$. Let $Q^n_P$ be the induced discrete probability of the empirical measure $P^n_X$ on the atoms of $Q$. Then, the event $P^n_X \in B(M, \epsilon)$ implies that the discrete probabilities $|Q_M(A_k) - Q_P(A_k)| < \delta_1$ for all atoms and $\delta_1 \to 0$ as $\epsilon \to 0$ by weak* convergence since $M(\partial A_k) = 0$.

Now, $Q^n_P$ is an empirical probability for a random sequence over a finite alphabet. Let $\Gamma$ be the set of all probability distributions $Q_P$ on the atoms of $Q$ such that $|Q_M(A_k) - Q_P(A_k)| \leq \delta_1$ for all $k$. In the finite alphabet case we have the following well-known large deviations result ( [6, Theorem 2.1.10])

$$
\limsup \frac{1}{n} \log \mu^n(P^n_X \in B(M, \epsilon)) \leq \limsup \frac{1}{n} \log \mu^n(Q^n_P \in \Gamma) \leq -\inf_{Q_P \in \Gamma} D(Q_P||Q_P).
$$

(31)

$^4$Since $D(M||P)$ is finite, this is straightforward by Lemma 5.
If $D(M||P)$ is finite, then $M \ll P$, and the divergence $D(Q_\nu||Q_P)$ is continuous on the compact set of $Q_\nu$ such that $Q_\nu \ll Q_P$ which includes $Q_M$ and $D(Q_\nu||Q_P)$ is infinite otherwise. Thus for small enough $\delta_1 > 0$, the inf can be bounded by $D(Q_M||Q_P) - \delta_2$ where $\delta_2 \to 0$ as $\delta_1 \to 0$. Hence, pick $\epsilon$ small enough that $\delta_2 < \delta/2$.

If $D(M||P) = \infty$, there are two cases to consider.

First, if $Q_M \ll Q_P$, then the same argument as the finite $D(M||P)$ case above shows that

$$- \inf_{Q_\nu \in \Gamma} D(Q_\nu||Q_P) \leq -2L + \delta_2,$$

where $\delta_2 < L$ can be ensured by choosing $\epsilon$ small enough.

Second, if we do not have $Q_M \ll Q_P$, then there is a set $A \in \mathcal{Q}$ such that $Q_M(A) > 0$ and $Q_P(A) = 0$. If $\epsilon$ is chosen sufficiently small that $\delta_1 < M(A)/2$, then $Q_\nu(A) > 0$ for all $Q_\nu \in \Gamma$ and $D(Q_\nu||Q_P) = \infty$ for all $Q_\nu \in \Gamma$.

Either way, $- \inf_{Q_\nu \in \Gamma} D(Q_\nu||Q_P) \leq -L$, where $L > 0$ is arbitrary.

**Theorem 21**: Let $P_{XY}$ be a joint distribution on $E_X \times E_Y$ and $P_X$ and $P_Y$ denote its marginals. Let $y^n$ be a sequence and $X^n$ a random sequence drawn i.i.d. according to $\mu^n = \otimes_{i=1}^n P_X$. If $D(P_{XY}||P_X \times P_Y)$ is finite then for each $\delta > 0$, there are $\epsilon(\delta)$ and $\bar{\epsilon}(\delta)$ such that if $\epsilon < \epsilon(\delta)$, $\bar{\epsilon} < \bar{\epsilon}(\delta)$ and $y^n \in A^\epsilon_\nu(P_X)$ and $y^n \in A^{\bar{\epsilon}}_\nu(P_Y)$ for all $n$ greater than some $N$, then

$$\limsup_n \frac{1}{n} \log \mu^n((X^n, y^n) \in A^n_\nu(P_{XY})) \leq -D(P_{XY}||P_X \times P_Y) + \delta.$$  \hfill (33)

If $D(P_{XY}||P_X \times P_Y) = \infty$, then for every $L > 0$, there is a sufficiently small $\epsilon, \bar{\epsilon} > 0$ such that holds with the right side replaced by $-L$.

**Proof**: See Appendix.

**Theorem 22**: Let $P_{XY}$ be a joint distribution on $E_X \times E_Y$ and $P_X$ and $P_Y$ denote its marginals. Let $y^n$ be a sequence and $X^n$ a random sequence drawn i.i.d. according to $\mu^n = \otimes_{i=1}^n P_X$. Then, for each $\delta > 0$ and $\epsilon > 0$ there is an $\bar{\epsilon}(\epsilon, \delta) > 0$ such that if $y^n \in A^n_{\epsilon(\epsilon, \delta)}(P_Y)$ for all $n$ greater than some $N$, then

$$\liminf_n \frac{1}{n} \log \mu^n((X^n, y^n) \in A^n_{\epsilon}(P_{XY})) \geq -D(P_{XY}||P_X \times P_Y) - \delta.$$  \hfill (34)

**Proof**: See Appendix.

**V. Examples**

We now apply the notion of weak* typical sequences to prove achievability results for two channel coding examples. The first is the traditional point-to-point channel. While more general results can be
obtained using information spectrum methods, the example highlights the application of weak* typical sequences.

In the second example, we apply weak* typical sequences to Gel’fand-Pinsker channels. These results cannot be obtained for arbitrary Polish spaces using weak/strong typical sequences.

In this section, the cost constraint $g(x)$ is continuous and bounded. In Section VI, we will consider the Gaussian case with power constraint.

### A. Point-to-Point Channel

We consider communicating over a channel $W_{Y|X}$, where the alphabets $E_X$ and $E_Y$ are Polish spaces. For completeness, we briefly state some definitions.

An \((n, M, P_e)\) code is a set of \(M\) codewords $x_1, \ldots, x_M$ and a decoder $\phi : E^n_Y \to \{1, \ldots, M\}$ such that the average probability of error is

$$P_e = \frac{1}{M} \sum_{v=1}^{M} P[\phi(Y) \neq v | X = x_v].$$

(35)

A rate $R$ is said to be achievable if there is a sequence of codes $\{(n, M_n, P^n_e)\}$ with block lengths $n$, $R = \lim_{n} \frac{1}{n} \log M_n$, and probability of error $P^n_e \to 0$.

We will show the following well known result using weak* typical sequences.

**Theorem 23:** Let $W_{Y|X}$ be a communication channel with Polish input and output alphabets and input constraint $(g(x), \Gamma)$. Then any rate

$$R < \sup_{P_X : E_{P_X}[g(X)] < \Gamma} I(X; Y)$$

(36)

is achievable.

