Achievability Results for Learning Under Communication Constraints

Maxim Raginsky
Department of Electrical and Computer Engineering
Duke University
Durham, NC 27708, USA
Email: m.raginsky@duke.edu

***Abstract*— The problem of statistical learning is to construct an accurate predictor of a random variable as a function of a correlated random variable on the basis of an i.i.d. training sample from their joint distribution. Allowable predictors are constrained to lie in some specified class, and the goal is to approach asymptotically the performance of the best predictor in the class. We consider two settings in which the learning agent only has access to rate-limited descriptions of the training data, and present information-theoretic bounds on the predictor performance achievable in the presence of these communication constraints. Our proofs do not assume any separation structure between compression and learning and rely on a new class of operational criteria specifically tailored to joint design of encoders and learning algorithms in rate-constrained settings. These operational criteria naturally lead to a learning-theoretic generalization of the rate-distortion function introduced recently by Kramer and Savari in the context of rate-constrained communication of probability distributions.**

I. Introduction

Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be jointly distributed random variables. The problem of statistical learning is to design an accurate predictor of the *output variable* Y from the *input variable* X on the basis of a number of independent *training samples* drawn from their joint distribution, with very little or no prior knowledge of that distribution. The present paper focuses on the achievable performance of learning schemes when the learning agent only has access to a finite-rate description of the training samples.

This problem of *learning under communication constraints* arises in a variety of contexts, such as distributed estimation using a sensor network, adaptive control, or repeated games. In these and other scenarios, it is often the case that the agents who gather the training data are geographically separated from the agents who use these data to make inferences and decisions, and communication between these two types of agents is possible only over rate-limited channels. Hence, there is a trade-off between the communication rate and the quality of the inference, and it is of interest to characterize this trade-off mathematically.

This paper follows on our earlier work [1] and presents improved bounds on the achievable performance of statistical learning schemes operating under two kinds of communication constraints: (a) the entire training sequence is delivered to the learning agent over a rate-limited noiseless digital channel, and (b) the input part of the training sequence is available to the learning agent with arbitrary precision, while the output part is delivered, as before, over a rate-limited channel. Whereas [1] has looked at schemes where the finite-rate description of the training data was obtained through vector quantization, effectively imposing a separation structure between compression and learning, here we remove this restriction.

We show that, under certain regularity conditions, there is no penalty for compression of the training sequence in the setting (a). This is due to the fact that the encoder can reliably estimate the underlying distribution (in the metric specifically tailored for the learning problem at hand) and then communicate the finite-rate description to the learning agent, who can then find the optimum predictor for the estimated distribution. The setting (b), however, is radically different: because the encoder has no access to the input part of the training sample, it cannot estimate the underlying distribution. Instead, the encoder constructs a finite-rate description of the output part using a specific kind of a vector quantizer, namely one designed to minimize the expected distance between the underlying distribution (whatever it may happen to be) and the empirical distribution of the input/quantized output pairs. Our achievability result for the setting (b) uses a learning-theoretic generalization of recent work by Kramer and Savari [2] on rate-constrained communication of probability distributions.

The problem of learning a pattern classifier under rate constraints was also treated in a recent paper by Westover and O'Sullivan [3]. They assumed that the underlying probability distribution is known, and the rate constraint arises from the limitations on the memory of the learning agent; then the problem is to design the best possible classifier (without any constraints on its structure). The motivation for the work in [3] comes from biologically inspired models of learning. The approach of the present paper is complementary to that of [3]. We consider a more general, decision-theoretic formulation of learning that includes regression as well as classification, but allow only vague prior knowledge of the underlying distribution and assume that the class of available predictors is constrained. Thus, while [3] presents information-theoretic bounds on the performance of *any* classifier (including ones that are fully cognizant of the generative model for the data), here we are concerned with the performance of constrained learning schemes that must perform well in the presence of uncertainty about the underlying distribution.

The novel element of our approach is that both the operational criteria used to design the encoders and the learning algorithm, and the regularity conditions that must hold for rate-constrained learning to be possible, involve a tight coupling between the available prior knowledge about the underlying distribution and the set of predictors available to the learning agent. Planned future work includes obtaining converse theorems (lower bounds) and applying our formalism to specific classes of predictors used in statistical learning theory.

II. Preliminaries and problem formulation

A very general decision-theoretic formulation of the learning problem, due to Haussler [4], goes as follows. We have a family $\mathcal{P}$ of probability distributions on $\mathcal{Z} \stackrel{\triangle}{=} \mathcal{X} \times \mathcal{Y}$ and a class $\mathcal{F}$ of measurable functions $f : \mathcal{Z} \to \mathbb{R}$. For any $P \in \mathcal{P}$, define

$$L(f, P) \stackrel{\triangle}{=} \mathbb{E}_P[f(Z)] \equiv \int_{\mathcal{Z}} f(z) dP(z), \qquad f \in \mathcal{F}$$

and

$$L^*(\mathcal{F}, P) \stackrel{\triangle}{=} \inf_{f \in \mathcal{F}} L(f, P),$$

where we assume that the infimum is achieved by some $f^* \in \mathcal{F}$. The family $\mathcal{P}$ represents prior knowledge about the joint distribution of X and Y; each function $f \in \mathcal{F}$ corresponds to the loss incurred by a particular predictor of Y based on X. This framework covers, for instance, the following standard scenarios:

- *classification* — $\mathcal{X} \subseteq \mathbb{R}^d$, $\mathcal{Y} = \{1, \ldots, M\}$, and $\mathcal{F}$ consists of functions of the form

  $$f(x, y) = I_{\{g(x) \neq y\}}, \qquad g \in \mathcal{G}$$

  where $I_{\{\cdot\}}$ is the indicator function, and $\mathcal{G}$ is a given family of *classifiers*, i.e., measurable functions $g : \mathcal{X} \to \{1, \ldots, M\}$. Any $f^* \in \mathcal{F}$ that achieves $L^*(\mathcal{F}, P)$ corresponds to some $g^* \in \mathcal{G}$ that has the smallest classification error:

  $$P(g^*(X) \neq Y) = \inf_{g \in \mathcal{G}} P(g(X) \neq Y).$$

- *regression* — $\mathcal{X} \subseteq \mathbb{R}^d$, $\mathcal{Y} \subseteq \mathbb{R}$, and $\mathcal{F}$ consists of functions of the form

  $$f(x, y) = (g(x) - y)^2, \qquad g \in \mathcal{G}$$

  where $\mathcal{G}$ is a given family of *estimators*, i.e., measurable functions $g : \mathcal{X} \to \mathbb{R}$. Any $f^* \in \mathcal{F}$ that achieves $L^*(\mathcal{F}, P)$ corresponds to some $g^* \in \mathcal{G}$ that has the smallest mean squared error:

  $$\mathbb{E}_P[(g^*(X) - Y)^2] = \inf_{g \in \mathcal{G}} \mathbb{E}_P[(g(X) - Y)^2].$$

