Universal Simulation with Fidelity Criteria

Neri Merhav, Marcelo J. Weinberger
Advanced Studies
HP Laboratories Palo Alto
HPL-2007-148
September 7, 2007*

universal simulation, distance measures, generalized divergence

We consider the problem of universal simulation of a memoryless source (with some partial extensions to Markov sources), based on a training sequence emitted from the source. The objective is to maximize the conditional entropy of the simulated sequence given the training sequence, subject to a certain distance constraint between the probability distribution of the output sequence and the probability distribution of the input, training sequence. We derive, for several distance criteria, single–letter expressions for the maximum attainable conditional entropy as well as corresponding universal simulation schemes that asymptotically attain these maxima.

* Internal Accession Date Only

Approved for External Publication

Universal Simulation with Fidelity Criteria*

Neri Merhav†
Department of Electrical Engineering
Technion – I.I.T., Haifa 32000, Israel
Email: merhav@ee.technion.ac.il

Marcelo J. Weinberger
Hewlett–Packard Laboratories
Palo Alto, CA 94304, U.S.A.
Email: marcelo@hpl.hp.com

***Abstract*— We consider the problem of universal simulation of a memoryless source (with some partial extensions to Markov sources), based on a training sequence emitted from the source. The objective is to maximize the conditional entropy of the simulated sequence given the training sequence, subject to a certain distance constraint between the probability distribution of the output sequence and the probability distribution of the input, training sequence. We derive, for several distance criteria, single–letter expressions for the maximum attainable conditional entropy as well as corresponding universal simulation schemes that asymptotically attain these maxima.**

Index Terms: Universal simulation, distance measures, generalized divergence, $\bar{\rho}$–distance, ϵ–contaminated model.

I. Introduction

Simulation of a source means artificial production of random data with some probability law, by using a certain device that is fed by a sequence of purely random bits. Simulation of sources and channels is a problem that has been studied in a series of works, see, e.g., [7], [15], [16], [17] and references therein. In all these works, it was assumed that the probability law of the desired process is perfectly known.

More recently, a universal version of this problem was studied in [12], [13] (see also [10]), where the assumption of perfect knowledge of the target probability law was relaxed. Instead, the target source P to be simulated was assumed in [12] to belong to a certain parametric family $\mathcal{P}$, but is otherwise unknown, and a training sequence $X^m = (X_1, \ldots, X_m)$, that has emerged from this source, is available. In addition, the simulator is provided with a sequence of ℓ random key bits $U^\ell = (U_1, \ldots, U_\ell)$, which is independent of X^m. The goal of the simulation scheme in [12] was to generate an output sequence $Y^n = (Y_1, \ldots, Y_n)$, $n \leq m$, corresponding to the simulated process, such that $Y^n = \psi(X^m, U^\ell)$, where ψ is a deterministic function that does not depend on the unknown source P, and which satisfies the following two conditions: (i) the probability distribution of Y^n is *exactly* the n-dimensional marginal of the probability law P corresponding to X^m for all $P \in \mathcal{P}$, and (ii) the mutual information $I(X^m; Y^n)$ is as small as possible, or equivalently (under (i)), the conditional entropy $H(Y^n|X^m)$ is as large as possible, simultaneously for all $P \in \mathcal{P}$ (so as to make the generated sample path Y^n as "original" as possible). In [12], the smallest achievable value of the mutual information (or, the largest conditional entropy) was characterized, and simulation schemes that asymptotically achieve these bounds were presented (see also [13]). It turns out that for these optimal schemes, for ℓ large enough, the normalized mutual information asymptotically vanishes. In [11], the same simulation problem was studied in the regime of a delay–limited system, in which the simulator produces output samples on–line, as the training data is fed into the system sequentially. The cost of limited delay was characterized and a strictly optimum simulation system was proposed. A different perspective on universal simulation was investigated in [14], where x^m was assumed to be an individual sequence not originating from any probabilistic source.

In this work, we extend the scope of the universal simulation problem in another direction, namely, relaxing the requirement of *exact* preservation of the probability law at the output of the simulator. In particular, we study the best achievable tradeoff between the performance of the simulation scheme and the distance (measured in terms of a certain metric) between the probability law of the output and that of the input. Observe that when the probability law of the simulated sequence is not constrained to be identical to that of the training sequence, the criteria $\min I(X^m; Y^n)$ and $\max H(Y^n|X^m)$ are no longer equivalent. While the former criterion aims at weak dependency, it should be emphasized that, for a large enough key rate, *vanishing* normalized mutual information was shown to be achievable with exact preservation of the probability law [12]. Therefore, under the $\min I(X^m; Y^n)$ criterion, the main objective of a relaxation of this requirement is to save on the key rate necessary for the normalized mutual information to vanish, as studied in [13] in the context of the $\bar{\rho}$–distance between probability distributions.[1] On the other hand, the asymptotic performance as given by the $\max H(Y^n|X^m)$ criterion (as a measure of the "originality" or the "diversity" of the typical sample paths generated by the simulator), on which we focus in this paper, can potentially benefit from the proposed relaxation.

For the class of discrete memoryless sources (DMSs), we derive single-letter formulas for the maximum achievable conditional entropy subject to various distance constraints

* The material in this paper was presented in part at the 2006 IEEE International Symposium on Information Theory, Seattle, WA, USA, July 2006.

† This work was done while N. Merhav was visiting Hewlett–Packard Laboratories, Palo Alto, CA, U.S.A.

[1] In addition, it is conceivable that, by deviating from the input probability law, a faster vanishing rate for the normalized mutual information is possible. However, this aspect of the problem is not discussed in [13].

(corresponding to different distance functions) and propose corresponding simulation schemes that universally achieve these bounds for large m and n. Some of the results have extensions to more general families of sources, like the family of Markov sources of a given order. We point out that here we limit ourselves to focus only on optimum tradeoffs between maximum achievable asymptotic values of $H(Y^n|X^m)/n$ and the distance between the true source and the simulated source, without an attempt to characterize optimum convergence rates, and without taking into account key rate limitations, as opposed to [12] and [13]. The assumption that the simulator has access to an unlimited stream of random bits is consistent with the setting in [14] and [11].

The remainder of this paper is organized as follows: In Section II, we establish notation conventions and formulate the problem. The other sections of the paper are devoted to single–letter characterizations of optimum performance for various kinds of distance measures between probability distributions: In Section III, we investigate the tradeoff between the maximum conditional entropy and a distance function that is referred to as *generalized divergence*, and a few examples are worked out in full detail, as well as an outline of a possible extension to Markov sources. In Section IV, we focus on the $\bar{\rho}$–distance measure (see also [13]), and finally, in Section V, we consider the ϵ–contaminated model, a notion rooted in robust statistics [8],[9].

II. Notation and Problem Formulation

Throughout the paper, random variables will be denoted by capital letters, specific values they may take will be denoted by the corresponding lower case letters, and their alphabets, as well as some other sets, will be denoted by calligraphic letters. Similarly, random vectors, their realizations, and their alphabets, will be denoted, respectively, by capital letters, the corresponding lower case letters, and calligraphic letters, all superscripted by their dimensions. For example, the random vector $X^m = (X_1, \ldots, X_m)$, (m – positive integer) may take a specific vector value $x^m = (x_1, \ldots, x_m)$ in $\mathcal{A}^m$, the m-th order Cartesian power of $\mathcal{A}$, which is the alphabet of each component of this vector. For $i \leq j$ (i, j – integers), x_i^j will denote the segment $(x_i, \ldots, x_j)$, where for $i = 1$ the subscript will be omitted. For typographical convenience, $\mathbf{x}$ and $\mathbf{X}$ will sometimes be used as alternative notations for x^m and X^m, respectively. Similar conventions will apply to the vector y^n (n – positive integers), its boldface notation $\mathbf{y}$, the corresponding uppercase notations Y^n and $\mathbf{Y}$, their empirical distributions, etc.

Let $\mathcal{P}$ denote the class of all DMS's with a finite alphabet $\mathcal{A}$, and let P denote a particular member of $\mathcal{P}$. For a given positive integer m, let $X^m = (X_1, X_2, \ldots, X_m)$, $X_i \in \mathcal{A}$, $i = 1, \ldots, m$, denote an m-vector drawn from P, namely,

$$\Pr\{X_i = x_i,\ i = 1, \ldots, m\} = \prod_{i=1}^{m} P(x_i) \stackrel{\Delta}{=} P(x^m)$$

for every $(x_1, \ldots, x_m)$, $x_i \in \mathcal{A}$, $i = 1, \ldots, m$. Let

$$H \equiv H(X) = -\sum_{x \in \mathcal{A}} P(x) \log P(x)$$

denote the entropy of the source P, where here and throughout the sequel $\log(\cdot) \stackrel{\Delta}{=} \log_2(\cdot)$. When it is the dependence of the entropy upon P that we wish to emphasize (rather than the name of the random variable X), we denote the entropy by $H(P)$, with a slight abuse of notation. Generic probability distributions of vectors of n symbols in $\mathcal{A}$ will be denoted by Q^n, when we wish to emphasize the dependence on n. When Q^n is a memoryless source (e.g., $Q^n = P^n$), then the superscript n obtains the meaning of the n–th power of Q, namely, the n–fold product of Q with itself n times. With a slight abuse of notation, however, we will normally omit the superscript n when referring to the probability of a certain vector or event, relying on the fact that the dimensionality n of the vector (or the event in the space of n–vectors) will be clear from the argument of the probability function $Q(\cdot)$.

We will denote the type class of x^m (with respect to $\mathcal{P}$) by T_{x^m} (or $T_{\mathbf{x}}$), i.e., the set of all $\tilde{x}^m \in \mathcal{A}^m$ such that $P(\tilde{x}^m) = P(x^m)$ simultaneously for all DMS's in $\mathcal{P}$. In other words, T_{x^m} is the set of all permutations of x^m, or equivalently, the set of all sequences with the same empirical distribution, P_{x^m}.