**Remark 24:** The converse can be obtained with the usual Fano inequality.

**Proof:** The proof follows the usual random coding argument with the exception that we now use the results derived for weak* typical sequences. Pick any $\gamma > 0$. We will bound the probability of error for a random code by $3\gamma$ for large enough $n$.

In particular, as usual, pick a $P_X$ which satisfies the constraint $E_{P_X}[g(X)] < \Gamma$. We generate $M_n = \lfloor 2^{nR} \rfloor$ codewords of length $n$ with each entry i.i.d. according to $P_X$, and denote these as $X_1, \ldots, X_{M_n}$.

The encoder transmits $X_V$ where $V$ is uniform among the indices $\{1, \ldots, M_n\}$. The decoder employs weak* typical decoding. Specifically, it looks for an index $v$ such that $(X_v, Y)$ are weak* $(P_X \otimes W_{Y|X}, \epsilon)$ typical for some $\epsilon > 0$ and declares $v$ as the transmitted index if such a $v$ exists and is unique. Otherwise, an error is declared.
By the usual symmetry, without loss of generality we assume that the index \( v = 1 \) is selected at the transmitter. The probability of error for a \((P_X \otimes W_{Y|X}, \epsilon)\)-typical decoder is then bounded as

\[
P_e^n \leq P[g(X_1) \geq \Gamma] + P[(X_1, Y) \notin A^n_\epsilon(P_X \otimes W_{Y|X})] + P[\cup_{v \neq 1} (X_v, Y) \in A^n_\epsilon (P_X \otimes W_{Y|X})].
\]  
(37)

By part 1 of Theorem 15 \( P_{e,n}^2 := P[(X_1, Y) \notin A^n_\epsilon (P_X \otimes W_{Y|X})] \to 0 \) as \( n \to 0 \). Thus \( P_{e,n}^2 < \gamma \) for all \( n \) larger than some \( N_2 \).

Second, we note that by Theorem 19 \( P_{e,n}^3 := P[\cup_{v \neq 1} (X_v, Y) \in A^n_\epsilon (P_X \otimes W_{Y|X})] \leq 2^{-n(I(X;Y) - \delta)} \) and \( \delta \to 0 \) as \( \epsilon \to 0 \). Thus, for large enough \( n \), the union bound implies

\[
P_{e,n}^3 := P[\cup_{v \neq 1} (X_v, Y) \in A^n_\epsilon (P_X \otimes W_{Y|X})] \leq 2^n R 2^{-n(I(X;Y) - \delta)},
\]  
(39)

and \( P_{e,n}^3 < \gamma \) for all \( n \) larger than some \( N_3 \) provided \( R < I(X;Y) - \delta \).

Finally, one can upper bound \( P_{e,n}^1 := P[g(X_1) \geq \Gamma] \) by \( P[X_1 \notin A^n_\alpha(P_X)] + P[g(X_1) \geq \Gamma | X_1 \in A^n_\alpha(P_X)] \) for any arbitrary \( \alpha > 0 \).

By part 2 of Theorem 15 and since \( E_{P_X}[g(X)] < \Gamma \), there is a sufficiently small \( \alpha > 0 \) such that

\[
P_{e,n}^5 := P[g(X_1) \geq \Gamma | X_1 \in A^n_\alpha(P_X)] = 0.
\]

By part 1 of Theorem 15 for any \( \alpha > 0 \), \( P_{e,n}^4 := P[X_1 \notin A^n_\alpha(P_X)] \) vanishes as \( n \to \infty \). Thus \( P_{e,n}^4 < \gamma \) for all \( n \) larger than some \( N_4 \).

Thus for any rate \( R < I(X;Y) - \delta \), the bound in (37) is at most \( 3\gamma \) for all \( n \) sufficiently large. Finally, \( \delta > 0 \) can be made arbitrarily small by choosing \( \epsilon \) small enough.

\textbf{Remark 25:} Since \( E_{P_X}[g(X)] < \Gamma \) and each codeletter of each codeword is i.i.d., one could have bounded \( P[g(X_1) \geq \Gamma] \) by the strong law of large numbers. However, this approach will not be possible for Gel’fand-Pinsker channels as the channel input is not generated by independent and randomly chosen codeletters.

\textbf{B. Gel’fand-Pinsker Channels}

We now consider proving an achievability result for Gel’fand-Pinsker channels assuming Polish alphabets. The achievability result for \( R < I(U;Y) - I(U;S) \) was proved in the discrete case in [9]. The Gaussian case with additive interference and noise was considered in [4] and further results on additive

\[\text{\footnotesize{Here we assume that } I(X;Y) \text{ is finite. The case that } I(X;Y) = \infty \text{ can be considered separately.}}\]
interference and noise can be found in [3], [14], [20], [21]. Here, we consider achievability for a general channel \( W_{Y|SX} \) with Polish alphabets directly using weak* typical sequences.

We start with a brief set of definitions. A source sends a message \( V \in \{1, \ldots, M\} \) selected uniformly at random to a receiver by transmitting a sequence \( x \). The channel \( W_{Y|XS} \) results in an output \( Y \) that depends stochastically on the input \( x \) as well as an interference sequence \( S \), where \( S \) is an i.i.d. random sequence drawn according to \( P_S \). Furthermore, the encoder is aware of the interference sequence \( S \) apriori and the decoder is unaware of the interference. Thus, the encoder is described by the mapping \( \phi^n_{tx} : \{1, \ldots, M\} \times E^n_S \to E^n_X \) while the decoder is the mapping \( \phi^n_{rx} : E^n_Y \to \{1, \ldots, M\} \).

A code is a tuple \( (\phi^n_{tx}, \phi^n_{rx}, P_e) \) where \( P_e = P[\phi^n_{rx}(Y) \neq V|X = \phi^n_{tx}(V,Y)] \). A rate \( R \) is said to be achievable if there exists a sequence of codes \( (\phi^n_{tx}, M_n, \phi^n_{rx}, P^n_e) \) with \( \lim_n \frac{1}{n} \log M_n = R \) and \( \lim_n P^n_e = 0 \).

We have the following achievability result for Polish alphabets.

**Theorem 26:** For Gel'fand-Pinsker channels with cost constraint \( (g(x), \Gamma) \) at the transmitter, any rate

\[
R < \sup_{P_{U|S}, W_{X|US}: E[g(X)] < \Gamma} I(U;Y) - I(U;S),
\]

is achievable where the supremum is over all transition kernels \( P_{U|S} \) and all channels \( W_{X|US} \).