These are instances of *supervised learning* problems. *Unsupervised* settings, where $\mathcal{Y} = \varnothing$ (such as density estimation or clustering), can also be accommodated by Haussler's framework. In this paper we focus only on the supervised case; thus, we will assume that $|\mathcal{Y}| \geq 2$. Then the learning problem is to construct, for each $n \in \mathbb{N}$, an approximation to f^* on the basis of a *training sequence* $Z^n = \{Z_i\}_{i=1}^n$, where $Z_i = (X_i, Y_i)$ are i.i.d. according to some unknown $P \in \mathcal{P}$.

Formally, a *learning scheme* (or *learner*, for short) is a sequence $\{\widehat{f}_n\}_{n=1}^\infty$ of maps $\widehat{f}_n : \mathcal{Z}^n \times \mathcal{Z} \to \mathbb{R}$, such that $\widehat{f}_n(z^n, \cdot) \in \mathcal{F}$ for all $z^n \in \mathcal{Z}^n$. Let $Z = (X, Y) \sim P$ be independent of the training sequence Z^n. The main quantity of interest is the *generalization error*

$$L(\widehat{f}_n, P) = \mathbb{E}\left[\widehat{f}_n(Z^n, Z)\Big|Z^n\right] \equiv \int_{\mathcal{Z}} \widehat{f}_n(Z^n, z) dP(z),$$

which is a random variable that depends on the training sequence Z^n. Under suitable regularity conditions on $\mathcal{P}$ and $\mathcal{F}$, one can show that there exist learning schemes that are *probably approximately correct* (PAC), i.e., for every $\epsilon > 0$ and $P \in \mathcal{P}$,

$$\lim_{n \to \infty} P\left(Z^n : L(\widehat{f}_n, P) > L^*(\mathcal{F}, P) + \epsilon\right) = 0 \qquad (2.1)$$

(see, e.g., Vidyasagar [5]). A more modest goal is to ensure that the *excess loss* $L(\widehat{f}_n, P) - L^*(\mathcal{F}, P)$ is small, either in probability or in expectation.

We are interested in the achievable excess loss in situations where there is a rate-constrained channel between the source of the training data and the learning agent. Specifically, we shall consider the following two scenarios, depicted in Figs. 1 and 2, respectively.

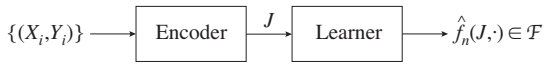

Fig. 1. Type I set-up: the encoder has full observation of the training samples.

In the first set-up, shown in Fig. 1, the learner observes the training data through a noiseless digital channel that can transmit a fixed finite number of bits per training pair $Z = (X, Y)$. A scheme for learning operating at rate R is specified by a sequence $\{(e_n, \widehat{f}_n)\}_{n=1}^\infty$, where

$$e_n : \mathcal{Z}^n \to \{1, 2, \ldots, M_n\}$$

is the *encoder* and

$$\widehat{f}_n : \{1, 2, \ldots, M_n\} \to \mathcal{F}$$

is the *learner*, such that

$$\limsup_{n \to \infty} n^{-1} \log M_n \leq R.$$

For each n, the output of the learner is a function $\widehat{f}_n(J, \cdot) \in \mathcal{F}$, where $J = e_n(Z^n)$ is the finite-rate description of Z^n provided by the encoder. We shall refer to this as Type I set-up.

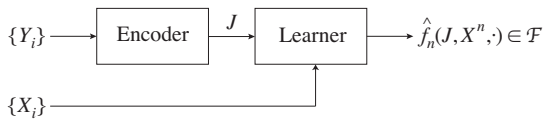

Fig. 2. Type II: the encoder sees only the output part of the training sequence.

In the second set-up, shown in Fig. 2, the learner has perfect observation of the input ($\mathcal{X}$-valued) part of the training sequence, while the output ($\mathcal{Y}$-valued part) is delivered over a rate-limited noiseless digital channel. A scheme for learning operating at rate R is a sequence $\{(e_n, \widehat{f}_n)\}_{n=1}^{\infty}$, where

$$e_n : \mathcal{Y}^n \to \{1, 2, \ldots, M_n\}$$

is the encoder and

$$\widehat{f}_n : \mathcal{X}^n \times \{1, 2, \ldots, M_n\} \to \mathcal{F}$$

is the learner, such that

$$\limsup_{n\to\infty} n^{-1} \log M_n \leq R.$$

For each n, the output of the learner is a function $\widehat{f}_n(J, X^n, \cdot) \in \mathcal{F}$, where $J = e_n(Y^n)$ is the finite-rate description of Y^n provided by the encoder.

We shall often abuse notation and let $\widehat{f}_n$ denote also the function in $\mathcal{F}$ returned by the learner. The main object of interest is the generalization error

$$L(e_n, \widehat{f}_n, P) \stackrel{\triangle}{=} \mathbb{E}\left[\widehat{f}_n(W_n, Z)\Big| Z^n\right], \qquad P \in \mathcal{P}$$

where $Z = (X, Y) \sim P$ is assumed independent of $\{Z_i\}_{i=1}^n$, and W_n is equal to $J = e_n(Z^n)$ in a Type I set-up and to (J, X^n) in a Type II set-up, where $J = e_n(Y^n)$. We are interested in the achievable values of the asymptotic expected excess loss. We say that a pair (R, Δ) is *achievable for* $(\mathcal{F}, \mathcal{P})$ if there exists a scheme $\{(e_n, \widehat{f}_n)\}_{n=1}^{\infty}$ operating at rate R, such that

$$\limsup_{n\to\infty} \mathbb{E}\, L(e_n, \widehat{f}_n, P) \leq L^*(\mathcal{F}, P) + \Delta, \qquad \forall P \in \mathcal{P}.$$

III. Achievability theorems

In this section, we prove two theorems about achievable pairs (R, Δ) in Type I and Type II settings. The key idea in both cases is that the encoder needs to provide enough information at rate R for the learner to estimate the expected value of each $f \in \mathcal{F}$ to within Δ.