For two positive integers m and n, let $W(y^n|x^m)$ denote the conditional probability of $Y^n = y^n$ given $X^m = x^m$ corresponding to the channel from X^m to Y^n that is induced by a simulation scheme, which has certain resources of randomness. Let $H(Y^n|X^m)$ denote the conditional entropy of Y^n given X^m induced by this channel. The expectation operator, denoted $\boldsymbol{E}\{\cdot\}$, will be understood to be taken with respect to the joint distribution $P \times W$ of (X^m, Y^n). As shown in [6], the *expected* number of key bits required to implement a channel W is approximately $H(Y^n|X^m)$, which can be achieved via arithmetic decoding. However, for some sample paths, the number of key bits required may be unlimited.

Let $\rho_n(P^n, Q^n)$ denote a distance function (not necessarily a metric) between two probability measures on $\mathcal{A}^n$, where P^n is defined as the product probability measure of n–vectors, whereas Q^n does not necessarily have product form.

This paper is about the quest for a channel $\{W(y^n|x^m),\ x^m \in \mathcal{A}^m,\ y^n \in \mathcal{A}^n\}$ that is independent of the unknown P generating X^m and that satisfies the following conditions:

C1. For every $P \in \mathcal{P}$, the probability distribution

$$Q^n(y^n) = \sum_{x^m \in \mathcal{A}^m} P(x^m) W(y^n|x^m)$$

of Y^n obeys $\rho_n(P^n, Q^n) \leq D$, where D is a prescribed constant.[2]

C2. The channel W maximizes $H(Y^n|X^m)$ simultaneously for all $P \in \mathcal{P}$ among all mappings satisfying C1.

This problem formulation assumes that the key–bit supply is in principle unlimited, and focuses only on the interplay

[2]We emphasize, once again, that Q^n need not be necessarily memoryless.

between conditional entropy and fidelity. However, notice that the expected key rate consumed by an efficient implementation of a simulation scheme is about $H(Y^n|X^m)$ bits output per symbol [6]. Therefore, our goal of maximizing $H(Y^n|X^m)$ will necessarily imply a maximization of the required expected key rate.

In the next three sections, we will study this problem for three different distortion functions $\rho_n(P^n, Q^n)$: (i) a generalized notion of the divergence between two distributions, (ii) the $\bar{\rho}$ distance function (see, e.g., [5]), and (iii) the distance function associated with the so called ϵ–contaminated model [8],[9].

III. The Generalized Divergence Distance

Let $\rho(P, Q)$ denote a distance function (not necessarily a metric) between two probability measures on $\mathcal{A}$, and define the distance between P^n and Q^n as

$$\rho_n(P^n, Q^n) = \frac{1}{n}\sum_{i=1}^{n}\sum_{a^{i-1}} Q(a^{i-1})\rho(P(\cdot|a^{i-1}), Q(\cdot|a^{i-1})).$$

For example, if

$$\rho(P(\cdot|a^{i-1}), Q(\cdot|a^{i-1})) = \sum_{a_i} Q(a_i|a^{i-1}) \log \frac{Q(a_i|a^{i-1})}{P(a_i|a^{i-1})},$$

then ρ_n is the normalized divergence between Q^n and P^n, hence the name "generalized divergence." In general, such additive distance functions between the conditional distributions $\{P(\cdot|a^{i-1})\}$ and $\{Q(\cdot|a^{i-1})\}$ may arise naturally in prediction and sequential decision problems, as they reflect the penalty for mismatch between the assumed probability law and the underlying one.

A. Main Result

Let us define the function:

$$\phi(D) = \sup\{H(Q) : \ \rho(P, Q) \le D\}. \tag{1}$$

We shall assume that $\rho(P, \cdot)$ is convex in Q (which is the case for many useful distance functions), and then it is easily seen that ϕ is concave. Our main result, in this section, is that $\phi(D)$ is the single–letter characterization of the highest possible normalized conditional entropy of Y^n given X^m subject to the constraint $\rho_n(P^n, Q^n) \le D$. Theorem 1 below is the converse theorem and Theorem 2 is the direct (achievability) theorem.

Notice that if ρ induces compact level sets $\{Q : \ \rho(P, Q) \le D\}$, then the supremum that defines $\phi(D)$ is actually a maximum. As we will assume throughout that $\rho(P, \cdot)$ is continuous, this will indeed be the case. Note that when $\rho(P, Q)$ is convex in Q for fixed P, the computation of $\phi(D)$ is a convex program, and hence can be solved by standard convex programming methods. Moreover, in some cases, the maximization can be carried out in closed form. A few examples will be outlined in Subsection III.B.

Theorem 1: (Converse): Let $\rho(P, Q)$ be convex in Q for fixed P. Then, for every simulation scheme W that satisfies condition C1, $H(Y^n|X^m) \le n\phi(D)$.

Proof. Consider first the conditional entropy of the i–th output symbol, Y_i, given Y^{i-1}. Then, we have:

$$\begin{aligned} H(Y_i|Y^{i-1}) &= \sum_{a^{i-1}\in\mathcal{A}^{i-1}} Q(a^{i-1}) H(Q(\cdot|a^{i-1})) \\ &\le \sum_{a^{i-1}\in\mathcal{A}^{i-1}} Q(a^{i-1})\phi(\rho(P(\cdot), Q(\cdot|a^{i-1})) \\ &\le \phi\left(\sum_{a^{i-1}\in\mathcal{A}^{i-1}} Q(a^{i-1})\rho(P(\cdot), Q(\cdot|a^{i-1}))\right), \end{aligned}$$

where in the last step we have used the concavity of ϕ (which follows from the postulated convexity of ρ) and Jensen's inequality. Thus, we obtain:

$$\begin{aligned} & \frac{1}{n}H(Y^n|X^m) \\ \le\ & \frac{1}{n}H(Y^n) \\ =\ & \frac{1}{n}\sum_{i=1}^{n} H(Y_i|Y^{i-1}) \\ \le\ & \frac{1}{n}\sum_{i=1}^{n} \phi\left(\sum_{a^{i-1}} Q(a^{i-1})\rho(P(\cdot), Q(\cdot|a^{i-1}))\right) \\ \le\ & \phi\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{a^{i-1}} Q(a^{i-1})\rho(P(\cdot), Q(\cdot|a^{i-1}))\right) \\ =\ & \phi(\rho_n(P^n, Q^n)) \\ \le\ & \phi(D), \end{aligned} \tag{2}$$

where the third inequality is again an application of Jensen's inequality. This completes the proof of Theorem 1. □

We now move on to the achievability part, which is essentially based on estimating P by the empirical distribution of x^m, finding the optimum Q that achieves $\phi(D)$ with respect to the estimated source rather than the unknown true source, and drawing Y^n according to Q^n. Although this approach seems natural and straightforward, there are nevertheless two non-trivial points to emphasize in this context.

The first point is the following: As can easily be seen from the proof of Theorem 1, even the *unconditional* entropy $H(Y^n)$ is upper bounded by $n\phi(D)$ (and even if the source is known). As we shall see below in Theorem 2 (the direct part), the fact that $H(Y^n|X^m)$, which is smaller, can still approach $n\phi(D)$ has to do with the fact that our proposed scheme assigns an output probability distribution that: (i) depends on x^m only via its type class $T_{\mathbf{x}}$, and (ii) given $T_{\mathbf{x}}$ the output distribution is memoryless and hence given an input type class $T_{\mathbf{x}}$ and an output type class T_{y^n}, the distribution within T_{y^n} is uniform, independently of $T_{\mathbf{x}}$. These two facts imply that $T_{X^m} \to T_{Y^n} \to Y^n$ is a Markov chain, and so,

$$\begin{aligned} & H(Y^n|X^m) \\ =\ & H(Y^n|T_{X^m}) \\ =\ & H(Y^n) - I(T_{X^m}; Y^n) \\ =\ & H(Y^n) - I(T_{X^m}; Y^n, T_{Y^n}) \\ =\ & H(Y^n) - I(T_{X^m}; T_{Y^n}) \end{aligned}$$

$$\geq \quad H(Y^n) - (|\mathcal{A}|-1)\log(\min\{m,n\}+1), \qquad (3)$$

where in the last line we have used the fact that the number of types of n–sequences does not exceed $(n+1)^{|\mathcal{A}|-1}$. Therefore, if $H(Y^n)$ is linear in n, then $H(Y^n|X^m)$ is essentially as large as $H(Y^n)$.

The second point is more related to the distance constraint $\rho_n(P^n, Q^n) \leq D$: Note that as the DMS, Q_{x^m}, applied by the proposed scheme, is estimated from training data x^m, then the overall induced probability distribution Q^n is the mixture of these DMS's weighted by the probabilities of all possible training sequences that could be emitted by the underlying DMS P, i.e.,

$$Q(y^n) = \sum_{x^m} P(x^m) Q_{x^m}(y^n) = \sum_{T_{x^m}} P(T_{x^m}) Q_{x^m}(y^n).$$

This mixture is, of course, no longer memoryless. Moreover, among the components of this mixture, there are contributions of type classes that correspond to empirical distributions that are fairly close to P, so that their weights are appreciably large, but since their associated DMS's have somewhat different single–letter probabilities than Q^*, the achiever of $\phi(D)$ for P, these differences may accumulate when products of n of these letter probabilities are taken to account for the n–vectors that are generated. It is not a–priori obvious then that these cumulative errors do not cause violation of the distance constraint. It turns out, fortunately, that this is not the case. The reason is rooted in a basic fact that is at the heart of the distance analysis in the proof of Theorem 2 below: Although the overall output probability distribution Q, induced by our scheme, is not memoryless, it has the property that $Q(y_i|y^{i-1})$ is close to $Q^*(y_i)$ whenever y^{i-1} is typical to Q^* (for large i). Thus, there is an interesting regenerative mechanism here: If the past is typical to Q^*, then subsequent symbols will continue to be drawn essentially under Q^*, and will then continue to create typical patterns (with high probability), which in turn continue to induce conditional probabilities close to Q^*, and so on.

