Empirical Processes and Typical Sequences

Maxim Raginsky
ECE Department, Duke University
Durham, NC 27708, USA
Email: m.raginsky@duke.edu

***Abstract*—This paper proposes a new notion of a typical sequence over an abstract alphabet based on approximation of memoryless sources by empirical distributions, uniformly over a class of measurable "test functions." In the finite-alphabet case, we can take all uniformly bounded functions and recover the usual notion of typicality under total variation distance. For a general alphabet, this function class is too large, and must be restricted. We develop our notion of typicality with respect to any Glivenko–Cantelli function class (which admits a Uniform Law of Large Numbers) and demonstrate its power by deriving fundamental limits on achievable rates in several settings that can be reduced to uniform approximation of general-alphabet memoryless sources with respect to a suitable function class.**

I. Introduction

The notion of *typical sequence* has been central to information theory since Shannon's original paper [1]. For finite alphabets, it leads to simple and intuitive proofs of achievability in a wide variety of source and channel coding settings, including multiterminal scenarios [2]. Another appealing aspect of typical sequences is that they provide a language for *approximation* of information sources in total variation distance using finite communication resources. Recent work of Cuff et al. [3] on coordination via communication serves as a particularly striking example of the power of this language.

For abstract alphabets, however, most of this power is lost; while such results as the asymptotic equipartition property carry over [4], in most other situations, particularly involving lossy codes, one has to resort to ergodic theory [5] or large deviations theory [6]. Direct approximation of abstract memoryless sources in total variation using empirical distributions is, in general, impossible (cf. Sec. III for details). However, it is precisely this direct approximation that renders typicality-based proofs of achievability so transparent.

The present paper is an attempt to revise the notion of typicality for *general* alphabets in a way that would allow for similarly transparent proofs of achievability in a variety of settings. When two probability measures are close in total variation, the corresponding expectations of *any* bounded measurable function are also close. For general alphabets, when one of the measures is discrete, this is too much to ask. Instead, we advocate an approach based on suitably *restricting* the class of functions on which we would like to match statistical expectations with sample (empirical) averages. Provided the Law of Large Numbers holds *uniformly* over the restricted function class, we can speak of typical sequences *with respect to this class* and develop typicality-based achievability arguments in close parallel to the finite-alphabet case. The central object of study is the *empirical process* [7]–[9] indexed by the function class, which gives information on the deviation of empirical means from statistical means for a given realization of the source under consideration, and the total variation distance is replaced by the supremum norm of this empirical process.

After collecting the preliminaries, notation, and basic definitions in Sec. II, we motivate and describe our approach to typicality in Sec. III, where we also present a key lemma on the preservation of empirical-process typicality in a Markov structure. Next, in Sec. IV, using this lemma as the main technical tool, we illustrate the power of the proposed new approach by proving three theorems concerning fundamental limits on minimal achievable rates for (i) two-node empirical coordination; (ii) lossy source coding under a family of distortion measures; and (iii) rate-constrained distributed approximation of empirical processes with side information at the decoder. Although these results apply to general (uncountably infinite) alphabets, the proofs are as intuitive and simple as in the finite-alphabet scenario. We follow up with some concluding remarks in Sec. V.

II. Preliminaries and notation

All spaces in this paper are assumed to be standard Borel spaces, i.e., Borel subsets of a complete separable metric space, and will be equipped with their Borel σ-algebras. Let $(\mathsf{Z}, \mathcal{B}_\mathsf{Z})$ be such a Borel space, and let $\mathcal{M}(\mathsf{Z})$ denote the space of all probability measures on it. For any $P \in \mathcal{M}(\mathsf{Z})$ and a P-integrable function $f : \mathsf{Z} \to \mathbb{R}$, we will let $P(f)$ denote the expectation of f w.r.t. P, i.e.,

$$P(f) \triangleq \int_\mathsf{Z} f dP \equiv \mathbb{E}_{Z\sim P}[f(Z)].$$

Given a class $\mathcal{F}$ of measurable functions $f : \mathsf{Z} \to [-1, 1]$, we can define a pseudometric on $\mathcal{M}(\mathsf{Z})$ via

$$\|P - P'\|_\mathcal{F} \triangleq \sup_{f\in\mathcal{F}} |P(f) - P'(f)|.$$

For example, when $\mathcal{F}$ consists of *all* measurable functions $f : \mathsf{Z} \to [-1, 1]$, then it can be shown that $\|P - P'\|_\mathcal{F}$ is precisely the *total variation distance*

$$\|P - P'\|_{\mathrm{TV}} \triangleq 2 \sup_{A\in\mathcal{B}_\mathsf{Z}} |P(A) - P'(A)|. \tag{1}$$

The *divergence* between P and P' in $\mathcal{M}(\mathsf{Z})$ is defined as

$$D(P\|P') \triangleq \begin{cases} P\left(\log(dP/dP')\right), & \text{if } P \ll P' \\ +\infty, & \text{otherwise} \end{cases}$$