**Remark 27:** Recall that a channel is a transition kernel that satisfies a weak* continuity condition.

**Proof:** Again, the random coding argument is followed, however we now invoke weak* typical sequences. Pick any \( \gamma > 0 \). We will show that for any \( n \) larger than some \( N \), the probability of error with a random codebook is at most \( 4\gamma \).

Specifically, pick a transition kernel \( P_{U|S} \) and channel \( W_{X|US} \) which satisfy the constraint \( E[g(X)] < \Gamma \) and let \( \delta > 0 \). First, construct \( M_n = [2^{nR}] \) bins, with each bin containing \( [2^{n(I(U;S) + \delta)}] \) sequences of length \( n \) with each codeletter generated i.i.d. according to the marginal \( P_Y \). We denote these sequences as \( U_1, U_2, \ldots, U_K \) where \( K = [2^{nR}] \times [2^{n(I(U;S) + \delta)}] \).

Following the usual argument, to encode message \( v \in \{1, \ldots, M_n\} \), the encoder looks in bin \( v \) for a sequence \( U_i \) such that \( (U_i, S) \in A^n_{\epsilon_1}(P_{US}) \), i.e. \( (U_i, S) \) are weakly* \( (P_{US}, \epsilon_1) \) typical for some appropriate \( \epsilon_1 > 0 \).

If there is no such \( U_i \) sequence, an error is declared, which is denoted by the event \( E_1 \). Otherwise, the encoder constructs a sequence \( X \) generated by the memoryless channel \( W_{X|US} \). If

\[
\frac{1}{n} \sum_{\ell=1}^{n} g(X_{\ell}) \geq \Gamma,
\]

(DRAFT)
then the transmission of $X$ would violate the channel input constraint and an error is declared, denoted by the event $E_2$. Otherwise, $X$ is transmitted over the channel.

The receiver obtains $Y$ and looks in all bins for a pair of sequences $(U_j, Y)$ that are jointly $(P_{US}, \epsilon_3)$-typical for some appropriate $\epsilon_3$. If there is a unique such pair, then the bin index in which $U_j$ is present is declared as the estimate $\hat{v}$ of $v$. If the index is not unique, or the bin incorrect, we denote this error event by $E_3$.

The probability of error is bounded by

$$P_e \leq P[E_0] + P[E_1 \cap \bar{E}_0] + P[E_2]P[\bar{E}_1] + P[E_3],$$

where $E_0$ is the event that $S$ is not $(\epsilon_0, P_S)$-typical for some suitable $\epsilon_0 > 0$.

**Error analysis:**

We start by analyzing $P[E_3|\bar{E}_1]$ and note that

$$P[E_3|\bar{E}_1] \leq P[E_4|\bar{E}_1] + P[E_5|\bar{E}_1],$$

where $E_4$ is the event that $(U_i, Y)$ are not jointly $(P_{US}, \epsilon_3)$-typical and $E_5$ is the event that there is an index $j \neq i$ such that $(U_j, Y)$ are jointly $(P_{US}, \epsilon_3)$-typical.

By applying Theorem 15 twice, first to the channel $W_{X|US}$ and then to the channel $W_{Y|XS}$ and using Part 3 of Theorem 15, there is an $\epsilon_{us}(\epsilon_3, \gamma/2)$ such that for any $\epsilon_{us} \leq \epsilon_{us}(\epsilon_3, \gamma/2)$, if $(U_i, S)$ is $(P_{US}, \epsilon_{us})$-typical, then $P_{e,n}^4 := P[E_4|\bar{E}_1] < \gamma/2$ for all $n$ greater than some $N_4$.

Now, conditioned on $\bar{E}_4$, by Part 3 of Theorem 15, $Y = y$ is $(P_Y, \epsilon_Y)$-typical, where $\epsilon_Y \to 0$ as $\epsilon_3 \to 0$. By Theorem 21 for any $\delta_3 > 0$, there is sufficiently small $\epsilon_3(\delta_3)$, $\epsilon_3(\delta_3)$ and large $N_5$ such that provided $\epsilon_3 < \epsilon_3(\delta_3)$ and $y \in A^n_{\epsilon_3(\delta_3)}$ for $n > N_5$, by the usual union bound

$$P[E_5|\bar{E}_1, \bar{E}_4] \leq 2^{n(I(U:S)+\delta)}2^{-n(I(U;Y)-\delta_3)},$$

Thus, selecting $\epsilon_3 < \epsilon_3(\delta_3)$ and $\epsilon_3$ small enough that $\epsilon_Y < \epsilon_3(\delta_3)$, $P_{e,n}^5 := P[E_5|\bar{E}_1, \bar{E}_4] < \gamma/2$ for all $n$ greater than some $N_5$ provided $R < I(U;Y) - I(U;S) - \delta_3 - \delta$. Note that $\delta_3 > 0$ and $\delta > 0$ are arbitrary.

We now analyze $P[E_2|\bar{E}_1]$. For any $\epsilon_2 > 0$, there is an $\epsilon_1(\epsilon_2, \gamma) > 0$ such that for $\epsilon_1 < \epsilon_1(\epsilon_2, \gamma)$ by Theorem 15

$$P_{e,n}^2 := P[(U, S, X) \notin A_{\epsilon_2}(P_{USX})] < \gamma$$

for all $n$ larger than some $N_2$. By Part 2 of Theorem 15, there is small enough $\epsilon_2$ such that $P[g(X) \geq \Gamma | (U, S, X) \in A^n_{\epsilon_2}(P_{USX})] = 0$. Select $\epsilon_1 < \min\{\epsilon_{us}(\epsilon_3, \gamma/2), \epsilon_1(\epsilon_2, \gamma)\}$. 