Finally, a more technical note: In Theorem 2 below, we will assume that $\log m = o(n)$. Operatively, this is, of course, not really a limitation (one can always use only part of the training sequence). But intuitively, it does not seem plausible that more training can harm performance. We believe, therefore, that the need for this technical assumption should be attributed to possible limitations of the bounding techniques, rather than to the real behavior of the simulation scheme (see also [12, Theorem 3]).

Theorem 2: (Direct): Assume that:
(i) The function $\rho(P,Q)$ is continuous at P uniformly in Q, continuous and bounded in Q for a given P, and convex in Q.
(ii) The function ρ induces a unique achiever Q^* of $\sup H(Q)$ subject to (s.t.) the constraints $\rho(P,Q) \leq D$ and $\min_{a\in\mathcal{A}} Q(a) \geq q_{\min}$ for all $q_{\min} \in [0, q_0]$ for some $q_0 > 0$.
(iii) The mapping from P to Q^* is continuous for all $q_{\min} \in [0, q_0]$.
Finally, assume that $\log m = o(n)$. Then, there exists a sequence of simulation schemes, independent of P, that asymptotically (as $m, n \to \infty$) satisfy Condition C1, and whose normalized conditional entropies tend to $\phi(D)$ for all $P \in \mathcal{P}$.

Note that since the direct part guarantees that $H(Y^n|X^m)/n$ approaches $\phi(D)$, which is in turn an upper bound to $H(Y^n)/n$, this means that the normalized mutual information $I(X^m;Y^n)/n \to 0$ (cf. the discussion after the proof of Theorem 1).

Proof of Theorem 2. Consider the following simulation scheme: Given $\mathbf{x} = x^m$, extract its empirical distribution, $P_{\mathbf{x}}$, and then find the achiever $Q_{\mathbf{x}}$ of $\max H(Q)$ s.t. $\rho(P_{\mathbf{x}}, Q) \leq D$ and an additional constraint that $\min_a Q(a) \geq q_{\min}$ for some arbitrarily small $q_{\min} > 0$.[3] Obviously, by elementary continuity arguments, this additional constraint on $\min_a Q(a)$ does not have much effect. In particular, if $q_{\min}$ is sufficiently small, this maximum is arbitrarily close, say, within $\mu(q_{\min})$, to the one obtained without this constraint, and $\lim_{q_{\min}\to 0}\mu(q_{\min}) = 0$. Finally, use $Q_{\mathbf{x}}$ as the target memoryless source that governs Y^n.

We next analyze both the conditional entropy and the distance level associated with the proposed scheme. For a positive real δ, let $\mathcal{T}_P(\delta)$ denote the set of sequences for which the empirical distribution satisfies $\mathcal{D}(P_{\mathbf{x}}\|P) \leq \delta$, where the sequence length will be understood from the context.

As for the conditional output entropy, we have:

$$\begin{aligned}
& \frac{1}{n}H(Y^n|X^m) \\
= \quad & \boldsymbol{E}\{H(Q_{\mathbf{X}})\} \\
\geq \quad & \sum_{\mathcal{T}_{\mathbf{x}}\subset\mathcal{T}_P(\delta)} P(T_{\mathbf{x}}) H(Q_{\mathbf{x}}) \\
= \quad & \sum_{T_{\mathbf{x}}\subset\mathcal{T}_P(\delta)} P(T_{\mathbf{x}}) \times \\
& \max\{H(Q):\ \rho(P_{\mathbf{x}},Q)\leq D,\ \min_a Q(a) \geq q_{\min}\} \\
\geq \quad & \sum_{T_{\mathbf{x}}\subset\mathcal{T}_P(\delta)} P(T_{\mathbf{x}}) \times \\
& \max\{H(Q):\ \rho(P,Q)+\epsilon(\delta)\leq D,\ \min_a Q(a) \geq q_{\min}\} \\
= \quad & P(\mathcal{T}_P(\delta))\cdot[\phi(D-\epsilon(\delta)) - \mu(q_{\min})] \\
\geq \quad & (1-\alpha_m(\delta))\cdot[\phi(D-\epsilon(\delta)) - \mu(q_{\min})], \qquad (4)
\end{aligned}$$

where in the second inequality, we have used the uniform continuity of $\rho(\cdot, Q)$ to argue that $\mathcal{D}(P'\|P) \leq \delta$ implies $|\rho(P,Q) - \rho(P',Q)| \leq \epsilon(\delta)$, with $\lim_{\delta\to 0}\epsilon(\delta) = 0$ independently of Q, and in the last inequality we used the weak law of large numbers (or, the asymptotic equipartition property (AEP)) to argue that $\alpha_m(\delta)$ tends to zero as $m \to \infty$ for every positive δ. Now, since ϕ is concave, it is also continuous (except, perhaps for the edgepoints), and thus $\phi(D)$ is asymptotically achieved for large m and small δ and $q_{\min}$.

It remains to show that $\rho_n(P^n, Q^n)$ is essentially less than D for large n (and m). To this end, we will need the following

[3]The reason for this additional constraint will become apparent in the sequel.

lemma:

Lemma 1: Let

$$Q(a_i|a^{i-1}) \triangleq \frac{\sum_{T_{\mathbf{x}}} P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^i)}{\sum_{T_{\mathbf{x}}} P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^{i-1})}. \tag{5}$$

For a given $\epsilon > 0$, let $i > \epsilon n$ and let $a^{i-1} \in \mathcal{T}_{Q^*}(\epsilon)$ where (with a slight abuse of notation), Q^* is the maximizer of $H(Q)$ s.t. $\rho(P,Q) \le D$ and the additional constraint $\min_{a\in\mathcal{A}} Q(a) \ge q_{\min}$. Then,

$$\max_{a_i\in\mathcal{A}} |Q(a_i|a^{i-1}) - Q^*(a_i)| \le \eta(m,n,q_{\min},\epsilon), \tag{6}$$

where for every given $q_{\min} > 0$,

$$\lim_{\epsilon\to 0}\lim_{n,m\to\infty} \eta(n,m,q_{\min},\epsilon) = 0,$$

with m and n tending to infinity under the regime $\log m = o(n)$.

The proof of Lemma 1 appears in the Appendix.

Now, recall that ρ is assumed uniformly continuous in Q. Since Lemma 1 tells us that $Q(\cdot|a^{i-1})$ is close to Q^* for a typical a^{i-1} and for small (but positive) ϵ and $q_{\min}$ and large enough n and m, then $\rho(P,Q(\cdot|a^{i-1})) \le \rho(P,Q^*) + \gamma(n,m,q_{\min},\epsilon)$, where $\lim_{\epsilon\to 0}\lim_{n,m\to\infty}\gamma(n,m,q_{\min},\epsilon) = 0$ under the regime $\log m = o(n)$. Consider now the i–th term of the distance function ρ_n, where $i > \epsilon n$. Then,

$$\begin{aligned}
&\sum_{a^{i-1}} Q(a^{i-1})\rho(P(\cdot),Q(\cdot|a^{i-1})) \\
&= \sum_{T_{\mathbf{x}}} P(T_{\mathbf{x}}) \sum_{a^{i-1}} Q_{\mathbf{x}}(a^{i-1})\rho(P(\cdot),Q(\cdot|a^{i-1})) \\
&= \sum_{T_{\mathbf{x}}\subseteq\mathcal{T}_P(\delta)} P(T_{\mathbf{x}}) \sum_{a^{i-1}} Q_{\mathbf{x}}(a^{i-1})\rho(P(\cdot),Q(\cdot|a^{i-1})) + \\
&\quad \sum_{T_{\mathbf{x}}\subseteq\mathcal{T}_P^c(\delta)} P(T_{\mathbf{x}}) \sum_{a^{i-1}} Q_{\mathbf{x}}(a^{i-1})\rho(P(\cdot),Q(\cdot|a^{i-1})),
\end{aligned}$$

where the second term vanishes, for any fixed $\delta > 0$, as $P(\mathcal{T}_P^c(\delta))$ vanishes by the weak law of large numbers and ρ is assumed bounded. Let us focus then on the first term, where we upper bound $P(\mathcal{T}_P(\delta))$ by unity. For each $\mathbf{x} \in \mathcal{T}_P(\delta)$,

$$\begin{aligned}
&\sum_{a^{i-1}} Q_{\mathbf{x}}(a^{i-1})\rho(P(\cdot),Q(\cdot|a^{i-1})) \\
&= \sum_{a^{i-1}\in\mathcal{T}_{Q^*}(\epsilon)} Q_{\mathbf{x}}(a^{i-1})\rho(P(\cdot),Q(\cdot|a^{i-1})) + \\
&\quad \sum_{a^{i-1}\in\mathcal{T}_{Q^*}^c(\epsilon)} Q_{\mathbf{x}}(a^{i-1})\rho(P(\cdot),Q(\cdot|a^{i-1})) \\
&\triangleq S_1 + S_2.
\end{aligned} \tag{7}$$