If Z has a product structure, e.g., $\mathsf{Z} = \mathsf{X} \times \mathsf{Y}$, then for any $Q \in \mathcal{M}(\mathsf{Z})$ we will denote by Q_X and Q_Y the corresponding marginal distributions in $\mathcal{M}(\mathsf{X})$ and $\mathcal{M}(\mathsf{Y})$. The (regular) conditional probability distribution of Y given X will be denoted by $Q_{Y|X}$. The *mutual information* between $X \in \mathsf{X}$ and $Y \in \mathsf{Y}$ with joint law Q is $I(Q) \triangleq D(Q\|Q_X \otimes Q_Y)$, where $Q_X \otimes Q_Y$ is the product of the marginals. Whenever Q is clear from context, we will write $I(X;Y)$ instead of $I(Q)$.

Given $z^n = (z_1, \ldots, z_n) \in \mathsf{Z}^n$, we will denote by P_{z^n} the induced *empirical measure*: $\mathsf{P}_{z^n} \triangleq n^{-1} \sum_{i=1}^n \delta_{z_i}$, where $\delta_{(\cdot)}$ is the Dirac measure. If $f : \mathsf{Z} \to [-1, 1]$ is a measurable function[1] and $\{Z_i\}_{i=1}^\infty$ is an i.i.d. sequence with common distribution $P \in \mathcal{M}(\mathsf{Z})$, then the Strong Law of Large Numbers says that the empirical mean $\mathsf{P}_{Z^n}(f) = n^{-1} \sum_{i=1}^n f(Z_i)$ converges to the true mean $P(f)$ almost surely. By the union bound, this holds for any *finite* family of functions. We can also consider *infinite* function classes that admit a Uniform Law of Large Numbers — that is, absolute deviations between empirical and true means converge to zero *uniformly* over the function class. This is known as the *Glivenko–Cantelli property* [7]–[9]:

Definition 1. *A class $\mathcal{F}$ of measurable functions $f : \mathsf{Z} \to [-1, 1]$ is called* Glivenko–Cantelli *(or GC, for short) if*

$$\|\mathsf{P}_{Z^n} - P\|_{\mathcal{F}} \xrightarrow{n\to\infty} 0 \qquad \textit{in probability} \tag{2}$$

for every $P \in \mathcal{M}(\mathsf{Z})$, where $\{Z_i\}_{i=1}^\infty$ is an i.i.d. random process with marginal distribution P.

Remark 1. If $\mathcal{F}$ is a GC class, the convergence in probability in (2) is equivalent to almost sure convergence [8].

Remark 2. When the class $\mathcal{F}$ is uncountable, $\|\mathsf{P}_{Z^n} - P\|_{\mathcal{F}}$ may not be a random variable (cf. [8]). In this case, one can either use the theory of outer measures and outer expectations, as in [8], or assume that $\mathcal{F}$ is "permissible" in the sense of Pollard [7]. We will sweep these issues under the rug.

Examples of GC classes include: all uniformly bounded, monotone functions on a compact subset of $\mathbb{R}$; all functions in a Sobolev ball on a compact subset of $\mathbb{R}^d$; all indicators of a Vapnik–Chervonenkis class of subsets of $\mathbb{R}^d$ [9].

Let $(\Omega, \mathcal{B}, \mathbb{P})$ be an underlying probability space for the random process $\{Z_i\}$. Then for each n we can construct another random process indexed by the elements of $\mathcal{F}$ via

$$\Delta_f^{(n)}(\omega) \triangleq \mathsf{P}_{Z^n(\omega)}(f) - P(f), \qquad f \in \mathcal{F}.$$

A random process of this type is called an *empirical process* [7]–[9]. Thus, a GC class is one for which the $\ell^\infty(\mathcal{F})$ norms $\left\|\Delta_f^{(n)}(\omega)\right\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \left|\Delta_f^{(n)}(\omega)\right|$ of the empirical processes $\{\Delta_f^{(n)}\}_{f\in\mathcal{F}}$, $n \geq 1$, converge to zero almost surely.

III. Rethinking typicality for general alphabets

There are many ways to define typicality, such as [3]:

[1]The restriction of the range of f to $[-1, 1]$ is not essential, but will be maintained in the sequel for the sake of convenience.

Definition 2. *Given a finite set Z and a probability distribution (mass function) P on it, the* typical set $\mathcal{T}_\varepsilon^{(n)}(P)$*, for $\varepsilon > 0$, is the set of all n-tuples $z^n \in \mathsf{Z}^n$ whose empirical distributions P_{z^n} are ε-close to P in total variation:*

$$\mathcal{T}_\varepsilon^{(n)}(P) \triangleq \{z^n \in \mathsf{Z}^n : \|\mathsf{P}_{z^n} - P\|_{\mathrm{TV}} < \varepsilon\}.$$

By the Law of Large Numbers, if $\{Z_i\}$ is a sequence of i.i.d. draws from P, then $\Pr\left(Z^n \notin \mathcal{T}_\varepsilon^{(n)}(P)\right) \to 0$ as $n \to \infty$. If Z is a Cartesian product $\mathsf{X} \times \mathsf{Y}$, then one can define *jointly* and *conditionally* typical sets and sequences [2].