A. Notation, preliminaries and assumptions

We assume that the space $\mathcal{Z}$ is equipped with an appropriate σ-algebra $\mathcal{A}$. Typical cases of interest in learning theory are $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y}$ finite (classification) or $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$ (regression), with the usual Borel σ-algebras. The space of all probability measures on $(\mathcal{Z}, \mathcal{A})$ will be denoted by $\mathcal{M}(\mathcal{Z})$. $\mathcal{F}$ is a class of measurable functions from $(\mathcal{Z}, \mathcal{A})$ into $[0, B]$ for some $0 < B < +\infty$; to avoid various measurability issues, we also assume throughout that $\mathcal{F}$ is countable. We shall identify signed measures μ on $(\mathcal{Z}, \mathcal{A})$ with real-valued linear functionals $f \mapsto \mu(f)$ on $\mathcal{F}$, where $\mu(f) \stackrel{\triangle}{=} \int_{\mathcal{Z}} f d\mu$. Thus, to each μ we can associate the $\ell^\infty(\mathcal{F})$-norm

$$\|\mu\|_{\mathcal{F}} \stackrel{\triangle}{=} \sup_{f\in\mathcal{F}} |\mu(f)|.$$

For an n-tuple $z^n \in \mathcal{Z}^n$, P_{z^n} will denote the corresponding empirical measure: $\mathsf{P}_{z^n} = n^{-1}\sum_{i=1}^n \delta_{z_i}$, where δ_z is the Dirac measure (point mass) concentrated at $z \in \mathcal{Z}$. We assume that $\mathcal{F}$ is a *Glivenko–Cantelli (GC) class* [6], i.e.,

$$\lim_{n\to\infty} \|\mathsf{P}_{Z^n} - P\|_{\mathcal{F}} = 0, \qquad \text{a.s.} \tag{3.2}$$

for every $P \in \mathcal{M}(\mathcal{Z})$. In other words, the class $\mathcal{F}$ is such that, for each $P \in \mathcal{M}(\mathcal{Z})$, the sample averages $\mathsf{P}_{Z^n}(f)$ converge to the theoretical averages $P(f)$ uniformly over $\mathcal{F}$. This is a standard assumption in statistical learning theory. In particular, if $\mathcal{F}$ is a GC class, then the well-known Empirical Risk Minimization (ERM) algorithm, defined by

$$\widehat{f}_n^{(\mathrm{ERM})} \stackrel{\triangle}{=} \arg\min_{f\in\mathcal{F}} \sum_{i=1}^n f(Z_i), \tag{3.3}$$

is PAC in the sense of (2.1) [5], [6].

B. Type I schemes

We now show that, in a Type I set-up, there is no penalty for compression of the training sequence, provided the family $\mathcal{P}$ is not too "rich." Our notion of richness will pertain to the geometry of $\mathcal{P}$ w.r.t. the $\|\cdot\|_{\mathcal{F}}$ norm. Given some $\epsilon > 0$, we say that a finite set $\{P_1, \ldots, P_M\} \subset \mathcal{P}$ is an *ϵ-net* for $\mathcal{P}$ if

$$\sup_{P\in\mathcal{P}} \min_{1\leq m\leq M} \|P - P_m\|_{\mathcal{F}} \leq \epsilon.$$

We define the *covering number* $N_{\mathcal{F}}(\epsilon, \mathcal{P})$ as the cardinality of the minimal ϵ-net of $\mathcal{P}$, and the *Kolmogorov ϵ-entropy* of $\mathcal{P}$ as $H_{\mathcal{F}}(\epsilon, \mathcal{P}) \stackrel{\triangle}{=} \log N_{\mathcal{F}}(\epsilon, \mathcal{P})$ [7].

Theorem 3.1. Suppose that there exists a monotone decreasing sequence $\{\epsilon_n\}_{n=1}^{\infty}$ of nonnegative reals, such that

$$H_{\mathcal{F}}(\epsilon_n, \mathcal{P}) = o(n). \tag{3.4}$$

Then the pair $(0, 0)$ is achievable for $(\mathcal{F}, \mathcal{P})$.

Proof: For each n, let $\mathcal{N}_n = \{P_1, P_2, \ldots, P_{M_n}\}$ be the minimal ϵ_n-net for $\mathcal{P}$ w.r.t. $\|\cdot\|_{\mathcal{F}}$, where $M_n = N_{\mathcal{F}}(\epsilon_n, \mathcal{P})$. Consider the following scheme:

- *encoder* — $e_n(Z^n) = \arg\min_{1\leq m\leq M_n} \|\mathsf{P}_{Z^n} - P_m\|_{\mathcal{F}}$
- *learner* — $\widehat{f}_n(J, \cdot) = \arg\min_{f\in\mathcal{F}} P_J(f)$

In other words, the encoder finds the element of $\mathcal{N}_n$ closest to the empirical distribution P_{Z^n} in the $\|\cdot\|_{\mathcal{F}}$ norm and transmits its index to the learner. The learner then finds the function in $\mathcal{F}$ that minimizes the expected loss assuming that the true distribution is the one estimated by the encoder.

It is easy to see that the resulting scheme operates at zero rate. Indeed, from (3.4),

$$\lim_{n\to\infty} \frac{\log M_n}{n} = \lim_{n\to\infty} \frac{H_{\mathcal{F}}(\epsilon_n, \mathcal{P})}{n} = 0.$$

To bound the expected loss, assume that $P \in \mathcal{P}$ is the true distribution and let $P_{m^*} \in \mathcal{N}_n$ be the element of the ϵ_n-net that is closest to P, i.e.,

$$\|P - P_{m^*}\|_{\mathcal{F}} = \min_{1\leq m\leq M} \|P - P_m\|_{\mathcal{F}} \leq \epsilon_n.$$