Once again, the second term S_2 on the right–most side of (7) vanishes as it pertains to a–typical sequences, provided that δ is chosen sufficiently small relative to ϵ. Specifically, denoting $\rho_{\max} \triangleq \sup_Q \rho(P,Q)$, S_2 is upper bounded as follows:

$$\begin{aligned}
S_2 &= \sum_{a^{i-1}\in\mathcal{T}_{Q^*}^c(\epsilon)} Q_{\mathbf{x}}(a^{i-1})\rho(P(\cdot),Q(\cdot|a^{i-1})) \\
&\le \rho_{\max}\cdot \sum_{T_{\mathbf{a}}\subset\mathcal{T}_{Q^*}^c(\epsilon)} Q_{\mathbf{x}}(T_{\mathbf{a}}) \\
&\le \rho_{\max}\cdot \sum_{T_{\mathbf{a}}\subset\mathcal{T}_{Q^*}^c(\epsilon)} 2^{-(i-1)\mathcal{D}(P_{\mathbf{a}}\|Q_{\mathbf{x}})} \\
&\le \rho_{\max}\cdot i^{|\mathcal{A}|-1} 2^{-(i-1)\{\mathcal{D}(P_{\mathbf{a}}\|Q^*)-\max_a \log[Q_{\mathbf{x}}(a)/Q^*(a)]\}}.
\end{aligned}$$

Now, since $\mathbf{a} \in \mathcal{T}_{Q^*}^c(\epsilon)$ we have $\mathcal{D}(P_{\mathbf{a}}\|Q^*) > \epsilon$. In addition,

$$\begin{aligned}
\max_a \log\frac{Q_{\mathbf{x}}(a)}{Q^*(a)} &\le \max_a \frac{|Q^*(a)-Q_{\mathbf{x}}(a)|}{\min\{Q^*(a),Q_{\mathbf{x}}(a)\}\ln 2} \\
&\le \max_a \frac{|Q^*(a)-Q_{\mathbf{x}}(a)|}{q_{\min}\ln 2}
\end{aligned} \tag{8}$$

where we used the fact that the logarithmic function is concave and both $Q_{\mathbf{x}}(a)$ and $Q^*(a)$ are lower bounded by $q_{\min} > 0$. Letting $\xi(\delta)$ designate the maximum variational distance between $Q_{\mathbf{x}}$ and Q^* when $\mathcal{D}(P_{\mathbf{x}}\|P) \le \delta$ (so that $\xi(\delta) \to 0$ as $\delta \to 0$ by the postulated continuity of the mapping from P to the maximum–entropy achiever Q), we conclude

$$S_2 \le \rho_{\max}\cdot n^{|\mathcal{A}|-1} 2^{-n\epsilon[\epsilon-\xi(\delta)/(q_{\min}\ln 2)]}.$$

Thus, if δ is sufficiently small such that $\xi(\delta) < \epsilon q_{\min}\ln 2$, this expression vanishes as n grows large. As for the first term S_1 on the right–most side of eq. (7), we have:

$$\begin{aligned}
S_1 &\triangleq \sum_{a^{i-1}\in\mathcal{T}_{Q^*}(\epsilon)} Q_{\mathbf{x}}(a^{i-1})\rho(P(\cdot),Q(\cdot|a^{i-1})) \\
&\le \sum_{a^{i-1}\in\mathcal{T}_{Q^*}(\epsilon)} Q_{\mathbf{x}}(a^{i-1})[\rho(P(\cdot),Q^*(\cdot)) + \gamma(n,m,q_{\min},\epsilon)] \\
&\le \rho(P(\cdot),Q^*(\cdot)) + \gamma(n,m,q_{\min},\epsilon) \\
&\le D + \gamma(n,m,q_{\min},\epsilon).
\end{aligned} \tag{9}$$

Finally, we should add to the distance yet another term that is proportional to ϵ to account for all $i < \epsilon n$. This completes the proof of Theorem 2. □

In the setting of [12], in which there is a "hard limit" nR to the number of key bits available to the simulation scheme, the optimal scheme of Theorem 2 can be faithfully *approximated* provided that $R > \phi(D)$ [7]. It can be shown that this approximation does not affect, asymptotically, the conditional entropy and the distance bound. On the other hand, if $R < \phi(D)$, since nR is an obvious upper bound to $H(Y^n|X^m)$, then it makes sense to decrease D to the level that gives $\phi(D) = R$, because larger values of D mean degrading the fidelity of the output distribution with respect to P, without any gain in the conditional entropy of the output.

B. Examples

We next provide three examples of sources, distortion measures, and their corresponding functions, $\phi(D)$.

1) The Binary Source and Absolute Difference Distortions: The first example is that of the binary source P, characterized by the symbol probabilities p and $1-p$, i.e., $P=\{p,1-p\}$. If $\rho(P,Q)=|p-q|$, with $Q=\{q,1-q\}$, then it is straightforward to see that $Q^*=\{q^*,1-q^*\}$ where

$$q^*=\begin{cases} p+D & p\le \frac{1}{2}-D \\ \frac{1}{2} & |p-\frac{1}{2}|\le D \\ p-D & p>\frac{1}{2}+D \end{cases} \tag{10}$$

Thus,

$$\phi(D)=\begin{cases} h(\hat{p}+D) & D<\frac{1}{2}-\hat{p} \\ 1 & D\ge \frac{1}{2}-\hat{p} \end{cases} \tag{11}$$

where $\hat{p}=\min\{p,1-p\}$ and h is the binary entropy function,

$$h(u)=-u\log u-(1-u)\log(1-u),\quad u\in[0,1]. \tag{12}$$

As can be seen, since $\rho(P,Q)=|p-q|$ is convex in q, then $\phi(D)$ is indeed concave in this example. If, however, we define $\rho(P,Q)=\sqrt{|p-q|}$, which is not convex in q, then $\phi(D)=h(\hat{p}+D^2)$ for $\hat{p}<\frac{1}{2}-D^2$ and $\phi(D)=1$ otherwise, which is not concave in the entire range. For example, if $p=0$, then $\phi(D)=h(D^2)$ is non–concave (actually, it is even convex) in the interval $[0,\Delta]$, where Δ is the solution to the equation[4]

$$\ln\frac{1-D^2}{D^2}=\frac{2}{1-D^2}.$$

which is $\Delta\approx 0.316$.

2) A General DMS with the Divergence Distortion Measure: In our second example, P is a general DMS and

$$\rho(P,Q)=\mathcal{D}(Q\|P)\stackrel{\Delta}{=}\sum_a Q(a)\log\frac{Q(a)}{P(a)}. \tag{13}$$

Using a simple Lagrange multiplier, the achiever Q^* of $\phi(D)$ is seen to take the form:

$$Q^*(a)=\frac{P(a)^\theta}{\sum_{a'\in\mathcal{A}}P(a')^\theta},\quad a\in\mathcal{A}$$

where $0\le\theta\le 1$ is chosen to meet the distance constraint with equality, i.e., $\mathcal{D}(Q\|P)=D$. The two extremes are $\theta=1$, corresponding to $D=0$, and $\theta=0$, corresponding to $D=\mathcal{D}(P_U\|P)$, P_U being the uniform probability distribution on $\mathcal{A}$. In this case, there is no apparent closed–form expression for $\phi(D)$, but we can easily express both D and $\phi(D)$ parametrically via θ, by

$$\begin{aligned} D_\theta &= -\log\Gamma_\theta-\frac{1-\theta}{\Gamma_\theta}\sum_a P(a)^\theta\log P(a) \\ \phi(D_\theta) &= \log\Gamma_\theta-\frac{\theta}{\Gamma_\theta}\sum_a P(a)^\theta\log P(a) \end{aligned}$$

where

$$\Gamma_\theta=\sum_a P(a)^\theta.$$

[4]The difference between the left–hand side and the right–hand side of this equation is proportional to the second derivative of $\phi(D)=h(D^2)$, and so, the solution $D=\Delta$ is the critical point below which the second derivative is positive.

3) A General DMS and the Quadratic Distortion Measure: The third and last example, is again for a general DMS, but now the distortion measure is quadratic, i.e.,

$$\rho(P,Q)=\sum_{a\in\mathcal{A}}[P(a)-Q(a)]^2. \tag{14}$$

We have not been able to find closed–form expressions for Q^* and for $\phi(D)$ in general, but we can characterize the behavior for small values of D. To this end, we will assume throughout that P corresponds to an internal point in the simplex, i.e., $P(a)>0$ for all $a\in\mathcal{A}$. We will also assume that D is so small so that Q is sufficiently close to P, and then $H(Q)$ can be approximated by

$$H(Q)\approx H(P)+\nabla H\cdot(Q-P) \tag{15}$$

where ∇H is the gradient of $H(\cdot)$ at P. I.e., labeling the alphabet letters by $1,2,\ldots,K\stackrel{\Delta}{=}|\mathcal{A}|$, we define:

$$\begin{aligned} \nabla H &= -(\log[eP(1)],\ldots,\log[eP(K)]) \\ &\stackrel{\Delta}{=} (g_1,\ldots,g_K). \end{aligned} \tag{16}$$

Denoting $e_i=Q(i)-P(i)$, $i=1,\ldots,K$, the problem of finding Q^* is, to the first order approximation, equivalent to maximizing $\sum_i g_i e_i$ subject to the constraints $\sum_i e_i^2\le D$ and $\sum_i e_i=0$. This is a standard convex program, which is easily solved using Lagrange multipliers, and the solution is given by

$$e_i^*=\sqrt{\frac{D}{\sum_j(g_j-\bar{g})^2}}\cdot(g_i-\bar{g}), \tag{17}$$

where

$$\bar{g}=\frac{1}{K}\sum_i g_i. \tag{18}$$

Substituting this into (15), we get the small distortion approximation:

$$\phi(D)\approx H(P)+\sqrt{D\cdot\sum_i(g_i-\bar{g})^2}, \tag{19}$$

that is, for small D, $\phi(D)$ exceeds $H(P)$ by an amount that is proportional to $\sqrt{D}$, and the constant of proportionality depends only on the "variance" of $\{g_i\}$, which is large for very skewed distributions and becomes smaller, as the uniform distribution is approached, since there is less room for enlarging the entropy as P is closer to be uniform.