However, all of this breaks down for general (uncountably infinite) alphabets. The reason is that the total variation distance between any discrete measure and a nonatomic measure is equal to 2. Indeed, if $(\mathsf{Z}, \mathcal{B}_\mathsf{Z})$ is a Borel space and $P \in \mathcal{M}(\mathsf{Z})$ assigns zero mass to singletons, $P(\{z\}) = 0, \forall z \in \mathsf{Z}$, then we can take any n-tuple $z^n \in \mathsf{Z}^n$ and let A be the set of its *distinct* elements, so that $\mathsf{P}_{z^n}(A) = 1$ and $P(A) = 0$. Using this and the definition (1), we deduce that $\|\mathsf{P}_{z^n} - P\|_{\mathrm{TV}} = 2$.

Of course, one could use typicality arguments by considering arbitrary finite quantizations of the underlying spaces, but, as long as we are dealing with nonatomic measures, this does not get rid of the above issue even in the limit of increasingly fine quantizations. While discretization is sufficient for many purposes [5], there is another issue that arises when dealing with Markov structures in multiterminal settings: quantization destroys the Markov property (cf. [10], Sec. VIII).

To resolve this conundrum, we recall (cf. Sec. II) that

$$\|P - P'\|_{\mathrm{TV}} = \sup_{|f|\leq 1} |P(f) - P'(f)|,$$

where the supremum is over *all* measurable functions $f : \mathsf{Z} \to [-1, 1]$. When the underlying measurable space supports nonatomic probability measures, this function class turns out to be too large to support uniform convergence of empirical averages to statistical expectations. Thus, we can either consider a weaker topology on the space of probability measures (e.g., the weak topology, which is amenable to large deviations techniques [11]), or we can restrict the class of functions. In this paper, we advocate the latter approach:

Definition 3. *Let $(\mathsf{Z}, \mathcal{B}_\mathsf{Z})$ be a Borel space and let $\mathcal{F}$ be a GC class of functions $f : \mathsf{Z} \to [-1, 1]$. Given a probability measure $P \in \mathcal{M}(\mathsf{Z})$, the* typical set $\mathcal{T}_{\varepsilon,\mathcal{F}}^{(n)}(P)$*, for $\varepsilon > 0$, is the set of all n-tuples $z^n \in \mathsf{Z}^n$ whose empirical distributions P_{z^n} are ε-close to P in the $\|\cdot\|_{\mathcal{F}}$ pseudometric:*

$$\mathcal{T}_{\varepsilon,\mathcal{F}}^{(n)}(P) \triangleq \{z^n \in \mathsf{Z}^n : \|\mathsf{P}_{z^n} - P\|_{\mathcal{F}} < \varepsilon\}.$$

By definition, $\Pr(Z^n \notin \mathcal{T}_{\varepsilon,\mathcal{F}}^{(n)}(P)) \to 0$ as $n \to \infty$. Moreover, GC typical sets share a key property with the total variation typical sets: empirical process typicality is preserved in a Markov structure. This is stated precisely in the following lemma, in the spirit of the so-called Piggyback Coding Lemma of Wyner [12, Lemma 4.3]. We omit the proof, which can be found in [13, Appendix].

Lemma 1. *Let $U \in \mathsf{U}$, $V \in \mathsf{V}$, and $W \in \mathsf{W}$ be random variables taking values in their respective Borel spaces according to a joint distribution P_{UVW}, such that $U \to V \to W$ is a Markov chain. Let $\{(U_i, V_i, W_i)\}_{i=1}^{\infty}$ be a sequence of i.i.d. draws from P_{UVW}. Let $\mathcal{F}$ be a GC class of measurable functions $f : \mathsf{U} \times \mathsf{W} \to [-1, 1]$. For a given $\varepsilon > 0$, there exists an $n = n(\varepsilon)$ and a mapping $\Phi_n : \mathsf{V}^n \to \mathsf{W}^n$, such that*

$$\log |\{\Phi_n(v^n) : v^n \in \mathsf{V}^n\}| \le n[I(V;W) + \varepsilon], \tag{3}$$

$$\Pr\Big((U^n, \Phi_n(V^n)) \notin \mathcal{T}^{(n)}_{\varepsilon,\mathcal{F}}(P_{UW})\Big) < \varepsilon. \tag{4}$$

IV. Examples of applications

We now show three sample applications of GC typicality. A useful feature of this approach is that the GC property is easy to check, and the list of known GC classes is catalogued fairly extensively [7]–[9]. With a judicious choice of $\mathcal{F}$ for a given application, we can develop particularly intuitive achievability arguments; moreover, convexity of the $\|\cdot\|_{\mathcal{F}}$ pseudometric is helpful for proving converse results.