DRAFT
We now analyze $P[E_1 \cap \tilde{E}_0]$. Let $\delta_1 > 0$ be arbitrary. By Theorem 22 there is an $\bar{N}_1$, $\epsilon_0 = \epsilon_0(\epsilon_1, \delta_1)$ such that conditioned on the fact that if $S = s$ is $(P_S, \epsilon_0)$-typical,

$$P[(U, s) \in A_{\epsilon_1}^n(P_{US})] \geq 2^{-n(I(U;S)+\delta_1)}, \quad (46)$$

for $n > \bar{N}_1$. Thus,

$$P[E_1 \cap \tilde{E}_0] = E_{P_S} \left[ P[E_1|S = s]1\{s \in A_{\epsilon_0}^n(P_S)\} \right] \quad (47)$$

$$= E_{P_S} \left[ 1 - P[(U, s) \in A_{\epsilon_1}^n(P_{US})] \right] 2^{n(I(U;S)+\delta_1)} 1\{s \in A_{\epsilon_0}^n(P_S)\} \quad (48)$$

$$\leq E_{P_S} \left[ 1 - 2^{-n(I(U;S)+\delta_1)} \right] 2^{n(I(U;S)+\delta_1)} 1\{s \in A_{\epsilon_0}^n(P_S)\} \quad (49)$$

$$\leq \left[ 1 - 2^{-n(I(U;S)+\delta_1)} \right] 2^{n(I(U;S)+\delta_1)} \quad (50)$$

$$\leq \exp \left(-2^{-n(\delta_1 - \delta)}\right), \quad (51)$$

where $P_S = \otimes_{i=1}^n P_S$. Thus, selecting $\delta_1 < \delta$ yields $P_{e,n}^1 := P[E_1|\tilde{E}_0] < \gamma$ for all $n$ greater than some $N_1$.

Finally, we analyze $P[E_0]$. However, by Part I of Theorem 15 for any $\epsilon_0 > 0$, $P_{e,0}^n := P[S \notin A_{\epsilon_0}^n(P_S)]$ vanishes as $n \to 0$ and thus $P_{e,0}^n < \gamma$ for all $n$ greater than some $N_0$. \hfill \blacksquare

Remark 28: We note the following remarks. First, we could not rely on the law of large numbers to argue that $X$ would satisfy the constraint pair $(g(x), \Gamma)$. This is because while $X$ was generated stochastically, it was done so based on the pair $(U_i, S)$, where $U_i$ was specifically chosen to satisfy a given property. Thus instead we argued via Theorem 18 that the triple $(U_i, S, X)$ is $\epsilon_2$-typical (by the channel consistency property or Markov Lemma) and then that $X$ must satisfy the power constraint for sufficiently small $\epsilon_2$.

Second, we again employed the channel consistency property (or Markov Lemma) to prove that the pair $(U_i, Y)$ are jointly typical with high probability.

VI. THE GAUSSIAN CASE

In Section V achievability results were proved for a point-to-point channel as well as the Gel’fand-Pinsker channel with input constraints. It was noted that due to the input constraint $(g(x), \Gamma)$, either the (continuous) cost function $g(x)$ should be bounded, or the input alphabet is compact (which trivially implies $g(x)$ is bounded).

This rules out the consideration of an input constraint $(g(x) = x^2, \sigma^2_X)$ with a Gaussian input distribution as neither the cost function nor the input alphabet is then bounded.
In this section, we show how one can recover the traditional achievability results in both cases for Gaussian distributions. Specifically, we will consider an input alphabet over the interval $E_{X_L} = [-L, L]$ and show that as $L \to \infty$, one can arbitrarily approach the well-known results in the Gaussian case. It should be noted that we consider all alphabets as subsets of $\mathbb{R}$ for simplicity of exposition only and the arguments apply equally well to alphabets over $\mathbb{R}^n$.

A. Point-to-Point Channel

Here the capacity of the channel $Y = X + Z$ with $Z \sim \mathcal{N}(0, \sigma^2_Z)$ is well known to be $C = I(Y; X)$ evaluated for $X \sim \mathcal{N}(0, \sigma^2_X)$.

Now, consider the family of random variables $X_L$ (indexed by $L > 0$) with densities

$$f_{X_L}(x) = \begin{cases} 
0 & x < -L \\
\frac{f_X(x)}{K(L)} & -L \leq x \leq L \\
0 & x > L 
\end{cases} \tag{52}$$

where $f_X(x)$ is the PDF of an $\mathcal{N}(0, \sigma^2_X)$ distribution and

$$K(L) = \int_{-L}^{L} f_X(x) \, dx \tag{53}$$

is a normalization constant with the property that $\lim_{L \to \infty} K(L) = 1$.

For notational convenience, let the output random variable be denoted by $Y_L$ when the input is $X_L$, i.e., $Y_L = X_L + Z$, and let the output random variable be $Y$ when the input is $X \sim \mathcal{N}(0, \sigma^2_X)$, i.e., $Y = X + Z$. It is straightforward to verify that $E[X_L^2] < P$ for all $L$, and thus this input distribution satisfies the input constraint, and the rate

$$R_L = I(Y_L; X_L) \tag{54}$$

is achievable for each $L$. We will show that as $L \to \infty$, that $R_L = I(Y_L; X_L) \to I(X; Y)$.

First, note that

$$I(Y_L; X_L) = h(Y_L) - h(Y_L|X_L) \tag{55}$$

$$= h(Y_L) - h(Z) \tag{56}$$

and thus it suffices to show that $\lim_{L \to \infty} h(Y_L) = h(Y)$.

\footnote{While $X \sim \mathcal{N}(0, \sigma^2_X)$ does not satisfy the input constraint, $X \sim \mathcal{N}(0, \sigma^2_X - \epsilon); 0 < \epsilon < \sigma^2_X$ does and the capacity follows by a simple limit argument.}

\[ \text{DRAFT} \]
Second,
\[ f_{Y_L}(y) = \int_{-\infty}^{\infty} f_{Y|X}(y|x) f_{X_L}(x) \, dx \]
\[ = \frac{1}{K(L)} \int_{-L}^{L} f_{Y|X}(y|x) f_{X}(x) \, dx \]  
Define
\[ g_{Y_L}(y) := \int_{-L}^{L} f_{Y|X}(y|x) f_{X}(x) \, dx, \]
then \( f_{Y_L}(y) = g_{Y_L}(y)/K(L), \) and
\[ h(Y_L) = \int_{-\infty}^{\infty} f_{Y_L}(y) \log 1/f_{Y_L}(y) \, dy \]
\[ = \frac{1}{K(L)} \left[ \int_{-\infty}^{\infty} g_{Y_L}(y) \log 1/g_{Y_L}(y) \, dy \right] - \int_{-\infty}^{\infty} f_{Y_L}(y) \log K(L) \, dy \]
\[ = - \frac{1}{K(L)} \left[ \int_{-\infty}^{\infty} g_{Y_L}(y) \log g_{Y_L}(y) \, dy \right] - \log K(L). \]  

Since \( \lim_{L \to \infty} K(L) = 1, \) it remains only to show that the term in the square brackets converges to \(-h(Y)\) as \( L \to \infty. \) However, because of (59), \( g_{Y_L}(y) \) is continuous, strictly positive, strictly increasing in \( L \) and converges pointwise to \( f_Y(y). \) Thus, the integrand \( g_{Y_L}(y) \log g_{Y_L}(y) \) is continuous and converges pointwise to \( f_Y(y) \log f_Y(y) \) as \( L \to \infty. \)

Let \( A = \{ y \in \mathbb{R} | f_Y(y) < e^{-1} \}. \) Since \( x \log x \) is decreasing in \( x \) for \( 0 \leq x < e^{-1}, \) then for \( y \in A, \)
\( f_{Y_L}(y) \log f_{Y_L}(y) \) is decreasing in \( L \) and
\[ \lim_{L \to \infty} \int_A g_{Y_L}(y) \log g_{Y_L}(y) \, dy = \int_A f_Y(y) \log f_Y(y) \, dy \]  
by the (Lebesgue) monotone convergence theorem.