C. Outline of an Extension to Markov Sources

Generalizing Theorem 1 and Theorem 2 to the Markov case requires some more caution. The converse part is fairly simple, as we show next. As for the direct part, we have not carried out an extension of Theorem 2 to the Markov case in full detail, but we will briefly outline how, we believe, this can be handled for first–order Markov sources (further extension to higher orders is then straightforward).

For simplicity, let us assume that Y^n is required to be stationary, which is a reasonable assumption when the input

is stationary. We will also assume now that ρ is convex in Q. Let us now define

$$\phi(D) = \max\{H(Y_1|Y_0) : \text{ dist}\{Y_0\} = \text{dist}\{Y_1\}, \sum_a Q(a)\rho(P(\cdot|a), Q(\cdot|a)) \le D\}, \quad (20)$$

where $H(Y_1|Y_0)$ is the conditional entropy of Y_1 given Y_0 under the first–order Markov probability measure Q, and the maximization is over the transition probabilities $\{Q(b|a),\ a, b \in \mathcal{A}\}$ and the unconditional marginal distributions, $\{Q(a),\ a \in \mathcal{A}\}$, subject to the constraints that the unconditional marginal distributions of Y_0 and Y_1 are the same (i.e., $\sum_{a\in\mathcal{A}} Q(a)Q(b|a) = Q(b)$ for all $b \in \mathcal{A}$), and the weighted distance constraint between the transition probability distributions $\{Q(\cdot|a)\}$ and $\{P(\cdot|a)\}$ is maintained. Also, let

$$\phi(D; Q_0) = \max\{H(Y_1|Y_0) : \text{ dist}\{Y_0\} = \text{dist}\{Y_1\} = Q_0, \sum_a Q_0(a)\rho(P(\cdot|a), Q(\cdot|a)) \le D\}, \quad (21)$$

and observe that for a given Q_0, $\phi(\cdot; Q_0)$ is concave (due to the convexity of ρ in Q).

As for the converse part, first observe that for every $i = 2, \ldots, n$, we have

$$\begin{aligned} D_i &\stackrel{\Delta}{=} \sum_{a^{i-1}} Q(a^{i-1})\rho(P(\cdot|a_{i-1}), Q(\cdot|a^{i-1})) \\ &= \sum_{a_{i-1}} Q(a_{i-1}) \sum_{a^{i-2}} Q(a^{i-2}|a_{i-1}) \times \\ &\quad \rho(P(\cdot|a_{i-1}), Q(\cdot|a_{i-1}, a^{i-2})) \\ &\ge \sum_{a_{i-1}} Q(a_{i-1}) \times \\ &\quad \rho(P(\cdot|a_{i-1}), \sum_{a^{i-2}} Q(a^{i-2}|a_{i-1})Q(\cdot|a_{i-1}, a^{i-2})) \\ &= \sum_{a_{i-1}} Q(a_{i-1})\rho(P(\cdot|a_{i-1}), Q(\cdot|a_{i-1})) \stackrel{\Delta}{=} D_i', \end{aligned} \quad (22)$$

where the inequality follows from the assumed convexity of ρ. Thus, for any simulation scheme with a given marginal Q_0 of each Y_i, we have

$$\begin{aligned} H(Y^n|X^m) &\le \sum_i H(Y_i|Y_{i-1}) \\ &\le \sum_i \phi(D_i'; Q_0) \\ &\le n\phi\left(\frac{1}{n}\sum_i D_i'; Q_0\right) \\ &\le n\phi(D; Q_0) \le n\phi(D), \end{aligned} \quad (23)$$

where the second to the last inequality follows from (22).

The achievability scheme may be constructed and analyzed in the same spirit as in Theorem 2 except that the memoryless structure is replaced by the Markov one: First, compute $\phi(D)$ of eq. (20) with P being replaced by the empirical Markov source extracted from x^m. Then, draw Y^n according to the achiever Q of $\phi(D)$. Here we impose the $q_{\min}$ constraints on the transition probabilities and hence they are met also by unconditional marginals of single symbols associated with $Q_{\mathbf{x}}$. This should apparently have an arbitrarily small effect on both the entropy and on the distance from P when $q_{\min} > 0$ is sufficiently small. Once again, since the details of this have not been worked out, we make no formal claims about this extension, but we find it plausible.

IV. The $\bar{\rho}$ Distance Measure

A related result is now developed for the $\bar{\rho}$ distance measure considered in [16] and [13], where distances between probability measures are induced by distortion measures between sequences of random variables (see, e.g., [5]). The results in this section apply to the memoryless case, and do not seem to lend themselves easily to extensions to sources with memory.

Let $\rho : \mathcal{A}^2 \to \mathbb{R}^+$ be a given single–letter distortion measure, and consider the Ornstein $\bar{\rho}$ distance, $\bar{\rho}_n(P^n, Q^n)$, between two measures P^n and Q^n of n–vectors in $\mathcal{A}^n$, i.e., the infimum of $\frac{1}{n}\sum_{i=1}^n \boldsymbol{E}\rho(\tilde{X}_i, \tilde{Y}_i)$ across all joint distributions of $(\tilde{X}^n, \tilde{Y}^n)$ for which the marginal of $\tilde{X}^n$ is P^n and the marginal of $\tilde{Y}^n$ is Q^n.[5] Thus, loosely speaking, the $\bar{\rho}$ distance gives the best explanation of $\tilde{Y}^n \sim Q^n$ as a distorted version of $\tilde{X}^n \sim P^n$ via some channel. Notice that, as shown in [5, Theorem 8.3.1], the infimum in the definition of the $\bar{\rho}$ distance is always achieved. For a given distortion level D, we will allow the probability law Q^n of Y^n to be at $\bar{\rho}$ distance at most D from P^n, i.e., $\bar{\rho}_n(P^n, Q^n) \le D$.

Define the single–letter function:

$$\gamma(D) = \max\{H(Y) : \boldsymbol{E}\rho(X, Y) \le D\} \quad (24)$$

where $X \sim P$ and the maximization is across conditional distributions $\{W(y|x),\ x, y \in \mathcal{A}\}$ that satisfy the distortion constraint. It is easy to see that $\gamma(\cdot)$ is concave (simply because the entropy is concave).

For example, referring to the first example in Subsection III.B, if P is binary with parameter p, and ρ is the Hamming distortion measure, then $\gamma(D) = h(\hat{p} + D)$ for $D < 1/2 - \hat{p}$ and $\gamma(D) = 1$ otherwise, where h is the binary entropy function defined in (12), and $\hat{p} = \min\{p, 1-p\}$ (thus, $\gamma(D) = \phi(D)$ in this example).

Our converse theorem asserts that $\gamma(D)$ is an upper bound to the per–symbol conditional entropy.

Theorem 3: (Converse): For every simulation scheme that satisfies $\bar{\rho}_n(P^n, Q^n) \le D$, we have $H(Y^n|X^m) \le n\gamma(D)$.

Proof. Given a simulation scheme W with $Y^n \sim Q^n$ that satisfies the $\bar{\rho}$ distance constraint, then by definition, there must exist random vectors $\tilde{X}^n \sim P^n$ and $\tilde{Y}^n \sim Q^n$ linked by a

[5] We are deliberately denoting here the random vector corresponding to P by $\tilde{X}^n$, because it may not coincide with the training sequence although both are governed by P. Similarly, $\tilde{Y}^n$ may not coincide with the simulated sequence although it is also governed by Q^n.

channel $\tilde{W}$ such that $\frac{1}{n}\sum_{i=1}^{n} E\rho(\tilde{X}_i, \tilde{Y}_i) \leq D$. Thus,

$$\begin{aligned} H(Y^n|X^m) &\leq \sum_{i=1}^{n} H(Y_i) = \sum_{i=1}^{n} H(\tilde{Y}_i) \\ &\leq \sum_{i=1}^{n} \gamma(\boldsymbol{E}\rho(\tilde{X}_i, \tilde{Y}_i)) \\ &\leq n\gamma\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{E}\rho(\tilde{X}_i, \tilde{Y}_i)\right) \\ &\leq n\gamma(D), \end{aligned} \tag{25}$$

where the first inequality is because conditioning reduces entropy, the second is by definition of $\gamma(\cdot)$, the third is due to the concavity of $\gamma(\cdot)$, and the fourth is due to its monotonicity and the aforementioned distortion constraint. This completes the proof of Theorem 3. □

Theorem 4: (Direct): There exists a sequence of simulation schemes, independent of P, that asymptotically (as $m, n \to \infty$) satisfy $\bar{\rho}_n(P^n, Q^n) \leq D$ (Condition C1), and whose normalized conditional entropies $H(Y^n|X^m)$ tend to $\gamma(D)$ for all $P \in \mathcal{P}$.