A. Empirical coordination

The problem of *coordination via communication*, recently formulated and studied by Cuff et al. [3], concerns joint generation of actions at the nodes of a network, such that the empirical distribution of the actions over time approximates, asymptotically, a desired joint distribution in total variation. We would like to extend this setting to general alphabets.

Let us consider a two-node network, where Node I (resp., Node II) generates actions from a Borel space X (resp., Y). At Node I, the actions are drawn i.i.d. from a fixed law $P_X \in \mathcal{M}(\mathsf{X})$. We also have a regular conditional probability measure $P_{Y|X}$ that describes the desired distribution of actions at Node II given the actions at Node I. Following the terminology of [3], we will also refer to the choice of $P_{Y|X}$ as a *coordination*. Node I can communicate with Node II over a rate-limited channel, and Node II uses the data it receives to choose its actions. For each n, let X^n and $\hat{Y}^n$ denote the action sequences at the two nodes. Given a class $\mathcal{F}$ of measurable "test fuctions" $f : \mathsf{X} \times \mathsf{Y} \to [-1, 1]$ and a desired distortion level $\Delta \ge 0$, the goal is for Node I to communicate with Node II at a minimal rate to guarantee that, asymptotically,

$$\mathbb{E}\|\mathsf{P}_{(X^n,\hat{Y}^n)} - P_X \otimes P_{Y|X}\|_{\mathcal{F}} \lesssim \Delta,$$

where $P_{XY} = P_X \otimes P_{Y|X}$ is the joint law induced by the source P_X and the coordination $P_{Y|X}$.

Definition 4. *An (n, M)-*code *is a pair (e_n, d_n), where $e_n : \mathsf{X}^n \to [M]$ is the* encoder *and $d_n : [M] \to \mathsf{Y}^n$ is the* decoder*, and $[M] \triangleq \{1, 2, \ldots, M\}$. We will denote $\hat{Y}^n = d_n(e_n(X^n))$.*

Definition 5. *Given a source P_X, a coordination $P_{Y|X}$, and a distortion Δ, let $\mathcal{E}(\Delta, P_{Y|X})$ denote the set of all $Q \in \mathcal{M}(\mathsf{X}\times\mathsf{Y})$, such that $Q_X = P_X$ and $\|Q - P_X \otimes P_{Y|X}\|_{\mathcal{F}} \le \Delta$. Define the* rate-distortion/coordination function *of P_X:*

$$R(\Delta, P_{Y|X}) \triangleq \inf_{Q \in \mathcal{E}(\Delta, P_{Y|X})} I(Q).$$

Theorem 1. *Let $P_{Y|X}$ be a given coordination and Δ a given distortion level.*

a) **Direct part:** *If $\mathcal{F}$ is a GC class and $R(\Delta, P_{Y|X}) < \infty$, then for any $\varepsilon > 0$ there exist $n \equiv n(\varepsilon)$ and an $(n, 2^{nR})$ code (e_n, d_n) with $R < R(\Delta, P_{Y|X}) + \varepsilon$ satisfying*

$$\mathbb{E}\left\|\mathsf{P}_{(X^n,\hat{Y}^n)} - P_X \otimes P_{Y|X}\right\|_{\mathcal{F}} \le \Delta + \varepsilon. \tag{5}$$

b) **Converse part:** *Suppose that there exists an $(n, 2^{nR})$-code $\hat{Y}^n(X^n) = d_n(e_n(X^n))$ satisfying*

$$\mathbb{E}\left\|\mathsf{P}_{(X^n,\hat{Y}^n)} - P_X \otimes P_{Y|X}\right\|_{\mathcal{F}} \le \Delta. \tag{6}$$

Then $R \ge R(\Delta, P_{Y|X})$.

Proof: To prove the direct part, fix $(\Delta, P_{Y|X})$ and pick any $Q \in \mathcal{E}(\Delta, P_{Y|X})$ such that $I(Q) < R(\Delta, P_{Y|X}) + \varepsilon/2$. Let $X \in \mathsf{X}$ and $U \in \mathsf{Y}$ have joint law Q. Then $X \to X \to U$ form a Markov chain, and Lemma 1 guarantees the existence of an n and a mapping $\Phi_n : \mathsf{X}^n \to \mathsf{Y}^n$, such that

$$n^{-1} \log |\{\Phi_n(\mathsf{X}^n)\}| \le I(Q) + \varepsilon/2 < R(\Delta, P_{Y|X}) + \varepsilon$$

and

$$\mathbb{E}\left\|\mathsf{P}_{(X^n,\Phi_n(X^n))} - Q\right\|_{\mathcal{F}} \le \varepsilon.$$

Let $\hat{Y}^n = \Phi_n(X^n)$. Then the triangle inequality gives

$$\begin{aligned}\mathbb{E}\left\|\mathsf{P}_{(X^n,\hat{Y}^n)} - P_{XY}\right\|_{\mathcal{F}} &\\ \le \mathbb{E}\left\|\mathsf{P}_{(X^n,\hat{Y}^n)} - Q\right\|_{\mathcal{F}} &+ \left\|Q - P_{XY}\right\|_{\mathcal{F}} \le \Delta + \varepsilon,\end{aligned}$$

which establishes (5).