Now consider the set \( B = A^c = \{ y \in \mathbb{R} | f_Y(y) \geq e^{-1} \} \) and note that \( B \) is closed (since \( f_Y(y) \) is continuous) and bounded (since \( f_Y(y) \) is a Gaussian pdf) and thus \( B \) is compact. Thus, on the set \( B, \)
\( f_{Y_L}(y) \) converges uniformly to \( f_Y(y) \) by Dini’s theorem. Hence, there is a large enough \( L \) such that for all \( L > \bar{L}, \) \( f_{Y_L}(y) > e^{-2} \) for all \( y \in B. \) Let \( K = \sup_{y \in \mathbb{R}} f_Y(y). \) Then for \( y \in B \) and \( L > \bar{L}, \)
\( |g_{Y_L}(y) \log g_{Y_L}(y)| \leq f_Y(y) \max\{2, |\log K|\} \) which is integrable. Thus, by the dominated convergence theorem,
\[ \lim_{L \to \infty} \int_B g_{Y_L}(y) \log g_{Y_L}(y) \, dy = \int_B f_Y(y) \log f_Y(y) \, dy, \]  

DRAFT
and therefore
\[
\lim_{L \to \infty} \int_{-\infty}^{\infty} g_{Y_L}(y) \log g_{Y_L}(y) \, dy = \lim_{L \to \infty} \int_A g_{Y_L}(y) \log g_{Y_L}(y) \, dy + \lim_{L \to \infty} \int_B g_{Y_L}(y) \log g_{Y_L}(y) \, dy \\
= \int_{-\infty}^{\infty} f_Y(y) \log f_Y(y) \, dy \\
= -h(Y),
\]
(65) as desired.

**B. Gel’fand-Pinsker Channels**

In the Gaussian case, it is well-known that the capacity is obtained with the choice
\[
U = X + \alpha S \\
Y = X + Z + S,
\]
(68) where \(X \sim \mathcal{N}(0, \sigma_X^2)\) and independent of \(S\), and \(\alpha\) is an appropriately chosen constant.

We follow the same strategy as in Section VI-A. Namely, we consider the family of truncated Gaussians \(X_L\) given in (52), and obtain
\[
U_L = X_L + \alpha S \\
Y_L = X_L + Z + S.
\]
(70) As previously, we will show that
\[
\lim_{L \to \infty} I(U_L; S) = I(U; S) \tag{72}
\]
\[
\lim_{L \to \infty} I(U_L; Y_L) = I(U; Y). \tag{73}
\]

First, note that \(I(U_L; S) = h(U_L) - h(X_L)\). Second, \(\lim_{L \to \infty} h(U_L) = h(U)\) follows by the same argument as in Section VI-A with \(U_L\) replaced by \(Y_L\) and \(\alpha S\) replaced by \(Z\). Third, \(\lim_{L \to \infty} h(X_L) = h(X)\) since
\[
\lim_{L \to \infty} \int_{-L}^{L} f_X(x) \log f_X(x) \, dx = \int_{-\infty}^{\infty} f_X(x) \log f_X(x) \, dx.
\]
(74) Next, we note that \(I(U_L; Y_L) = h(U_L) + h(Y_L) - h(U_L, Y_L)\) and \(\lim_{L \to \infty} h(U_L) = h(U)\) was already argued, and \(\lim_{L \to \infty} h(Y_L) = h(Y)\) follows similarly. We provide an outline of \(\lim_{L \to \infty} h(U_L, Y_L) = h(U, Y)\). Because of the additive nature, \(f_{U || X}(u, y|x)\) denotes the conditional PDF of \(U\) and \(Y\) given \(X\), then
\[
f_{U_L, Y_L}(u, y) = \frac{1}{K(L)} \int_{-L}^{L} f_{U || X}(u, y|x)f_X(x) \, dx.
\]
(75)
Thus, following the same argument as in Section VI-A one can apply the monotone and dominated convergence theorems and obtain $\lim_{L \to \infty} h(U_L, Y_L) = h(U, Y)$.

VII. CONCLUSION

In this paper, a notion of typical sequences based on the weak* topology was defined. This notion of typical sequence applies to discrete, continuous and mixed distribution and was shown to satisfy consistency properties normally associated with strongly typical sequences. As examples of applying these notions of typical sequences, achievable rates were proved for the traditional point-to-point channel and Gel’fand-Pinsker channels with Polish alphabets and input constraints.

APPENDIX

Proof of Theorem 21: Pick two fields $\mathcal{F}_X$ and $\mathcal{F}_Y$ as described in Corollary 8 and let $Q_X \subset \mathcal{F}_X$ and $Q_Y \subset \mathcal{F}_Y$ be two partitions of size $|Q_X| = N_X$ and $|Q_Y| = N_Y$ and elements $Q_X = \{A_1, \ldots, A_{N_X}\}$, $Q_Y = \{B_1, \ldots, B_{N_Y}\}$.

Let $P^n_{XY}$ be the empirical measure induced by the pair of sequences $(X^n, y^n)$. Let $Q^n_{XY}, Q^n_X$ and $Q^n_Y$ denote the empirical measures and empirical marginals induced on the partitions $Q_X \times Q_Y$. Furthermore, for $B_j$ such that $Q^n_Y(B_j) > 0$, define the conditional measure $Q^n_{X|Y}(A_i|B_j) = Q^n_{XY}(A_i \times B_j)/Q^n_Y(B_j)$, otherwise $Q^n_{X|Y}(A_i|B_j)$ is arbitrary.