Proof. If $m > n$, we will ignore the training samples $X_{n+1}, \ldots, X_m$, and so, reduce m to the value of n. Thus, from this point, we will assume $m = n$ and denote both integers by n. For a given P, let $f(P)$ denote the output marginal induced by P and by the channel W_P that attains $\gamma(D)$. For a given training sequence $x^n = \mathbf{x}$, let $Q_n = [f(P_{\mathbf{x}})]_n$, where the operation $[\cdot]_n$ means quantization to a rational distribution with denominator n, in the following manner: Given $P_{\mathbf{x}}$, find the channel $\{W_{P_{\mathbf{x}}}(y|x),\ x, y \in A\}$ that maximizes $H(\hat{Y})$ subject to the constraint $\boldsymbol{E}\rho(\hat{X}, \hat{Y}) \leq D$, where $\hat{X}$ is a random variable drawn according to $P_{\mathbf{x}}$. Next, for each $a \in \mathcal{A}$, quantize the transition probabilities $\{W_{P_{\mathbf{x}}}(b|a), b \in \mathcal{A}\}$ to the nearest rational numbers (say, in the Euclidean sense) with denominator $n(a|x^n)$ – the number of occurrences of a in x^n, keeping the constraint that they sum up to unity. This determines a channel W_n and guarantees that the output marginal Q_n induced by $P_{\mathbf{x}}$ and W_n will be rational with denominator n. According to the proposed simulation scheme, Y^n is drawn uniformly from the type class $T(Q_n)$ corresponding to Q_n.[6] We now have to show that: (i) the output distribution of Y^n is within $\bar{\rho}$–distance D from P^n, and (ii) the performance is close to $\gamma(D)$ for large enough n.

As for (i), consider the following argument: For a given $y^n = \mathbf{y}$, let $T_{\mathbf{y}}$ denote its type class and let $Q_{\mathbf{y}}$ denote its empirical distribution. Let $f^{-1}(Q_{\mathbf{y}})$ denote the set of $\{T_{\mathbf{x}}\}$ such that $Q_{\mathbf{y}} = [f(P_{\mathbf{x}})]_n$. Then, on the one hand, we obviously have:

$$\begin{aligned} \Pr\{Y^n = \mathbf{y}\} &= \sum_{T_{\mathbf{x}} \in f^{-1}(Q_{\mathbf{y}})} P(T_{\mathbf{x}}) \cdot \frac{1}{|T_{\mathbf{y}}|} \\ &= \sum_{T_{\mathbf{x}} \in f^{-1}(Q_{\mathbf{y}})} P(\mathbf{x}) \cdot \frac{|T_{\mathbf{x}}|}{|T_{\mathbf{y}}|}. \end{aligned} \tag{26}$$

[6]Here the proposed approach is somewhat different from the one in Section 3, the reason being mostly convenience and simplicity of the proof.

On the other hand, we would like to show that this distribution of Y^n can be represented as the distribution of the output $\tilde{Y}^n$ of a channel $\tilde{W}(\tilde{y}^n|\tilde{x}^n)$, whose input $\tilde{X}^n$ is drawn by P, and which satisfies $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{E}\{\rho(\tilde{X}_i, \tilde{Y}_i)\} \leq D$. Consider the channel $\tilde{W}(\tilde{y}^n|\tilde{x}^n)$ that puts all its mass uniformly within the *conditional* type class $T(W_n)$ corresponding to the quantized channel W_n described in the previous paragraph. Since $T(W_n)$ depends on $\tilde{\mathbf{x}}$ only through $T_{\tilde{\mathbf{x}}}$, we have

$$\Pr\{\tilde{Y}^n = \mathbf{y}\} = \sum_{T_{\mathbf{x}}} \frac{P(\mathbf{x})}{|T(W_n)|} \sum_{\tilde{\mathbf{x}} \in T_{\mathbf{x}}} 1\{\tilde{\mathbf{x}} : T_{\mathbf{y}|\tilde{\mathbf{x}}} = T(W_n)\}\,.$$

Clearly, the set $\{\tilde{\mathbf{x}} \in T_{\mathbf{x}} : T_{\mathbf{y}|\tilde{\mathbf{x}}} = T(W_n)\}$ is empty if $T_{\mathbf{x}} \notin f^{-1}(Q_{\mathbf{y}})$, or, by Bayes' rule, has cardinality $|T(W_n)||T_{\mathbf{x}}|/|T_{\mathbf{y}}|$ otherwise. Therefore,

$$\Pr\{\tilde{Y}^n = \mathbf{y}\} = \sum_{T_{\mathbf{x}} \in f^{-1}(Q_{\mathbf{y}})} P(\mathbf{x}) \cdot \frac{|T_{\mathbf{x}}|}{|T_{\mathbf{y}}|} = \Pr\{Y^n = \mathbf{y}\}\,.$$

Since joint typicality guarantees that the distortion between $\tilde{X}^n$ and $\tilde{Y}^n$, induced by this channel, is always within D (hence, *a–fortiori* its expectation), this means that the distribution of Y^n satisfies the $\bar{\rho}$ distance constraint.

As for (ii), we have:

$$\begin{aligned} H(Y^n|X^n) &= \boldsymbol{E}\{\log|T(Q_n)|\} \\ &= n\boldsymbol{E}\{H([f(P_{X^n})]_n)\} - O(\log n) \\ &= n[\gamma(D) - \epsilon_n] \end{aligned} \tag{27}$$

where ϵ_n tends to 0 as n grows without bound and where the last passage is due to the law of large numbers, the continuity of f, the vanishing effect of the operation $[\cdot]_n$, and the fact that $H(f(P)) = \gamma(D)$. This completes the proof of Theorem 4. □

The following comments are in order regarding the scheme of Theorem 4:

1) The scheme is different from the one that was described in [13] in the context of the $\bar{\rho}$ distance measure. As discussed in the Introduction, the scheme in [13] aims at minimizing the mutual information between the input training vector and the output vector. While *both* schemes guarantee a vanishingly small mutual information as n (and m) grow without bound, the scheme proposed in [13] is inferior to the one proposed herein in terms of the conditional output entropy (about $nR(D)$ as opposed to $n\gamma(D)$, respectively, where $R(D)$ is the rate–distortion function of the source). On the other hand, the scheme in [13] is more economical in terms of consuming key bits, an aspect of the problem that we do not study in this work.
2) Our comment in Section III regarding the behavior of the simulation scheme in case only nR key bits are available remains valid for the $\bar{\rho}$ distance measure. Here, a uniformly random draw from a type must be approximated following the ideas in [12], provided $R > \gamma(D)$.

V. The ϵ–Contaminated Model

Consider now another distortion function between two probability distributions of n–vectors, P^n and Q^n:

$$\rho_n(P^n, Q^n) = 1 - \min_{y^n \in \mathcal{A}^n} \frac{Q(y^n)}{P(y^n)}, \tag{28}$$

where $0/0 \stackrel{\Delta}{=} 1$. Our first observation is that $\rho_n(P^n, Q^n) \le D$ if and only if $Q(y^n)$ can be represented in the form

$$Q(y^n) = (1-D)P(y^n) + D \cdot R(y^n), \tag{29}$$

where $R(\cdot)$ is an arbitrary probability distribution on $\mathcal{A}^n$. The "if" part follows immediately since (29) implies that $Q(y^n) \ge (1-D)P(y^n)$, and so, $1 - Q(y^n)/P(y^n) \le D$ for all $y^n \in \mathcal{A}^n$. The "only if" part follows from the fact that $\rho_n(P^n, Q^n) \le D$ implies that

$$D \cdot R(y^n) \stackrel{\Delta}{=} Q(y^n) - (1-D)P(y^n) \ge 0 \quad \forall y^n \in \mathcal{A}^n \tag{30}$$

and moreover, taking the summation over all $y^n \in \mathcal{A}^n$, we get

$$D \cdot \sum_{y^n} R(y^n) = 1 - (1-D) = D, \tag{31}$$

that is, $\sum_{y^n} R(y^n) = 1$. Since $R(y^n)$ is non–negative and it sums up to unity, it is a probability distribution.

Having established this equivalence, we will refer, from now on, to eq. (29) as our model, which is well–known as the *ϵ–contaminated model* (with the notation D being replaced by ϵ) in the literature of robust statistics (see, e.g., [8], [9] and references therein) and it is customarily used for describing a small uncertainty with regard to the actual probability distribution Q about the nominal (desired) distribution P.

Eq. (29) can be interpreted as the result of the action of an underlying switch, i.e., a binary random variable, S, taking the values g and b (standing for "good" and "bad", respectively) with probabilities $1-D$ and D, respectively. This switch multiplexes between two distributions of y^n, namely, $\Pr\{Y^n = y^n | S = g\} = P(y^n)$ and $\Pr\{Y^n = y^n | S = b\} = R(y^n)$. Thus, $Q(y^n)$ is the marginal of y^n derived from the joint distribution of Y^n and S.

When a simulation system enters the picture, we will have in mind a joint distribution of (X^m, Y^n, S), which is, in general, given by

$$\begin{aligned} &\Pr(X^m = x^m, Y^n = y^n, S = s) \\ = \;& P(x^m) M(s|x^m) W(y^n|x^m, s)\,, \end{aligned} \tag{32}$$

with the interpretation that the simulator, upon receiving x^m, first randomly chooses either $S = g$ or $S = b$, with probabilities $M(g|x^m)$ and $M(b|x^m) = 1 - M(g|x^m)$, respectively, and then applies the corresponding simulation channel $W_g(y^n|x^m) = W(y^n|x^m, g)$ or $W_b(y^n|x^m) = W(y^n|x^m, b)$. A simulation scheme, ψ, in this setting, is then defined by the choice of $M(g|x^m)$ (or, equivalently, $M(b|x^m)$) for every $x^m \in \mathcal{A}^m$ as well as the channels $W_g(y^n|x^m)$ and $W_b(y^n|x^m)$ under the constraint that $M(b) = \sum_{x^m} P(x^m) M(b|x^m) \le D$. We would like then to select these ingredients in a way that is independent of the unknown source P, and that maximizes $H(Y^n|X^m)$ without violating the distortion constraint $\rho_n(P^n, Q^n) \le D$, or equivalently, keeping $Q(y^n)$ in the form (29). As in Sections II–IV, we will assume that the distribution P that governs X^m is i.i.d. Extensions to more general families of sources are quite straightforward. We will also assume here that $m \ge n$.

Theorem 5 below characterizes the best achievable performance and suggests a conceptually simple way to approach it.