For the converse, we will use the time mixing technique (cf. [3]). Let $\hat{Y}^n(X^n)$ be an $(n, 2^{nR})$-code such that (6) holds. Let T be a random variable uniformly distributed over the set $[n]$, independently of X^n, and let $\hat{Q}$ denote the joint distribution of $(X_T, \hat{Y}_T)$. A standard argument (cf. proof of the converse part of Theorem 3 in [3]) can be used to show that $R \ge I(\hat{Q})$. Since the X_i's are i.i.d., X_T is independent of T and has the same distribution as X_1, namely P_X. Moreover, the expected empirical distribution $\mathbb{E}\mathsf{P}_{(X^n,\hat{Y}^n)}$ is equal to $\hat{Q}$. Thus, we can write

$$\begin{aligned}\|\hat{Q} - P_X \otimes P_{Y|X}\|_{\mathcal{F}} &= \|\mathbb{E}\mathsf{P}_{(X^n,\hat{Y}^n)} - P_{X\otimes Y}\|_{\mathcal{F}} \\ &\le \mathbb{E}\|\mathsf{P}_{(X^n,\hat{Y}^n)} - P_X \otimes P_{Y|X}\|_{\mathcal{F}} \le \Delta.\end{aligned}$$

Hence, $\hat{Q} \in \mathcal{E}(\Delta, P_{Y|X})$, so $R \ge I(\hat{Q}) \ge R(\Delta, P_{Y|X})$. ∎

B. Lossy coding with respect to a class of distortion measures

Next, we consider the problem of lossy coding with respect to a class of distortion measures (fidelity criteria). For general (Polish) alphabets, it was solved by Dembo and Weissman [14], but the finite-alphabet variant appears already as Problem 14 in [15]. Let X and Y denote the source and the reproduction alphabets, respectively. Suppose a class Γ of distortion measures $\rho : \mathsf{X} \times \mathsf{Y} \to [0, 1]$ is given, together with a class of nonnegative reals indexed by $\rho \in \Gamma$, $\{\Delta_\rho\}_{\rho\in\Gamma}$. The goal is to find a block code of minimal rate whose expected distortion under each $\rho \in \Gamma$ is bounded by the corresponding

Δ_ρ. We use the same definition of an (n, M)-code as in the preceding section.

Define a mapping $F(\cdot, \{\Delta_\rho\}) : \mathcal{M}(\mathsf{X} \times \mathsf{Y}) \to \mathbb{R}$ by

$$F(Q, \{\Delta_\rho\}) \triangleq \sup_{\rho \in \Gamma} [Q(\rho) - \Delta_\rho],$$

where $Q(\rho) = \int \rho dQ = \int \rho(x, y) Q(dx, dy)$ is the expected distortion between X and Y when they have joint law Q.

Definition 6. *Given a source $P_X \in \mathcal{M}(\mathsf{X})$, let $\mathcal{E}(\{\Delta_\rho\})$ denote the set of all $Q \in \mathcal{M}(\mathsf{X} \times \mathsf{Y})$ such that $Q_X = P_X$ and $F(Q, \{\Delta_\rho\}) \leq 0$. Define the* rate-distortion function

$$R(\{\Delta_\rho\}) \triangleq \inf_{Q \in \mathcal{E}(\{\Delta_\rho\})} I(Q).$$

Theorem 1 of [14] shows that any rate $R \geq R(\{\Delta_\rho\})$ is achievable, provided the mapping $Q \mapsto F(Q, \{\Delta_\rho\})$ is upper semicontinuous (u.s.c.) under the weak topology on $\mathcal{M}(\mathsf{X} \times \mathsf{Y})$. Moreover, no rate $R < R(\{\Delta_\rho\})$ is achievable. We now show that the u.s.c. requirement can be replaced by a GC condition:

Theorem 2. *Let Γ be a class of distortion measures and $\{\Delta_\rho\}_{\rho \in \Gamma}$ a class of nonnegative distortion levels.*

a) **Direct part:** *If Γ is a GC class and $R(\{\Delta_\rho\}) < \infty$, then for any $\varepsilon > 0$, there exist an $n \equiv n(\varepsilon)$ and an $(n, 2^{nR})$ code with $R < R(\{\Delta_\rho\}) + \varepsilon$ satisfying*

$$\mathbb{E} \sup_{\rho \in \Gamma} \Big[\rho(X^n, \hat{Y}^n) - \Delta_\rho \Big] \leq \varepsilon, \tag{7}$$

where $\rho(X^n, \hat{Y}^n) \triangleq \mathsf{P}_{(X^n, \hat{Y}^n)}(\rho)$.

b) **Converse part:** *Suppose that there exists an $(n, 2^{nR})$-code $\hat{Y}^n = d_n(e_n(X^n))$ satisfying*

$$\mathbb{E}\rho(X^n, \hat{Y}^n) \leq \Delta_\rho, \qquad \forall \rho \in \Gamma. \tag{8}$$

Then $R \geq R(\{\Delta_\rho\})$.