Let $J = e_n(Z^n)$. We then have

$$\begin{aligned}
L(e_n, \widehat{f}_n, P) &= P(\widehat{f}_n) \\
&\le \|P - P_J\|_{\mathcal{F}} + P_J(\widehat{f}_n) \\
&= \|P - P_J\|_{\mathcal{F}} + L^*(\mathcal{F}, P_J) \\
&\overset{(a)}{\le} 2\|P - P_J\|_{\mathcal{F}} + L^*(\mathcal{F}, P) \\
&\le 2\|P - \mathsf{P}_{Z^n}\|_{\mathcal{F}} + 2\|\mathsf{P}_{Z^n} - P_J\|_{\mathcal{F}} + L^*(\mathcal{F}, P) \\
&\overset{(b)}{\le} 2\|P - \mathsf{P}_{Z^n}\|_{\mathcal{F}} + 2\|\mathsf{P}_{Z^n} - P_{m^*}\|_{\mathcal{F}} + L^*(\mathcal{F}, P) \\
&\le 4\|P - \mathsf{P}_{Z^n}\|_{\mathcal{F}} + 2\|P - P_{m^*}\|_{\mathcal{F}} + L^*(\mathcal{F}, P) \\
&\le 4\|P - \mathsf{P}_{Z^n}\|_{\mathcal{F}} + 2\epsilon_n + L^*(\mathcal{F}, P), \qquad (3.5)
\end{aligned}$$

where (a) follows from the fact that

$$|L^*(\mathcal{F}, P) - L^*(\mathcal{F}, P')| \le \|P - P'\|_{\mathcal{F}}$$

for any two $P, P' \in \mathcal{P}$, and (b) is by construction of the encoder. The remaining steps are consequences of various definitions and the triangle inequality. Taking expectations and the limit as $n \to \infty$, we get

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{E}\, L(e_n, \widehat{f}_n, P) \\
&\quad\le \quad 4 \lim_{n\to\infty} \mathbb{E}\, \|\mathsf{P}_{Z^n} - P\|_{\mathcal{F}} + 2 \lim_{n\to\infty} \epsilon_n + L^*(\mathcal{F}, P).
\end{aligned}$$

The first limit on the right-hand side of this inequality is zero by the GC property, while the second one is zero since $\epsilon_n \to 0$. Thus, $\lim_{n\to\infty} \mathbb{E}\, L(e_n, \widehat{f}_n, P) \le L^*(\mathcal{F}, P)$. ∎

We can give one particular example when condition (3.4) will hold. Given any two probability measures P, Q on $(\mathcal{Z}, \mathcal{A})$, define the *variational distance* between them as

$$\|P - Q\|_V \triangleq \sup_{\{A_i\} \subseteq \mathcal{A}} \sum_i |P(A_i) - Q(A_i)|,$$

where the supremum is over all finite $\mathcal{A}$-measurable partitions of $\mathcal{Z}$. Then we can define the covering numbers $N_V(\epsilon, \mathcal{P})$ and the Kolmogorov ϵ-entropy $H_V(\epsilon, \mathcal{P})$. Now suppose that there exist some constants $C > 0$ and $k > 0$, such that

$$N_V(\epsilon, \mathcal{P}) \le C(1/\epsilon)^k$$

for small enough ϵ. This will be the case, for instance, when $\mathcal{P}$ is a parametric family of distributions, $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^k$ is a compact set. Then, since

$$\|P - P'\|_{\mathcal{F}} \le B\|P - P'\|_V$$

for all $P, P' \in \mathcal{P}$, we will have

$$N_V(\epsilon, \mathcal{P}) \le C'(1/\epsilon)^k$$

with $C' = C'(C, B, k)$. Then, choosing $\epsilon_n = 1/n$, we will have

$$\log H_{\mathcal{F}}(\epsilon_n, \mathcal{P}) \le k \log n + C' = o(n),$$

as required by the theorem. Moreover, when the function class $\mathcal{F}$ is sufficiently regular (e.g., a Vapnik–Chervonenkis subgraph class [6]), we can identify explicitly the rate at which the expected generalization error $\mathbb{E}\, L(e_n, \widehat{f}_n, P)$ converges to $L^*(\mathcal{F}, P)$ as $n \to \infty$. For such a class $\mathcal{F}$, we will have

$$\mathbb{E}\, \|\mathsf{P}_{Z^n} - P\|_{\mathcal{F}} \le \frac{C_{\mathcal{F}}}{\sqrt{n}},$$

where $C_{\mathcal{F}} > 0$ is some constant that depends on $\mathcal{F}$. Using this and (3.5), we obtain

$$\sup_{P\in\mathcal{P}} \mathbb{E}\left[L(e_n, \widehat{f}_n, P) - L^*(\mathcal{F}, P)\right] \le \frac{4C_{\mathcal{F}}}{\sqrt{n}} + \frac{2}{n} = O\left(\frac{1}{\sqrt{n}}\right).$$

This is the same rate of convergence one would obtain without any compression. In fact, without making additional assumptions on $\mathcal{F}$ beyond it being a VC subgraph class or something similar, this is the best convergence rate possible. Thus, in a Type I setting, we can get the same learning rates as in the infinite-precision ($R = \infty$) scenario. When $\mathcal{P}$ is a nonparametric class, we get slower rates. For example, if $H_V(\epsilon, \mathcal{P}) \le C(1/\epsilon)^\alpha$ for some $\alpha > 0$ (as would be the case when $\mathcal{Z}$ is a compact subset of a Euclidean space, all distributions in $\mathcal{P}$ have Lipschitz-continuous densities w.r.t. a common dominating measure, and their Lipschitz constants are uniformly bounded [7]), then we can choose $\epsilon_n = (1/\log n)^\alpha$.

C. Type II schemes

The case of Type II schemes is radically different. Whereas in a Type I scheme the encoder can use the training data to estimate the underlying distribution and then communicate its finite-rate description to the learner, in a Type II situation the encoder can only estimate the Y-marginal. Unless the distributions in $\mathcal{P}$ can be reliably identified from their Y-marginals (which is a very restrictive condition), the encoder does not have enough "learning" ability to estimate the underlying distribution. Instead, we will take the following approach.

Given $\Delta \ge 0$, let us suppose that, for each n, the encoder can implement a mapping $Y^n \mapsto \widehat{Y}^n$, such that, whenever the training data are drawn from some $P \in \mathcal{P}$ (unknown to both the encoder and the learner), the empirical distribution $\mathsf{P}_{(X^n, \widehat{Y}^n)}$ is, on average, at most $\Delta/4$ away from P in the $\|\cdot\|_{\mathcal{F}}$ sense, and that $n^{-1} \log |\widehat{Y}^n(\mathcal{Y}^n)| \le R$. Then the encoder communicates a binary description J of $\widehat{Y}^n$ at rate $\le R$ to the learning agent, who decodes it to get $\widehat{Y}^n$ and then implements the following two-step procedure:

$$\begin{aligned}
\widehat{P} &= \arg\min_{P\in\mathcal{P}} \|\mathsf{P}_{(X^n, \widehat{Y}^n)} - P\|_{\mathcal{F}}, \\
\widehat{f}_n &= \arg\min_{f\in\mathcal{F}} \widehat{P}(f).
\end{aligned}$$

Then essentially the same technique as in the proof of Theorem 3.1 will give us $\mathbb{E}\, L(e_n, \widehat{f}_n, P) \le L^*(\mathcal{F}, P) + \Delta$ for every $P \in \mathcal{P}$, thus establishing the existence of a scheme operating at rate R and achieving an excess loss of $\le \Delta$ on each $P \in \mathcal{P}$.