Theorem 5: Let X^m be drawn from a memoryless source P with entropy H, and define

$$\psi(D) = (1-D)H + D \log |\mathcal{A}|\,. \tag{33}$$

(a) (Converse part): For any simulation scheme that satisfies $\rho_n(P^n, Q^n) \le D$, we have

$$H(Y^n|X^m) \le n\psi(D) + 1. \tag{34}$$

(b) (Direct part): There exists a sequence of simulation schemes, independent of P, that asymptotically satisfy $\rho_n(P^n, Q^n) \le D$, and at the same time:

$$H(Y^n|X^m) \ge n\psi(D) - \mu_{n,m} + C + o(1) \tag{35}$$

where C is a constant and

$$\mu_{m,n} = \begin{cases} \frac{|\mathcal{A}|-1}{2} \log \frac{m}{m-n} & m > n \\ \frac{|\mathcal{A}|-1}{2} \log n & m = n. \end{cases} \tag{36}$$

Proof. As for part (a), we have:

$$\begin{aligned} & H(Y^n|X^m) \\ \le\;& H(Y^n) \\ =\;& H(Y^n|S) + I(S; Y^n) \\ \le\;& H(Y^n|S) + 1 \\ =\;& (1-D)H(Y^n|S=g) + D \cdot H(Y^n|S=b) + 1 \\ =\;& (1-D)nH + D \cdot H(Y^n|S=b) + 1 \\ \le\;& n[(1-D)H + D \log |\mathcal{A}|] + 1 \\ =\;& n\psi(D) + 1, \end{aligned} \tag{37}$$

where the second inequality is due to the fact that S is binary.

For part (b), consider the following simulation scheme: Let $M^*(b|x^m) = D$ for all x^m (i.e., S is independent of X^m), $W_b^*(y^n|x^m) = 1/|\mathcal{A}|^n$ for all $x^m \in \mathcal{A}^m$ and $y^n \in \mathcal{A}^n$, and let $W_g^*(y^n|x^m)$ be the optimum simulation channel derived in [12], which preserves the input distribution P in the case of unlimited key rate, that is, the channel induced by letting Y^n be the first n symbols of a random permutation of X^m. The induced output distribution will then be

$$Q(y^n) = (1-D)P(y^n) + D \cdot \frac{1}{|\mathcal{A}|^n}, \tag{38}$$

which complies with (29). As for the conditional output entropy, we have

$$\begin{aligned} & H(Y^n|X^m) \\ \geq\ & H(Y^n|X^m,S) \\ =\ & (1-D)H(W_g^*(\cdot|X^m)) + D\cdot H(W_b^*(\cdot|X^m)) \\ =\ & (1-D)H(W_g^*(\cdot|X^m)) + D\cdot n\log|\mathcal{A}|. \end{aligned} \tag{39}$$

The proof is completed by using two facts shown in [12, eqs. (26)–(28)]: The first fact is that

$$E\{H(W_g^*(\cdot|X^m))\} = E\log|T_{X^m}| - E\log|T_{X^{m-n}}|, \tag{40}$$

where if $m=n$ the second term is defined to be zero, and the second fact is that for a general positive integer k, when $k\to\infty$,

$$E\log|T_{X^k}| = kH - \frac{|A|-1}{2}\log(2\pi ek) + \text{const} + o(1). \tag{41}$$

Therefore, (35) follows both for $(m-n)\to\infty$ and when $(m-n)$ is bounded by a constant.

APPENDIX

Proof of Lemma 1.

For a given $\epsilon>0$ and a positive integer $i>n\epsilon$, let $a^{i-1}\in\mathcal{T}_{Q^*}(\epsilon)$, i.e., $\mathbf{a}=a^{i-1}$ has an empirical distribution $P_{\mathbf{a}}$ that satisfies $\mathcal{D}(P_{\mathbf{a}}\|Q^*)\leq\epsilon$. Next define the set $\mathcal{S}_\epsilon$ (depending on a^{i-1}) as the set of types $\{T_{\mathbf{x}}\}$ for which:

$$P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^{i-1}) \geq 2^{-n\epsilon^2}\max_{\mathbf{x}'}[P(T_{\mathbf{x}'})Q_{\mathbf{x}'}(a^{i-1})]. \tag{A.1}$$

Generally speaking, $\mathcal{S}_\epsilon$ contains the dominant terms of the denominator of the right–hand side of eq. (5). Obviously, it contains at least the largest term, $\max_{\mathbf{x}}P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^{i-1})$. It follows then, by definition, that

$$\begin{aligned} & \sum_{T_{\mathbf{x}}\in\mathcal{S}_\epsilon^c} P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^{i-1}) \\ \leq\ & 2^{-n\epsilon^2}\cdot|\mathcal{S}_\epsilon^c|\max_{\mathbf{x}}P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^{i-1}) \\ \leq\ & 2^{-n\epsilon^2}(m+1)^{|\mathcal{A}|-1}\cdot\sum_{T_{\mathbf{x}}\in\mathcal{S}_\epsilon}P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^{i-1}) \\ \stackrel{\Delta}{=}\ & \zeta_{n,m}\cdot\sum_{T_{\mathbf{x}}\in\mathcal{S}_\epsilon}P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^{i-1}), \end{aligned} \tag{A.2}$$

where we have again used the fact [4] that the number of type classes of m–sequences does not exceed $(m+1)^{|\mathcal{A}-1}$, and where $\zeta_{n,m}\to0$ as $m,n\to\infty$ because $\log m = o(n)$ as postulated. Thus,

$$\begin{aligned} & \sum_{T_{\mathbf{x}}}P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^{i-1}) \\ =\ & \sum_{T_{\mathbf{x}}\in\mathcal{S}_\epsilon}P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^{i-1}) + \sum_{T_{\mathbf{x}}\in\mathcal{S}_\epsilon^c}P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^{i-1}) \\ \leq\ & (1+\zeta_{n,m})\cdot\sum_{T_{\mathbf{x}}\in\mathcal{S}_\epsilon}P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^{i-1}). \end{aligned} \tag{A.3}$$

In a similar manner, referring to the numerator of (5), we have:

$$\begin{aligned} & \sum_{T_{\mathbf{x}}\in\mathcal{S}_\epsilon^c}P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^i) \\ \leq\ & \sum_{T_{\mathbf{x}}\in\mathcal{S}_\epsilon^c}P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^{i-1}) \\ \leq\ & \zeta_{n,m}\cdot\sum_{T_{\mathbf{x}}\in\mathcal{S}_\epsilon}P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^{i-1}) \\ \leq\ & \frac{\zeta_{n,m}}{q_{\min}}\cdot\sum_{T_{\mathbf{x}}\in\mathcal{S}_\epsilon}P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^{i-1})Q_{\mathbf{x}}(a_i) \\ =\ & \frac{\zeta_{n,m}}{q_{\min}}\cdot\sum_{T_{\mathbf{x}}\in\mathcal{S}_\epsilon}P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^i) \end{aligned} \tag{A.4}$$

and similarly as before, we now get

$$\begin{aligned} & \sum_{T_{\mathbf{x}}}P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^i) \\ \leq\ & \left(1+\frac{\zeta_{n,m}}{q_{\min}}\right)\cdot\sum_{T_{\mathbf{x}}\in\mathcal{S}_\epsilon}P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^i). \end{aligned} \tag{A.5}$$

We therefore obtain the following upper and lower bounds to $Q(a_i|a^{i-1})$:

$$\begin{aligned} Q(a_i|a^{i-1}) &= \frac{\sum_{T_{\mathbf{x}}}P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^i)}{\sum_{T_{\mathbf{x}}}P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^{i-1})} \\ &\leq \frac{(1+\zeta_{n,m}/q_{\min})\cdot\sum_{T_{\mathbf{x}}\in\mathcal{S}_\epsilon}P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^i)}{\sum_{T_{\mathbf{x}}\in\mathcal{S}_\epsilon}P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^{i-1})} \\ &\leq \left(1+\frac{\zeta_{n,m}}{q_{\min}}\right)\cdot\max_{T_{\mathbf{x}}\in\mathcal{S}_\epsilon}\frac{P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^i)}{P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^{i-1})} \\ &= \left(1+\frac{\zeta_{n,m}}{q_{\min}}\right)\cdot\max_{T_{\mathbf{x}}\in\mathcal{S}_\epsilon}Q_{\mathbf{x}}(a_i), \end{aligned} \tag{A.6}$$

where we have used the inequality [2, Lemma 1]

$$\frac{\sum_{i=1}^N\alpha_i}{\sum_{i=1}^N\beta_i} \leq \max_{1\leq i\leq N}\frac{\alpha_i}{\beta_i} \tag{A.7}$$

for a positive integer N and for positive $\{\alpha_i\}_{i=1}^N$ and $\{\beta_i\}_{i=1}^N$. On the other hand,

$$\begin{aligned} Q(a_i|a^{i-1}) &= \frac{\sum_{T_{\mathbf{x}}}P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^i)}{\sum_{T_{\mathbf{x}}}P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^{i-1})} \\ &\geq \frac{\sum_{T_{\mathbf{x}}\in\mathcal{S}_\epsilon}P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^i)}{(1+\zeta_{n,m})\cdot\sum_{T_{\mathbf{x}}\in\mathcal{S}_\epsilon}P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^{i-1})} \\ &\geq \frac{1}{1+\zeta_{n,m}}\cdot\min_{T_{\mathbf{x}}\in\mathcal{S}_\epsilon}\frac{P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^i)}{P(T_{\mathbf{x}})Q_{\mathbf{x}}(a^{i-1})} \\ &= \frac{1}{1+\zeta_{n,m}}\cdot\min_{T_{\mathbf{x}}\in\mathcal{S}_\epsilon}Q_{\mathbf{x}}(a_i), \end{aligned} \tag{A.8}$$

where we have now used the inequality

$$\frac{\sum_{i=1}^N\alpha_i}{\sum_{i=1}^N\beta_i} \geq \min_{1\leq i\leq N}\frac{\alpha_i}{\beta_i}, \tag{A.9}$$

which is obviously equivalent to (A.7).