Proof: To prove the direct part, pick any $Q \in \mathcal{E}(\{\Delta_\rho\})$ such that $I(Q) < R(\{\Delta_\rho\}) + \varepsilon/2$. Let $X \in \mathsf{X}$ and $U \in \mathsf{Y}$ have joint law Q. The same argument as in the proof of Theorem 1 can be used to show the existence of a large enough n and a mapping $\Phi_n : \mathsf{X}^n \to \mathsf{Y}^n$, such that

$$n^{-1} \log |\{\Phi_n(\mathsf{X}^n)\}| \leq I(Q) + \varepsilon/2 < R(\{\Delta_\rho\}) + \varepsilon$$

and

$$\mathbb{E} \big\| \mathsf{P}_{(X^n, \hat{Y}^n)} - Q \big\|_\Gamma \leq \varepsilon,$$

where $\hat{Y}^n = \Phi_n(X^n)$. Now, for any $\rho \in \Gamma$ we have

$$\rho(X^n, \hat{Y}^n) - \Delta_\rho \leq \| \mathsf{P}_{(X^n, \hat{Y}^n)} - Q \|_\Gamma + F(Q, \{\Delta_\rho\}).$$

Consequently, taking the supremum of both sides over Γ and then the expectation w.r.t. P_{X^n}, we get (7).

The proof of the converse is exactly the same as in [14]. ■

C. Distributed compression of empirical processes

The last application we consider also concerns distributed approximation of an empirical process. We have a joint law $P = P_{XY} \in \mathcal{M}(\mathsf{X} \times \mathsf{Y})$. Let $\{(X_i, Y_i)\}_{i=1}^\infty$ be an infinite sequence of independent draws from P. Consider a two-node network, just as in Sec. IV-A, except now Node I (resp., Node II) has perfect observations of $\{X_i\}$ (resp., $\{Y_i\}$). Node I can transmit information to Node II over a rate-limited channel. The goal is for Node I to communicate with Node II at a minimal rate, so that Node II can approximate the desired empirical process to within a given distortion level Δ. More precisely, given a block length n and denoting by $\hat{X}^n$ the reconstruction of X^n at Node II, we wish to guarantee that

$$\mathbb{E} \| \mathsf{P}_{(\hat{X}^n, Y^n)} - P \|_\mathcal{F} \lesssim \Delta.$$

This setting generalizes the problem of *communication of probability distributions*, recently formulated and studied by Kramer and Savari [16]. Here, we allow general alphabets and decoder side information. As we will see, the minimum achievable rate admits a single-letter characterization reminiscent of the Wyner–Ziv rate-distortion function for lossy source coding with decoder side information [17], [18].

Definition 7. *An (n, M)-*code *is a pair (e_n, d_n), where $e_n : \mathsf{X}^n \to [M]$ is the* encoder *and $d_n : [M] \times \mathsf{Y}^n \to \mathsf{X}^n$ is the* decoder. *We will denote $\hat{X}^n = d_n(e_n(X^n), Y^n)$.*

Definition 8. *Given a source $P_{XY} \in \mathcal{M}(\mathsf{X} \times \mathsf{Y})$, let $\mathcal{E}(\Delta)$ denote the set of all $Q \in \mathcal{M}(\mathsf{X} \times \mathsf{Y} \times \mathsf{U})$, where U is an arbitrary Borel space, such that:*

1) *$Q_{XY} = P_{XY}$*
2) *$Q_{U|XY} = Q_{U|X}$ (i.e., $Y \to X \to U$ is a Markov chain)*
3) *There is a function $g : \mathsf{Y} \times \mathsf{U} \to \mathsf{X}$, such that*

$$\tilde{Q}_{WY}(f) - P(f) \leq \Delta, \qquad \forall f \in \mathcal{F}$$

where $W = g(Y, U)$ and $\tilde{Q}$ is the induced joint law of (X, W, Y) when $(X, Y, U) \sim Q$.

With this, define the rate-distortion function

$$R(\Delta) \triangleq \inf_{Q \in \mathcal{E}(\Delta)} [I(Q_{XU}) - I(Q_{YU})].$$

Theorem 3. *Let $\mathcal{F}$ be a class of functions $f : \mathsf{X} \times \mathsf{Y} \to [-1, 1]$ and Δ a nonnegative distortion level.*

a) **Direct part:** *Suppose that $\mathcal{F}$ is a GC class, and that, for any $\varepsilon > 0, \mu \in \mathcal{M}(\mathsf{X} \times \mathsf{Y})$, one can find a finite set $\{\hat{x}_j\}_{j=1}^N \subset \mathsf{X}$ and a* quantizer *$q : \mathsf{X} \to \{\hat{x}_j\}$, such that*

$$\mathbb{E}_\mu[f(q(X), Y)] \leq (1 + \varepsilon) \mathbb{E}_\mu[f(X, Y)], \forall f \in \mathcal{F}. \tag{9}$$