Likewise, starting with $P_{XY}$, let $Q_{XY}, Q_X$ and $Q_Y$ denote the induced measures and marginals on the partitions $Q_X \times Q_Y$. For $B_j$ such that $Q_Y(B_j) > 0$, define $Q_{X|Y}(A_i|B_j) = Q_{XY}(A_i \times B_j)/Q_Y(B_j)$, and note that $Q_{X|Y}(.,B_j) \ll Q_X(\cdot)$. Otherwise pick $Q_{X|Y}(A_i|B_j)$ to be some arbitrary distribution for which $Q_{X|Y}(.,B_j) \ll Q_X(\cdot)$.

Now, pick an $\epsilon_1 > 0$ and note that $(X^n, y^n) \in A^n_{\epsilon}(P_{XY})$ implies $P^n_{XY} \in B(P_{XY}, \epsilon)$ which for sufficiently small $\epsilon$, itself implies that for all $i$ and $j$,

$$|Q^n_{XY}(A_i \times B_j) - Q_{XY}(A_i \times B_j)| < \epsilon_1.$$  

Therefore, with this choice of $\epsilon$,

$$\mu^n((X^n, y^n) \in A^n_{\epsilon}(P_{XY})) \leq \mu^n \left( \bigcap_{i,j} \{|Q^n_{XY}(A_i \times B_j) - Q_{XY}(A_i \times B_j)| < \epsilon_1\} \right).$$  

(76)
Let $\mathcal{J}_{>0}$ denote the set of $j$ such that $Q_Y(B_j) > 0$ and select $\bar{\epsilon}(\delta) > 0$ small enough such that $y^n \in A^n_{\bar{\epsilon}(\delta)}(P_Y)$ implies

$$
|Q^n_{Y}(B_j) - Q_Y(B_j)| < \epsilon_1 \quad \forall j
$$

(78)

$$
1 - \epsilon_1 < \frac{Q^n_{Y}(B_j)}{Q_Y(B_j)} \leq 1 + \epsilon_1 \quad \forall j \in \mathcal{J}_{>0}.
$$

(79)

Now, by (78), for all $j \notin \mathcal{J}_{>0}$ and all $i$, (76) is satisfied and the right side of the bound in (77) can be limited to the intersection of all $i$ and all $j \in \mathcal{J}_{>0}$. Furthermore, since for $j \in \mathcal{J}_{>0}$, (76) implies

$$
\left| Q^n_{\mathcal{X}|\mathcal{Y}}(A_i|B_j)Q^n_{\mathcal{Y}}(B_j) - Q_{\mathcal{X}|\mathcal{Y}}(A_i|B_j)Q^n_{\mathcal{Y}}(B_j) + Q_{\mathcal{X}|\mathcal{Y}}(A_i|B_j)Q^n_{\mathcal{Y}}(B_j) - Q_{\mathcal{X}|\mathcal{Y}}(A_i|B_j)Q_{\mathcal{Y}}(B_j) \right| < \epsilon_1,
$$

(80)

together with (78), this implies

$$
\left| Q^n_{\mathcal{X}|\mathcal{Y}}(A_i|B_j) - Q_{\mathcal{X}|\mathcal{Y}}(A_i|B_j) \right| Q^n_{\mathcal{Y}}(B_j) < 2\epsilon_1,
$$

(81)

and for $j \in \mathcal{J}_{>0}$, with (72) this implies

$$
\left| Q^n_{\mathcal{X}|\mathcal{Y}}(A_i|B_j) - Q_{\mathcal{X}|\mathcal{Y}}(A_i|B_j) \right| < \frac{2\epsilon_1}{Q_{\mathcal{Y}}(B_j)(1 - \epsilon_1)}.
$$

(82)

Let $\epsilon_2 = \max_{j \in \mathcal{J}_{>0}} \frac{2\epsilon_1}{Q_{\mathcal{Y}}(B_j)(1 - \epsilon_1)}$. Then $\epsilon_2 \to 0$ as $\epsilon_1 \to 0$ and for all $i$ and $j \in \mathcal{J}_{>0}$,

$$
\left| Q^n_{\mathcal{X}|\mathcal{Y}}(A_i|B_j) - Q_{\mathcal{X}|\mathcal{Y}}(A_i|B_j) \right| < \epsilon_2.
$$

(83)

Therefore, we have shown that

$$
\mu^n ((\mathcal{X}^n, y^n) \in A^n_{\epsilon}(P_{\mathcal{X}Y})) \leq \mu^n \left( \cap_{i,j \in \mathcal{J}_{>0}} \left\{ \left| Q^n_{\mathcal{X}|\mathcal{Y}}(A_i|B_j) - Q_{\mathcal{X}|\mathcal{Y}}(A_i|B_j) \right| < \epsilon_2 \right\} \right) \cap_i \left( \left| Q^n_{\mathcal{X}|\mathcal{Y}}(A_i|B_j) - Q_{\mathcal{X}|\mathcal{Y}}(A_i|B_j) \right| < \epsilon_2 \right) \\
= \prod_{j \in \mathcal{J}_{>0}} \mu^n \left( \cap_i \left\{ \left| Q^n_{\mathcal{X}|\mathcal{Y}}(A_i|B_j) - Q_{\mathcal{X}|\mathcal{Y}}(A_i|B_j) \right| < \epsilon_2 \right\} \right).
$$

(84)

Now, let $N_{n,j} = \sum_{\ell=1}^n 1_{\{y^\ell \in B_j\}}$ for any $j \in \mathcal{J}_{>0}$, i.e., the number of letters of $y^n$ that are in $B_j$. Then for a given $j \in \mathcal{J}_{>0}$,

$$
\limsup_{n \to \infty} \frac{1}{n} \log \mu^n \left( \cap_i \left\{ \left| Q^n_{\mathcal{X}|\mathcal{Y}}(A_i|B_j) - Q_{\mathcal{X}|\mathcal{Y}}(A_i|B_j) \right| < \epsilon_2 \right\} \right) \\
= \limsup_{n \to \infty} \frac{N_{n,j}}{n} \times \limsup_{n \to \infty} \frac{1}{N_{n,j}} \log \mu^n \left( \cap_i \left\{ \left| Q^n_{\mathcal{X}|\mathcal{Y}}(A_i|B_j) - Q_{\mathcal{X}|\mathcal{Y}}(A_i|B_j) \right| < \epsilon_2 \right\} \right) \leq -(1 - \epsilon_1)Q_{\mathcal{Y}}(B_j) \times \left[ D(Q_{\mathcal{X}|\mathcal{Y}}(\cdot|B_j)||Q_{\mathcal{X}}(\cdot)) - \delta_{1,j} \right].
$$

(86)
where \( \delta_{1,j} \to 0 \) as \( \epsilon_2 \to 0 \) since \( Q_{X|Y}(\cdot|B_j) \ll Q_X(\cdot) \) and we have used Theorem 2.1.10 of [6]. Therefore,

\[
\limsup_{n \to \infty} \frac{1}{n} \log \mu^n ((X^n, Y^n) \in A^n_{\epsilon_1}(P_{XY})) \leq -(1 - \epsilon_1)D(Q_{XY}||Q_X \times Q_Y) + \delta_2
\]

(88)

\[
= -(1 - \epsilon_1)H_{P_{XY}}|P_X \times P_Y (Q_X \times Q_Y) + \delta_2,
\]

(89)

where \( \delta_2 = (1 - \epsilon_1) \sum_{j \in \mathcal{J} > 0} \delta_{1,j} \).