Next, we argue that for $\mathbf{a}=a^{i-1}\in\mathcal{T}_{Q^*}(\epsilon)$, $i>\epsilon n$, every $T_{\mathbf{x}}\in\mathcal{S}_\epsilon$ must induce $Q_{\mathbf{x}}$ which is close to Q^*, and so, in

particular, the minimizer and the maximizer of $Q_{\mathbf{x}}(a_i)$ within $\mathcal{S}_\epsilon$ must be close to $Q^*(a_i)$. To see why this is true, observe that since [4]

$$(m+1)^{-|\mathcal{A}|+1}2^{-mD(P_{\mathbf{x}}\|P)} \leq P(T_{\mathbf{x}}) \leq 2^{-mD(P_{\mathbf{x}}\|P)},$$

and since $Q_{\mathbf{x}}(\mathbf{a}) = 2^{-(i-1)[H(P_{\mathbf{a}})+\mathcal{D}(P_{\mathbf{a}}\|Q_{\mathbf{x}})]}$, the condition that defines $\mathcal{S}_\epsilon$ implies that

$$\begin{aligned}&m\mathcal{D}(P_{\mathbf{x}}\|P) + (i-1)\mathcal{D}(P_{\mathbf{a}}\|Q_{\mathbf{x}}) \leq \\ &\min_{\mathbf{x}'}[m\mathcal{D}(P_{\mathbf{x}'}\|P) + (i-1)\mathcal{D}(P_{\mathbf{a}}\|Q_{\mathbf{x}'})] + n\epsilon^2 + \xi_m,\end{aligned}$$

where $\xi_m = (|\mathcal{A}|-1)\log(m+1)$. The left–hand side is, of course, lower bounded by $(i-1)\mathcal{D}(P_{\mathbf{a}}\|Q_{\mathbf{x}})$. As for the right–hand side, let $P_{\mathbf{x}^*} = \mathrm{argmin}_{P_{\mathbf{x}}}\mathcal{D}(P_{\mathbf{x}}\|P)$, where the minimum obviously exists since $P_{\mathbf{x}}$ belongs to a finite set. Then,

$$\begin{aligned}&\min_{\mathbf{x}'}[m\mathcal{D}(P_{\mathbf{x}'}\|P) + (i-1)\mathcal{D}(P_{\mathbf{a}}\|Q_{\mathbf{x}'})] \\ \leq\ & m\mathcal{D}(P_{\mathbf{x}^*}\|P) + (i-1)\mathcal{D}(P_{\mathbf{a}}\|Q_{\mathbf{x}^*}) \\ \leq\ & m\boldsymbol{E}\{\mathcal{D}(P_{\mathbf{X}}\|P)\} + (i-1)\mathcal{D}(P_{\mathbf{a}}\|Q^*) + \\ & (i-1)\cdot\sum_a P_{\mathbf{a}}(a)\log\frac{Q^*(a)}{Q_{\mathbf{x}^*}(a)} \\ \leq\ & (|\mathcal{A}|-1)\log e + (i-1)\epsilon + \\ & (i-1)\cdot\max_a\log\frac{Q^*(a)}{Q_{\mathbf{x}^*}(a)},\end{aligned} \tag{A.10}$$

where, in the last passage, we have used the inequality $m\boldsymbol{E}\{\mathcal{D}(P_{\mathbf{X}}\|P)\} \leq (|\mathcal{A}|-1)\log e$ (see [1] and ref. [19, Proposition 5.2 therein]) for the first term, and the assumption $\mathbf{a} \in \mathcal{T}_{Q^*}(\epsilon)$ for the second term, ϵ. As for the third term, we have the following consideration: Under the assumptions of the theorem, the maximizer of $H(Q)$ subject to the constraints $\rho(P,Q) \leq D$ and $\min_a Q(a) \geq q_{\min}$ is unique and continuous in P. Since $\mathcal{D}(P_{\mathbf{x}^*}\|P) \leq [(|\mathcal{A}|-1)\log e]/m$ as we have just shown, and hence the variational distance between $P_{\mathbf{x}^*}$ and P vanishes with m (by Pinsker's inequality [3, Lemma 11.6.1]), then so does the variational distance $\sum_a |Q_{\mathbf{x}^*}(a) - Q^*(a)| \stackrel{\Delta}{=} \delta_m$ as well. Now, similarly as in (8),

$$\max_a\log\frac{Q^*(a)}{Q_{\mathbf{x}^*}(a)} \leq \frac{\delta_m}{q_{\min}\ln 2}. \tag{A.11}$$

Putting all these facts together, we obtain

$$\begin{aligned}(i-1)\mathcal{D}(P_{\mathbf{a}}\|Q_{\mathbf{x}}) \quad\leq\quad & (|\mathcal{A}|-1)\log e + (i-1)\epsilon + n\epsilon^2 + \\ & (i-1)\frac{\delta_m}{q_{\min}\ln 2} + \xi_m,\end{aligned} \tag{A.12}$$

and so,

$$\begin{aligned}&\mathcal{D}(P_{\mathbf{a}}\|Q_{\mathbf{x}}) \\ \leq\ & \epsilon + \frac{\delta_m}{q_{\min}\ln 2} + \frac{(|\mathcal{A}|-1)\log e + n\epsilon^2 + \xi_m}{i-1} \\ \leq\ & \epsilon + \frac{\delta_m}{q_{\min}\ln 2} + \frac{(|\mathcal{A}|-1)\log e + n\epsilon^2 + \xi_m}{\epsilon n} \\ =\ & 2\epsilon + \frac{\delta_m}{q_{\min}\ln 2} + \frac{(|\mathcal{A}|-1)\log e + \xi_m}{\epsilon n}\end{aligned} \tag{A.13}$$

which is arbitrarily small for sufficiently small ϵ and sufficiently large n, since $\log m = o(n)$. Hence, also the variational distance between $Q_{\mathbf{x}}$ and $P_{\mathbf{a}}$ is bounded by a small quantity depending on ϵ, $q_{\min}$, ξ_m/n, and δ_m. In turn, the divergence (as well as the variational distance) between $P_{\mathbf{a}}$ and Q^* is bounded in terms of ϵ for $\mathbf{a} \in \mathcal{T}_{Q^*}(\epsilon)$. It follows then by the triangle inequality that the variational distance between $Q_{\mathbf{x}}$ and Q^* is upper bounded by the sum of these two terms. In particular, as claimed earlier, the maximizer and the minimizer in eqs. (A.6) and (A.8) are close to $Q^*(a_i)$, implying that the variational distance between $Q(\cdot|a^{i-1})$ and $Q^*(\cdot)$ is upper bounded in terms of the above terms as well as $\zeta_{n,m}/q_{\min}$. This bound, denoted $\eta(n,m,q_{\min},\epsilon)$, vanishes under the regime specified in the assertion of the lemma, which completes the proof of Lemma 1. □

REFERENCES

[1] B. S. Clarke and A. R. Barron, "Information–theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*, vol. 36, no. 3, pp. 453–471, May 1990.

[2] T. M. Cover and E. Ordentlich, "Universal portfolios with side information," *IEEE Trans. Inform. Theory*, vol. 42, no. 2, pp. 348–363, March 1996.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* John Wiley & Sons, Second Edition, 2006.

[4] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* New York: Academic, 1981.

[5] R. M. Gray, *Probability, Random Processes, and Ergodic Properties.* New York: Springer–Verlag, 1988.

[6] T. S. Han, M. Hoshi, "Interval algorithm for random number generation," *IEEE Trans. Inform. Theory*, vol. 43, pp. 599–611, March 1997.

[7] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Inform. Theory*, vol. IT–39, no. 3, pp. 752–772, May 1993.

[8] P. J. Huber, "Robust statistical procedures," *Society for Industrial and Applied Mathematics*, no. 27, 1977.

[9] S. A. Kassam and H. V. Poor, "Robust techniques in signal processing: a survey," *Proc. of the IEEE*, vol. 73, no. 3, pp. 433–481, March 1985.

[10] N. Merhav, "Achievable key rates for universal simulation of random data with respect to a set of statistical tests," *IEEE Trans. Inform. Theory*, vol. 50, no. 1, pp. 21–30, January 2004.

[11] N. Merhav, G. Seroussi, and M. J. Weinberger, "Universal delay–limited simulation," *Proc. ISIT 2005*, pp. 765–769, Adelaide, Australia, September 2005.

[12] N. Merhav and M. J. Weinberger, "On universal simulation of information sources using training data," *IEEE Trans. Inform. Theory*, vol. 50, no. 1, pp. 5–20, January 2004.

[13] N. Merhav and M. J. Weinberger, "Addendum to "On universal simulation of information sources using training data"," *IEEE Trans. Inform. Theory*, vol. 51, no. 9, pp. 3381–3383, September 2005.

[14] G. Seroussi, "On universal types," *IEEE Trans. Inform. Theory*, vol. 52, no. 1, pp. 171–189, January 2006.

[15] Y. Steinberg and S. Verdú, "Channel simulation and coding with side information," *IEEE Trans. Inform. Theory*, vol. IT–40, no. 3, pp. 634–646, May 1994.

[16] 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.

[17] K. Visweswariah, S. R. Kulkarni, and S. Verdú, "Separation of random number generation and resolvability," *IEEE Trans. Inform. Theory*, vol. 46, pp. 2237–2241, September 2000.