These considerations motivate the definition of the following nth-order operational distortion-rate function:

$$\widehat{\mathbb{D}}_n(\mathcal{P}, \mathcal{F}, R) \triangleq \inf_{\widehat{Y}^n} \sup_{P\in\mathcal{P}} \mathbb{E}_P\, \|\mathsf{P}_{(X^n, \widehat{Y}^n(Y^n))} - P\|_{\mathcal{F}}, \qquad (3.6)$$

where the infimum is over all $\widehat{Y}^n : \mathcal{Y}^n \to \mathcal{Y}^n$, such that $n^{-1}\log|\{\widehat{Y}^n(y^n) : y^n \in \mathcal{Y}^n\}| \le R$. We also define the limiting operational distortion-rate function

$$\widehat{\mathbb{D}}(\mathcal{P},\mathcal{F},R) \triangleq \lim_{n\to\infty} \widehat{\mathbb{D}}_n(\mathcal{P},\mathcal{F},R).$$

We now state the achievability result for Type II schemes in terms of these operational quantities:

Theorem 3.2. Given any $R \ge 0$, the pair $(R, 4\widehat{\mathbb{D}}(\mathcal{P},\mathcal{F},R))$ is achievable.

Proof: For each n, let $\widehat{Y}^n_* : \mathcal{Y}^n \to \mathcal{Y}^n$ be the encoder that achieves the infimum in (3.6). Let $\{\widehat{y}^n(1),\ldots,\widehat{y}^n(M_n)\}$ be some arbitrary enumeration of its codewords. Then we construct the following scheme:

- *encoder* — $e_n(Y^n) = J$, such that $\widehat{Y}^n_*(Y^n) = \widehat{y}^n(J)$.
- *learner* — $\widehat{f}_n(J, X^n, \cdot) = \arg\min_{f\in\mathcal{F}} \widehat{P}(f)$, where

$$\widehat{P} = \arg\min_{P\in\mathcal{P}} \|\mathsf{P}_{(X^n,\widehat{y}^n(J))} - P\|_{\mathcal{F}}.$$

The scheme $\{(e_n,\widehat{f}_n)\}_{n=1}^{\infty}$ operates at rate R owing to the fact that $n^{-1}\log M_n \le R$. As for the excess loss, we have

$$\begin{aligned}
L(e_n,\widehat{f}_n,P) &= P(\widehat{f}_n)\\
&\le 2\|P-\widehat{P}\|_{\mathcal{F}} + L^*(\mathcal{F},P)\\
&\le 2\|P-\mathsf{P}_{(X^n,\widehat{y}^n(J))}\|_{\mathcal{F}}\\
&\qquad +2\|\mathsf{P}_{(X^n,\widehat{y}^n(J))}-\widehat{P}\|_{\mathcal{F}} + L^*(\mathcal{F},P)\\
&\le 4\|P-\mathsf{P}_{(X^n,\widehat{y}^n(J))}\|_{\mathcal{F}} + L^*(\mathcal{F},P)\\
&= 4\|P-\mathsf{P}_{(X^n,\widehat{Y}^n_*(Y^n))}\|_{\mathcal{F}} + L^*(\mathcal{F},P).
\end{aligned}$$

Taking expectations, using the fact that each $\widehat{Y}^n_*$ achieves the nth-order optimum $\widehat{\mathbb{D}}_n(\mathcal{P},\mathcal{F},R)$, and then taking the limit as $n\to\infty$, we get

$$\mathbb{E}\, L(e_n,\widehat{f}_n,P) \le L^*(\mathcal{F},P) + 4\widehat{\mathbb{D}}(\mathcal{P},\mathcal{F},R), \qquad \forall P\in\mathcal{P}$$

which proves the theorem. ∎

We would like to express $\widehat{\mathbb{D}}(\mathcal{P},\mathcal{F},R)$ purely in terms of information-theoretic quantities. It is relatively straightforward to derive an information-theoretic lower bound on $\widehat{\mathbb{D}}(\mathcal{P},\mathcal{F},R)$. To that end, we will draw upon recent work of Kramer and Savari [2] on rate-constrained communication of probability distributions. The following properties of $\|\cdot\|_{\mathcal{F}}$ are immediate:

1) $\|P-Q\|_{\mathcal{F}} \le 2B$ for all $P,Q\in\mathcal{M}(\mathcal{Z})$.
2) For a fixed P, the mapping $Q\mapsto\|Q-P\|_{\mathcal{F}}$ is Lipschitz in the variational norm $\|\cdot\|_V$: for all $Q,Q'\in\mathcal{M}(\mathcal{Z})$

$$|\|P-Q\|_{\mathcal{F}} - \|P-Q'\|_{\mathcal{F}}| \le B\|Q-Q'\|_V.$$

3) The mapping $Q\mapsto\|Q-P\|_{\mathcal{F}}$ is convex: for any $Q=\lambda Q_1+(1-\lambda)Q_2$ with some $\lambda\in[0,1]$ and $Q_1,Q_2\in\mathcal{M}(\mathcal{Z})$,

$$\|Q-P\|_{\mathcal{F}} \le \lambda\|Q_1-P\|_{\mathcal{F}} + (1-\lambda)\|Q_2-P\|_{\mathcal{F}}.$$

Then for each $P\in\mathcal{P}$ the mapping $Q\in\mathcal{M}(\mathcal{Z})\mapsto\|Q-P\|_{\mathcal{F}}$ satisfies the requirements listed in Section III of [2]. Thus, following Kramer and Savari, we can define, for every $P\in\mathcal{P}$ and every $R\ge 0$, the distortion-rate function

$$D_{\mathrm{KS}}(P,\mathcal{F},R) \triangleq \inf\|P_{XU}-P\|_{\mathcal{F}}, \tag{3.7}$$

where the infimum is over all distributions of the triple $(X,Y,U)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}$, such that

$$P_{XY}=P, \quad X\to Y\to U, \quad I(Y;U)\le R.$$

Here, $X\to Y\to U$ denotes that X, Y, and U form a Markov chain, i.e., X and U are conditionally independent given Y. Kramer and Savari deal only with the case when $\mathcal{X}$ and $\mathcal{Y}$ are both finite. Their results can be generalized to the case when $\mathcal{X}$ and $\mathcal{Y}$ are *standard* measurable spaces [8], so that regular versions of conditional probability distributions exist. Specifically, assume that $\mathcal{P}$ is a singleton, $\mathcal{P}=\{P\}$ for some $P\in\mathcal{M}(\mathcal{X}\times\mathcal{Y})$. Then we have the following:

Theorem 3.3.

$$\widehat{\mathbb{D}}(P,\mathcal{F},R) \equiv \widehat{\mathbb{D}}(\{P\},\mathcal{F},R) \le D_{\mathrm{KS}}(P,\mathcal{F},R)$$

Proof: Let (X,Y,U) achieve the infimum in (3.7). For each n, define the function $\xi_n : \mathcal{X}^n\times\mathcal{Y}^n\to[0,B]$ via

$$\xi_n(x^n,u^n) \triangleq \|\mathsf{P}_{(x^n,u^n)} - P_{XU}\|_{\mathcal{F}}. \tag{3.8}$$

Since $\mathcal{F}$ is a GC class (cf. Eq. (3.2)), we have

$$\lim_{n\to\infty}\mathbb{E}\,\xi_n(X^n,U^n) = 0.$$

Hence, we can apply Lemma A.1 in the Appendix to show that, for any $\epsilon>0$ there exists a sequence $\{\widehat{Y}^n\}_{n=1}^{\infty}$ of mappings $\widehat{Y}^n : \mathcal{Y}^n\to\mathcal{Y}^n$ and some $n_0 = n_0(\epsilon)$, such that for all $n\ge n_0$,

$$\frac{1}{n}\log\left|\left\{\widehat{Y}^n(y^n) : y^n\in\mathcal{Y}^n\right\}\right| \le I(Y;U)+\epsilon \le R+\epsilon$$

and

$$\mathbb{E}\,\|\mathsf{P}_{(X^n,\widehat{Y}^n(Y^n))} - P_{XU}\|_{\mathcal{F}} \le \epsilon.$$

Hence,

$$\begin{aligned}
&\mathbb{E}\,\|\mathsf{P}_{(X^n,\widehat{Y}^n(Y^n))} - P\|_{\mathcal{F}}\\
&\quad\le \|P_{XU}-P\|_{\mathcal{F}} + \mathbb{E}\,\|\mathsf{P}_{(X^n,\widehat{Y}^n(Y^n))} - P_{XU}\|_{\mathcal{F}}\\
&\quad\le D_{\mathrm{KS}}(P,\mathcal{F},R)+\epsilon.
\end{aligned}$$

The theorem is proved. ∎

Moreover, when $|\mathcal{P}|\ge 2$, we have the following lower bound:

Theorem 3.4.

$$\widehat{\mathbb{D}}(\mathcal{P},\mathcal{F},R) \ge \sup_{P\in\mathcal{P}} D_{\mathrm{KS}}(P,\mathcal{F},R)$$

Proof: Fix any code $\widehat{Y}^n(\cdot)$ of rate R that achieves $\widehat{\mathbb{D}}_n(\mathcal{P},\mathcal{F},R)$:

$$\sup_{P\in\mathcal{P}}\mathbb{E}_P\,\|\mathsf{P}_{(X^n,\widehat{Y}^n(Y^n))} - P\|_{\mathcal{F}} = \widehat{\mathbb{D}}_n(\mathcal{P},\mathcal{F},R).$$

Fix some $P\in\mathcal{P}$ and let $P_{X_i,Y_i,\widehat{Y}_i}$ denote the joint distribution of $(X_i,Y_i,\widehat{Y}_i)$ when $(X_1,Y_1),\ldots,(X_n,Y_n)$ are i.i.d. according to P, and $\widehat{Y}_i$ denotes the ith component of $\widehat{Y}^n(Y^n)$. Also,

$$\widehat{\mathbb{D}}(\mathcal{P},\mathcal{F},R) \le \sup_{\alpha>0}\inf_{\delta>0}\sup_{P'\in\mathcal{M}(\mathcal{Y})}\inf_{\substack{Q_{U|Y}:\\ I(P'\times Q_{U|Y})\le R+\alpha}}\sup_{\substack{P\in\mathcal{P}:\\ \|P_Y-P'\|_V\le\delta}}\mathbb{E}_{P\times Q_{U|Y}}\,\|\delta_{(X,U)}-P\|_{\mathcal{F}}. \tag{3.10}$$

define the random variables $\bar{X}\in\mathcal{X}$, $\bar{Y}\in\mathcal{Y}$, and $\bar{U}\in\mathcal{Y}$ with the joint distribution

$$P_{\bar{X},\bar{Y},\bar{U}} \triangleq \frac{1}{n}\sum_{i=1}^{n} P_{X_i,Y_i,\widehat{Y}_i}.$$

Then $P_{\bar{X},\bar{Y}}=P$ and that $\bar{X}\to\bar{Y}\to\bar{U}$. Using convexity and the fact that $\mathbb{E}\sup_{f\in\mathcal{F}}[\cdot]\ge\sup_{f\in\mathcal{F}}E[\cdot]$, we get

$$\begin{aligned}
&\mathbb{E}_P\,\|\mathsf{P}_{(X^n,\widehat{Y}^n)}-P\|_{\mathcal{F}}\\
&\quad= \mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}f(X_i,\widehat{Y}_i)-P(f)\right|\\
&\quad\ge \sup_{f\in\mathcal{F}}\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n}f(X_i,\widehat{Y}_i)-P(f)\right|\\
&\quad\ge \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\,f(X_i,\widehat{Y}_i)-P(f)\right|\\
&\quad= \sup_{f\in\mathcal{F}}|P_{\bar{X},\bar{U}}(f)-P(f)|\\
&\quad= \|P_{\bar{X},\bar{U}}-P\|_{\mathcal{F}}.
\end{aligned}$$