If \( D(P_{XY}||P_X \times P_Y) \) is finite, the result then follows by first choosing appropriate fine quantizers \( Q_X \) and \( Q_Y \) such that

\[
-H_{P_{XY}}|P_X \times P_Y (Q_X \times Q_Y) < -D(P_{XY}||P_X \times P_Y) + \delta/2,
\]

(90)

and then choosing \( \epsilon_1 \) small enough (thus \( \epsilon(\delta) \) and \( \bar{\epsilon}(\delta) \) small enough) such that

\[
-(1 - \epsilon_1)H_{P_{XY}}|P_X \times P_Y (Q_X \times Q_Y) + \delta_2 \leq -D(P_{XY}||P_X \times P_Y) + \delta.
\]

(91)

If \( D(P_{XY}||P_X \times P_Y) = \infty \) then for every \( L > 0 \), we can find a pair of quantizers such that

\[
H_{P_{XY}}|P_X \times P_Y (Q_X \times Q_Y) > 2L.
\]

The result follows by choosing \( \epsilon_1 \) small enough (thus \( \epsilon \) and \( \bar{\epsilon} \) small enough) such that \( (1 - \epsilon_1)H_{P_{XY}}|P_X \times P_Y (Q_X \times Q_Y) - \delta_2 > L \). Thus the right side of (33) is less than \(-L\) for any positive \( L \).

\[\square\]

**Proof of Theorem 22.** The case that \( D(P_{XY}||P_X \times P_Y) = \infty \) is trivial, thus we only consider finite \( D(P_{XY}||P_X \times P_Y) \).

We first show that for each \( \epsilon > 0 \), there is a finite partition \( Q_X = \{A_i\} \) and \( Q_Y = \{B_j\} \) of \( E_X \) and \( E_Y \), and \( \lambda(\epsilon) \) such that

\[
\{P^n_{XY} \in B(P_{XY}, \epsilon)\} \supset \bigcap_{i,j} \{||P^n_{XY}(A_i \times B_j) - P_{XY}(A_i \times B_j)|| < \lambda(\epsilon)\}.
\]

(92)

and \( \lambda(\epsilon) \to 0 \) as \( \epsilon \to 0 \).

To see this, let \( \mathcal{F}_X \) and \( \mathcal{F}_Y \) be fields as described in Corollary 8. For \( k = 1, 2, \ldots \), let \( \mathcal{Q}^k_X \subset \mathcal{F}_X \) be a sequence of successively finer finite partitions of \( E_X \) in the sense that if \( A \in \mathcal{Q}^k_X \) then \( A \) is the union of atoms of \( \mathcal{Q}^{k+1}_X \) and for each \( A \in \mathcal{F}_X \), \( A \) is the union of atoms of \( \mathcal{Q}^k_X \) for some \( k \). Likewise for \( \mathcal{Q}^k_Y \subset \mathcal{F}_Y \). We denote the atoms of \( \mathcal{Q}^k_X \) by \( A^k_i : i = 1, \ldots, |\mathcal{Q}^k_X| := N^k_X \), and likewise for the atoms \( B^k_j \) of \( \mathcal{Q}^k_Y \).

Consider any sequence \( M^n_{XY} \) of distributions that satisfy the sequence of events

\[
\bigcap_{i,j} \{||M^n_{XY}(A^k_i \times B^k_j) - P_{XY}(A^k_i \times B^k_j)|| < \lambda_k\}
\]

(93)

DRAFT
where $\lambda_k := 1/(k \times N_X^k \times N_Y^k)$. Then for any $A \in \mathcal{F}_X$ and $B \in \mathcal{F}_Y$, $M_{XY}^k(A \times B) \to P_{XY}(A \times B)$. Thus by [1, Chapter 1, Theorem 2.2], the sequence of events in (93) implies $P_{XY} = \text{w-lim}_k M_{XY}^k$. This implies $\lim_k d(M_{XY}^k, P_{XY}) = 0$. Let $\epsilon_k = \sup d(M_{XY}^k, P_{XY})$, where the supremum is over $M_{XY}^k$ such that (93) holds at the $k$th step. We must have $\epsilon_k \to 0$ or there would be a choice of $M_{XY}^k$ satisfying (93) such that $P_{XY} = \text{w-lim}_k M_{XY}^k$ does not hold. Let $K$ be such that $\epsilon_K < \epsilon$.

Hence, we can pick $Q_X = Q_X^K$, $Q_Y = Q_Y^K$ and any $\lambda(\epsilon) \leq \lambda_K$. Therefore, with these choices,

$$
\liminf_{n \to \infty} \frac{1}{n} \log \mu^n((X^n, Y^n) \in A^n_k(P_{XY})) \\
\geq \liminf_{n \to \infty} \frac{1}{n} \log \mu^n \left( \bigcap_{i,j} \{|P_{XY}^n(A_i \times B_j) - P_{XY}(A_i \times B_j)| < \lambda(\epsilon)\} \right) \\
\geq (a) - D(P_{XY} \parallel P_X \times P_Y) - \delta,
$$

(94)

where inequality $(a)$ is justified below.