If $R(\Delta) < \infty$, then for any $\varepsilon > 0$ there exist an $n \equiv n(\varepsilon)$ and an $(n, 2^{nR})$ code with $R < R(\Delta) + \varepsilon$ satisfying

$$\mathbb{E} \big\| \mathsf{P}_{(\hat{X}^n, Y^n)} - P \big\|_\mathcal{F} \leq \Delta + \varepsilon, \tag{10}$$

where $\hat{X}^n = d_n(e_n(X^n), Y^n)$.

b) **Converse part:** *Suppose that there exists an $(n, 2^{nR})$-code $\hat{X}^n = d_n(e_n(X^n), Y^n)$ satisfying*

$$\mathbb{E} \big\| \mathsf{P}_{(\hat{X}^n, Y^n)} - P \big\|_\mathcal{F} \leq \Delta. \tag{11}$$

Then $R \geq R(\Delta)$.

Remark 3. The quantization assumption (9) is a "smoothness" condition on $\mathcal{F}$. A similar assumption was made by Wyner [18] in order to extend the (achievability part of) the finite-alphabet result of [17] to abstract alphabets.

Proof: For the direct part, pick any $Q \in \mathcal{E}(\Delta)$ such that $I(Q_{XU}) - I(Q_{YU}) < R(\Delta) + \varepsilon/2$. Using (9), we can develop an approximation argument similar to Lemma 5.3 of Wyner [18] to show that, for any $\delta > 0$, there exist finite measurable partitions $\{A_i\}_{i=1}^{N_1}$ and $\{B_j\}_{j=1}^{N_2}$ of Y and U and a function $g_1 : \mathsf{Y} \times \mathsf{X} \to \mathsf{X}$ such that:

a) $\mathbb{E}[f(g_1(Y,W),Y)] \leq \Delta + \delta, \forall f \in \mathcal{F}$
b) g_1 is constant on the rectangles $A_i \times B_j$, $1 \leq i \leq N_1, 1 \leq j \leq N_2$
c) $I(Q_{X\tilde{U}}) - I(Q_{\tilde{Y}\tilde{U}}) \leq I(Q_{XU}) - I(Q_{YU}) + \delta$ where $\tilde{Y} = i$ for $Y \in A_i$ and $\tilde{U} = j$ for $U \in B_j$.

Thus, without loss of generality we may assume that both Y and the auxiliary alphabet U are finite. Let $\tilde{Q}$ be the induced distribution of (X, W, Y), where $W = g(Y, U)$. Since $\mathcal{F}$ is a GC class, Lemma 1 guarantees the existence of a large enough n_1 and a mapping $\Phi_{n_1} : \mathsf{X}^{n_1} \to \mathsf{U}^{n_1}$, such that

$$n_1^{-1} \log |\{\Phi_{n_1}(\mathsf{X}^{n_1})\}| \leq I(Q_{XU}) + \varepsilon/2$$

and

$$\mathbb{E} \left\| \mathsf{P}_{(\hat{W}^n, Y^n)} - \tilde{Q}_{WY} \right\|_{\mathcal{F}} \leq \varepsilon/2,$$

where

$$\hat{W}^{n_1} = \big(g(Y_1, \hat{U}_1), \ldots, g(Y_{n_1}, \hat{U}_{n_1})\big) \text{ and } \hat{U}^{n_1} = \Phi_{n_1}(X^{n_1}).$$

We can use a blocking argument along the lines of Lemmas 3 and 5 of Wyner and Ziv [17] to show that a sufficiently long sequence $\hat{U}^{n_1}(1), \ldots, \hat{U}^{n_1}(n_2)$ of i.i.d. realizations of $\hat{U}^{n_1}$ can be losslessly encoded, using a Slepian–Wolf code, at a rate of

$$n_1^{-1} H\big(\hat{U}^{n_1} \big| Y^{n_1}\big) \leq I(Q_{XU}) - I(Q_{YU}) + \varepsilon/2 < R(\Delta) + \varepsilon.$$

Let $n = n_1 n_2$, and let $\{\tilde{U}_i\}_{i=1}^n$ denote the resulting decoding. Then, if n_2 is large enough, we can guarantee that

$$\mathbb{E} \left\| \mathsf{P}_{(\hat{X}^n, Y^n)} - \tilde{Q}_{WY} \right\|_{\mathcal{F}} \leq \varepsilon/2,$$

where $\hat{X}^n = \big(g(Y_1, \tilde{U}_1), \ldots, g(Y_n, \tilde{U}_n)\big)$. The triangle inequality then yields

$$\begin{aligned} &\mathbb{E} \left\| \mathsf{P}_{(\hat{X}^n, Y^n)} - P \right\|_{\mathcal{F}} \\ &\quad \leq \mathbb{E} \left\| \mathsf{P}_{(\hat{X}^n, Y^n)} - \tilde{Q}_{WY} \right\|_{\mathcal{F}} + \left\| \tilde{Q}_{WY} - P \right\|_{\mathcal{F}} \leq \Delta + \varepsilon. \end{aligned}$$

This gives us the desired $(n, 2^{nR})$-code with $R < R(\Delta) + \varepsilon$.