That is, $\|P_{\bar{X},\bar{U}}-P\|_{\mathcal{F}}\le\widehat{\mathbb{D}}_n(\mathcal{P},\mathcal{F},R)$ for all $P\in\mathcal{P}$. Moreover, steps similar to those in [2, Thm. 1] give

$$\begin{aligned}
nR &\ge H(\widehat{Y}^n)\\
&\ge I(Y^n;\widehat{Y}^n)\\
&= H(Y^n)-H(Y^n|\widehat{Y}^n)\\
&= \sum_{i=1}^{n}[H(Y_i|Y^{i-1})-H(Y_i|\widehat{Y}^n,Y^{i-1})]\\
&= \sum_{i=1}^{n}[H(Y_i)-H(Y_i|\widehat{Y}^n,Y^{i-1})]\\
&\ge \sum_{i=1}^{n}[H(Y_i)-H(Y_i|\widehat{Y}_i)]\\
&= \sum_{i=1}^{n}I(Y_i;\widehat{Y}_i)\\
&\ge nI(\bar{Y};\bar{U}).
\end{aligned}$$

Thus, we have found a triple of random variables $(\bar{X},\bar{Y},\bar{U})\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}$, such that:

1) $P_{\bar{X},\bar{Y}}=P$,
2) $\bar{X}\to\bar{Y}\to\bar{U}$,
3) $\|P_{\bar{X},\bar{U}}-P\|_{\mathcal{F}}\le\widehat{\mathbb{D}}_n(\mathcal{P},\mathcal{F},R)$,
4) $I(\bar{Y};\bar{U})\le R$.

Hence, for every $P\in\mathcal{P}$, $\widehat{\mathbb{D}}_n(\mathcal{P},\mathcal{F},R)\ge D_{\mathrm{KS}}(P,\mathcal{F},R)$. Taking the supremum over all $P\in\mathcal{P}$ and then the limit as $n\to\infty$, we get the desired result. ∎

However, it is not straightforward to derive an information-theoretic upper bound on $\widehat{\mathbb{D}}(\mathcal{P},\mathcal{F},R)$ when $|\mathcal{P}|>1$. This would require constructing a rate-R code that asymptotically achieves $\widehat{\mathbb{D}}(\mathcal{P},\mathcal{F},R)$. In order to prove achievability, one could take a rate-R code for each "representative" distribution in $\mathcal{P}$ (assuming $\mathcal{P}$ is not too rich, so that it can be represented by a slowly, e.g., subexponentially, growing number of distributions), combine the codes into a union code (which will result in an asymptotically negligible rate overhead), and then devise a rule for mapping the sequence Y^n into one of the codewords. However, the difficulty here is that the encoder can only estimate the Y-marginal of the underlying distribution and cannot select the right code based on this information alone. One (suboptimal) strategy is to bound the distortion $\|\mathsf{P}_{(X^n,\widehat{Y}^n)}-P\|_{\mathcal{F}}$ by the average of single-letter functions of the form

$$\rho_{\mathcal{F},P}(X_i,\widehat{Y}_i)\triangleq\|\delta_{(X_i,\widehat{Y}_i)}-P\|_{\mathcal{F}},$$

where $\delta_{(X_i,\widehat{Y}_i)}$ is the Dirac measure concentrated on $(X_i,\widehat{Y}_i)$, and consider the new problem of finding

$$\inf_{\widehat{Y}^n}\sup_{P\in\mathcal{P}}\mathbb{E}_P\left[\frac{1}{n}\sum_{i=1}^{n}\rho_{\mathcal{F},P}(X_i,\widehat{Y}_i)\right] \tag{3.9}$$

where the infimum is over all rate-R codes $\widehat{Y}^n:\mathcal{Y}^n\to\mathcal{Y}^n$. Then (3.9) will be an upper bound on $\widehat{\mathbb{D}}_n(\mathcal{P},\mathcal{F},R)$. Note that the problem of minimizing (3.9) is an instance of minimax noisy source coding [9]: given a sequence of i.i.d. samples $(X_1,Y_1),(X_2,Y_2),\ldots$ from an unknown $P\in\mathcal{P}$ and a blocklength n, we wish to code Y^n using a rate-R code, such that the sequence X^n is reconstructed from the encoded data with small average $\rho_{P,\mathcal{F}}(\cdot,\cdot)$ distortion. When $\mathcal{Y}$ is finite, a type-covering argument, as in [9], can be used to show (3.10) at the top of this page. Given any $\alpha>0$, $\delta>0$, and $P'\in\mathcal{M}(\mathcal{Y})$, the second infimum in (3.10) is over all conditional probability distributions (transition kernels) from $\mathcal{Y}$ to $\mathcal{Y}$, such that the mutual information between Y and U when $Y\sim P'$ and $U|Y\sim Q_{U|Y}$, is at most $R+\alpha$. The inner supremum is over all probability distributions $P\in\mathcal{P}$, such that their Y-marginal P_Y is within δ from P' in the variational norm $\|\cdot\|_V$, $P\times Q_{U|Y}$ denotes the joint distribution of X, Y and U when $(X,Y)\sim P$ and $U|Y\sim Q_{U|Y}$, and $\delta_{(X,U)}$ denotes the Dirac measure concentrated at $(X,U)\in\mathcal{X}\times\mathcal{Y}$. We leave the problem of tightening (3.10) for future work. Evidently, the difficulties involved in extending this technique to general $\mathcal{Y}$ are of the same nature as in [9] and have to do with finding the right topology on $\mathcal{M}(\mathcal{Y})$ that would give the same uniform error bounds as for the variational distance in the finite case.

IV. Discussion

We have derived information-theoretic bounds on the performance achievable by statistical learning algorithms in the presence of two types of communication constraints:

- Type I — the entire training sequence is delivered to the learning agent over a noiseless digital channel with finite capacity.
- Type II — the input part of the training sequence is available to the learning agent exactly, while the output part is delivered over a rate-limited channel.

We have shown that, under certain regularity conditions on the underlying distribution and the family of predictors available to the learning algorithm, the Type I setting is not very different from the traditional setting of learning theory, where there are no communication constraints, and the learning agent sees the training data with arbitrary precision. In fact, the learning algorithm converges to the optimal predictor at the same rate as in the infinite-precision setting. There is, however, a price to be paid for this. In the infinite-precision setting one can have learning algorithms, such as the ERM algorithm defined in (3.3), that have *distribution-free* performance guarantees. That is, the optimal predictor is learned for *any* probability distribution of X and Y, and the convergence rate depends only on the sample size n and on the geometry of the class of predictors $\mathcal{F}$. However, once we impose the communication constraints, we can no longer guarantee distribution-free performance. Instead, the class $\mathcal{P}$ has to be sufficiently well-behaved, so that we can cover it by finitely many balls in the metric induced by the class $\mathcal{F}$.