To justify inequality $(a)$, first let $Q_{XY}, Q_{XY}^n$, etc denote the appropriate induced distributions on the partitions of $Q_X$ and $Q_Y$ as in the proof of Theorem 21. Then, for any $\alpha > 0$,

$$
\left| Q_{X|Y}^n(A|B) - Q_{X|Y}(A|B) \right| < \alpha
$$

(96)

implies

$$
\left| Q_{XY}^n(A \times B) - Q_{XY}(A \times B) \right| < \alpha Q_Y(B) + |Q_Y^n(B) - Q_Y(B)|.
$$

(97)

If $\epsilon(\epsilon, \delta)$ is sufficiently small that $y^n \in A^n_{\epsilon(\epsilon, \delta)}(P_Y)$ implies $|Q_Y^n(B) - Q_Y(B)| < \alpha$ for all $B \in \mathcal{Q}_Y$ with $\alpha < \lambda(\epsilon)/2$, then the event

$$
\bigcap_{i,j} \left| Q_{X|Y}^n(A_i|B_j) - Q_{X|Y}(A_i|B_j) \right| < \alpha
$$

(98)

is a subset of

$$
\bigcap_{i,j} \{|P_{XY}^n(A_i \times B_j) - P_{XY}(A_i \times B_j)| < \lambda(\epsilon)\}.
$$

(99)
Thus
\[
\liminf_{n \to \infty} \frac{1}{n} \log \mu_n \left( \bigcap_{i,j} \{|P_{XY}^n(A_i \times B_j) - P_{XY}(A_i \times B_j)| < \lambda(\epsilon)\} \right) \\
\geq \liminf_{n \to \infty} \frac{1}{n} \log \mu_n \left( \bigcap_{i,j} \{|Q_{X|Y}^n(A_i|B_j) - Q_{X|Y}(A_i|B_j)| < \alpha\} \right) \\
\geq - \sum_j (Q(Y(B_j) + \alpha) D(Q_{X|Y}(|B_j)||Q_X(\cdot)) \\
= -H_{P_{XY}}||P_X \times P_Y - \alpha \sum_j D(Q_{X|Y}(|B_j)||Q_X(\cdot)) \\
\geq -D(P_{XY}||P_X \times P_Y) - \delta, \tag{103}
\]
where \(\delta = \alpha \sum_j D(Q_{X|Y}(|B_j)||Q_X(\cdot))\) is finite as \(D(P_{XY}||P_X \times P_Y) < \infty\). Furthermore, \(\delta\) can be made arbitrarily small by choosing \(\alpha\) small enough, which can be assured by choosing \(\varepsilon(\epsilon, \delta)\) small enough.

\[7\] If \(Q_Y(B_j) > 0\) then \(D(Q_{X|Y}(|B_j)||Q_X(\cdot))\) must be finite as otherwise, \(D(P_{XY}||P_X \times P_Y) \geq D(Q_{XY}||Q_X \times Q_Y) = \infty\). If \(Q_Y(B_j) = 0\), then \(Q_{X|Y}(|B_j)\) is arbitrary, and selecting \(Q_{X|Y}(|B_j) = P_X(\cdot)\) results in \(D(Q_{X|Y}(|B_j)||Q_X(\cdot)) = 0\).

References

[1] P. Billingsley, Convergence of Probability Measures, 2nd ed., ser. Probability and Statistics. Wiley, 1999.
[2] J. A. Bucklew, T. G. Kurtz, and W. A. Sethares, “Weak convergence and local stability properties of fixed step size recursive algorithms,” IEEE Trans. Inform. Theory, vol. 39, no. 3, pp. 966 – 978, Nov. 1993.
[3] A. S. Cohen and A. Lapidoth, “The Gaussian watermarking game,” IEEE Trans. Inform. Theory, vol. 48, no. 6, pp. 1639–1667, June 2002.
[4] M. H. M. Costa, “Writing on dirty paper,” IEEE Trans. Inform. Theory, vol. IT-29, pp. 439–441, May 1983.
[5] I. Csiszár, “Arbitrarily varying channels with general alphabets and states,” IEEE Trans. Inform. Theory, vol. 38, pp. 1725–1742, Nov. 1992.
[6] A. Dembo and O. Zeitouni, Large Deviations Techniques and Applications, 2nd ed. Springer, 1998.
[7] R. M. Dudley, Real Analysis and Probability, 2nd ed. Cambridge University Press, 2003.
[8] M. Fozudal, S. W. McLaughlin, and R. W. Schaefer, “Capacity analysis for continuous-alphabet channels with side information, part I: A general framework,” IEEE Trans. Inform. Theory, vol. 51, no. 9, pp. 3075 – 3085, Sept. 2005.
[9] S. I. Gel’fand and M. S. Pinsker, “Coding for channels with random parameters,” Probl. Contr. and Inform. Theory, vol. 9, no. 1, pp. 19–31, 1980.
[10] R. M. Gray, Probability, random processes, and ergodic properties. New York: Springer-Verlag, 1988, revised 16 July 2009 edition available electronically at http://www-ee.stanford.edu/~gray/arp.html.
[11] T. S. Han, *Information-Spectrum Methods in Information Theory*, ser. Stochastic Modelling and Applied Probability. Springer, 2003, vol. 50.

[12] J. C. Kieffer, “Block coding for weakly continuous channels,” *IEEE Trans. Inform. Theory*, vol. IT-27, no. 6, pp. 721 – 727, Nov. 1981.

[13] ——, “Sliding-block coding for weakly continuous channels,” *IEEE Trans. Inform. Theory*, vol. IT-28, no. 1, pp. 2 – 10, Jan. 1982.

[14] Y.-H. Kim, A. Sutivong, and S. Sigurjonsso, “Multiple user writing on dirty paper,” in *Proc. IEEE Int. Symp. Inform. Theory*, Chicago, IL, June 2004, p. 534.

[15] A. Klenke, *Probability Theory: A Comprehensive Course*. Springer, 2008.

[16] G. Kramer, “Topics on multi-user information theory,” *Foundations and Trends in Communications and Information Theory*, vol. 4, no. 4-5, pp. 265 – 444, 2007.

[17] H. Schwarte, “Approaching capacity of a continuous channel by discrete input distributions,” *IEEE Trans. Inform. Theory*, vol. 42, no. 2, pp. 671 – 675, Mar. 1996.

[18] ——, “On weak covergence of probability measures, channel capacity and code error probabilities,” *IEEE Trans. Inform. Theory*, vol. 42, no. 5, pp. 1549 – 1551, Sept. 1996.

[19] S. Verdú and T. S. Han, “A general formula for channel capacity,” *IEEE Trans. Inform. Theory*, vol. 40, no. 4, pp. 1147 – 1157, July 1994.

[20] W. Yu, A. Sutivong, D. Julian, T. M. Cover, and M. Chiang, “Writing on colored paper,” in *Proc. IEEE Int. Symp. Inform. Theory*, Washington, DC, June 2001, p. 302.

[21] R. Zamir, S. Shamai, and U. Erez, “Nested linear/lattice codes for structured multiterminal binning,” *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1250–1276, June 2002.