To prove the converse, we again use time mixing. Let (e_n, d_n) be an $(n, 2^{nR})$ code, let $J = e_n(X^n)$ and $\hat{X}^n = d_n(J, Y^n)$, and let T be uniformly distributed on $[n]$ independently of X^n. Define an auxiliary random variable $U = (J, X^{T-1}, Y^{T-1}, Y_{T+1}^n, T)$ (cf. [3], [17], [18]) and note that $Y_T \to X_T \to U$ is a Markov chain. Following the same steps as in the proof of the Wyner–Ziv converse, we can show $R \geq I(X_T; U | Y_T)$. Since $\{(X_i, Y_i)\}_{i=1}^n$ are i.i.d., (X_T, Y_T) has the same joint law as (X_1, Y_1), namely P_{XY}. Moreover, $\hat{X}_T$ is a deterministic function of (Y_T, U), and $\mathbb{E}\mathsf{P}_{(\hat{X}^n, Y^n)} = P_{(\hat{X}_T, Y_T)}$. By convexity,

$$\big\| P_{(\hat{X}_T, Y_T)} - P \big\|_{\mathcal{F}} \leq \mathbb{E} \big\| \mathsf{P}_{(\hat{X}^n, Y^n)} - P \big\|_{\mathcal{F}} \leq \Delta,$$

which implies that $P_{(\hat{X}_T, Y_T)}(f) \leq P(f) + \Delta$ for all $f \in \mathcal{F}$. Hence, the joint law of X_T, Y_T, and U belongs to $\mathcal{E}(\Delta)$, which means that $R \geq I(X_T; U | Y_T) \geq R(\Delta)$. ∎

V. Conclusion

We have proposed a new notion of typical sequence over an abstract alphabet which retains many useful properties of total-variation typicality for finite alphabets. In the future, we plan to investigate the finite block length behavior using the $\Theta(\frac{1}{\sqrt{n}})$ behavior of empirical process supremum norms over sufficiently "regular" function classes [7]–[9], as well as the relationship between our approach and the information spectrum ideas of Verdú and co-workers [19], [20].

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.
[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley, 2006.
[3] P. Cuff, H. Permuter, and T. Cover, "Coordination capacity," *IEEE Trans. Inform. Theory*, 2009, submitted.
[4] A. R. Barron, "The strong ergodic theorem for densities: generalized Shannon–McMillan–Breiman theorem," *Ann. Probab.*, vol. 13, no. 4, pp. 1292–1303, 1985.
[5] R. M. Gray, *Entropy and Information Theory*. New York: Springer-Verlag, 1990.
[6] A. Dembo and O. Zeitouni, *Large Deviations: Techniques and Applications*. New York: Springer, 1998.
[7] D. Pollard, *Convergence of Stochastic Processes*. New York: Springer, 1984.
[8] A. W. van der Waart and J. A. Wellner, *Weak Convergence and Empirical Processes*. New York: Springer-Verlag, 1996.
[9] S. van de Geer, *Empirical Processes in M–Estimation*. Cambridge Univ. Press, 2000.
[10] I. Csiszár, "The method of types," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2505–2523, June 1998.
[11] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*. New York: Wiley, 1997.
[12] A. D. Wyner, "On source coding with side information at the decoder," *IEEE Trans. Inform. Theory*, vol. IT-21, no. 3, pp. 294–300, May 1975.
[13] M. Raginsky, "Achievability results for learning under communication constraints," in *Proc. Inform. Theory and Applications Workshop*, San Diego, CA, 2009, pp. 272–279.
[14] A. Dembo and T. Weissman, "The minimax distortion redundancy in noisy source coding," *IEEE Trans. Inform. Theory*, vol. 49, no. 11, pp. 3020–3030, November 2003.
[15] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Sources*. Budapest: Akadémiai Kiadó, 1981.
[16] G. Kramer and S. A. Savari, "Communicating probability distributions," *IEEE Trans. Inform. Theory*, vol. 53, no. 2, pp. 518–525, February 2007.
[17] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inform. Theory*, vol. IT-22, no. 1, pp. 1–10, January 1976.
[18] A. D. Wyner, "The rate-distortion function for source coding with side information at the decoder II: general sources," *Inform. Control*, vol. 38, pp. 60–80, 1978.
[19] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Inform. Theory*, vol. 39, no. 3, pp. 752–772, March 1993.
[20] Y. Steinberg and S. Verdú, "Simulation of random processes and rate-distortion theory," *IEEE Trans. Inform. Theory*, vol. 42, no. 1, pp. 63–86, January 1996.