In practice, this loss of distribution-free guarantees will not cause many problems. For example, considering the regression setting, suppose that X and Y are related via $Y = f^*(X)+W$, where f^* is some unknown function, and the additive noise W is independent of X. Then, assuming that we have prior knowledge that f^* lies in some sufficiently regular function class $\mathcal{F}^*$ (such as a Hölder or a Besov ball), we can deduce the covering properties of the resulting class $\mathcal{P}$ from the covering properties of $\mathcal{F}^*$. Typical nonparametric classes have ϵ-covering numbers that behave like $(1/\epsilon)^\alpha$, where $\alpha > 0$ is the smoothness index. Provided that the functions in $\mathcal{F}^*$ are uniformly bounded, we can guarantee similar covering properties for $\mathcal{P}$.

However, in a Type II setting, things are radically different. The encoder has access only to the output part of the training sequence, and therefore does not have enough information to be able to estimate the underlying distribution. Instead, we must guarantee that the rate-R description provided by the encoder is sufficient for the learner to estimate the sample averages of the prediction losses for every predictor in $\mathcal{F}$ to within a desired level of distortion. To achieve this objective, we have extended recent work of Kramer and Savari [2] on rate-constrained communication of probability distributions in two directions: (1) we have allowed for general (not necessarily finite) alphabets, and (2) we have allowed for uncertainty in the underlying source distribution. We have formulated an appropriate operational version of this problem and derived upper and lower single-letter information-theoretic bounds on the achievable performance. Currently, our upper bound is specialized to the classification setting, and there is a gap between the upper bound and the lower bound. It still remains to develop appropriate extensions to the regression setting and to investigate the possibility of closing the gap.

Our approach to rate-constrained learning in a Type II setting is reminiscent of *noisy source coding* (see [9] and references therein). In noisy source coding, the objective is to construct a low-distortion reproduction of a given source when the source sequence can only be viewed through a noisy channel. In the setting of this paper, the input part of the training sequence is the clean source of interest, and the output part (available to the encoder) is the noisy source. The noisy source coding perspective assumes that the original (clean) source sequence is unavailable. This is, of course, not the case in the problem of interest because the input part of the training sequence is available to the learner. Thus, strictly speaking, it is more appropriate to view the Type II scenario from the Wyner–Ziv perspective of rate-distortion coding with side information at the decoder [10], [11]. The encoder should be able to exploit the fact that the learner has access to *side information* in the form of the input part of the training sequence in order to further reduce the distortion. However, as pointed out recently by Merhav and Ziv [12], there is no hope for obtaining a universal Wyner–Ziv coding scheme when the encoder does not know the distribution of the side information sequence. Even in the case when such knowledge is available, universality can be guaranteed only under fairly restrictive assumptions [13]. These considerations have led us to focus on the noisy source coding formulation.

In the future, we plan to obtain closed-form bounds on the distortion-rate functions introduced in the Type II setting in the specific instances of distribution families and predictor classes used in statistical learning theory.

Acknowledgment

The author gratefully acknowledges stimulating discussions with Todd Coleman, Sayan Mukherjee, David Neuhoff, Andrew Nobel, Sandeep Pradhan, Clayton Scott, Sergio Verdú, and Rebecca Willett.

Appendix

In this appendix we prove the following lemma, which is an extension of the Piggyback Coding lemma of Wyner [14, Lemma 4.3] to general alphabets:

Lemma A.1. Let $\mathcal{X}, \mathcal{Y}, \mathcal{U}$ be standard measurable spaces, so that regular versions of conditional distributions exist [8], and let $(X, Y, U) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{U}$ be a triple of random variables forming a Markov chain: $X \to Y \to U$. Let $\{(X_i, Y_i, U_i)\}_{i=1}^\infty$ be a sequence of independent drawings from the joint distribution P_{XYU}. Let $\{\psi_n\}_{n=1}^\infty$ be a sequence of measurable functions $\psi_n : \mathcal{X}^n \times \mathcal{U}^n \to [0, 1]$, such that

$$\lim_{n\to\infty} \mathbb{E}\, \psi_n(X^n, U^n) = 0.$$

For a given $\epsilon > 0$, there exists $n_0 = n_0(\epsilon)$, such that for every $n \geq n_0$ we can find a mapping $F_n : \mathcal{Y}^n \to \mathcal{U}^n$ that satisfies

$$\frac{1}{n} \log \left| \left\{ F_n(y^n) : y^n \in \mathcal{Y}^n \right\} \right| \leq I(Y; U) + \epsilon$$

and

$$\mathbb{E}\,\psi_n(X^n, F_n(Y^n)) \le \epsilon,$$

where $I(Y;U)$ is the mutual information between Y and U.

Proof: The proof is very similar to Wyner's proof for finite alphabets [14]. Fix $n \ge 1$ and define a function $\phi_n : \mathcal{Y}^n \times \mathcal{U}^n \to [0,1]$ by

$$\begin{aligned}\phi_n(y^n,u^n) &\triangleq \mathbb{E}\left[\psi_n(X^n,U^n)\middle|Y^n=y^n,U^n=u^n\right]\\ &= \int_{\mathcal{X}^n}\psi_n(x^n,u^n)dP_{X^n|Y^n,U^n}(x^n|y^n,u^n).\end{aligned}$$

Owing to the Markov chain condition, we can write

$$\phi_n(y^n,u^n) = \int_{\mathcal{X}^n}\psi_n(x^n,u^n)dP_{X^n|Y^n}(x^n|y^n).$$

Letting $\delta_n \triangleq \mathbb{E}\,\psi_n(X^n,U^n)$, we define the set

$$\mathcal{S}_n \triangleq \left\{(y^n,u^n)\in\mathcal{Y}^n\times\mathcal{U}^n : \phi_n(y^n,u^n)\le\sqrt{\delta_n}\right\}.$$

Then by the Markov inequality we have

$$\mathbb{P}\left((Y^n,U^n)\notin\mathcal{S}_n\right) \le \frac{\mathbb{E}\,\phi_n(Y^n,U^n)}{\sqrt{\delta_n}} = \sqrt{\delta_n}.$$

Consider an arbitrary measurable mapping $G : \mathcal{Y}^n \to \{u_1^n,\ldots,u_M^n\}\subset\mathcal{U}^n$ for some $M<\infty$. Then, since $\psi_n(\cdot,\cdot)\le 1$, we can write

$$\begin{aligned}&\mathbb{E}\,\psi_n(X^n,G(Y^n))\\ &\quad= \mathbb{E}\,\phi_n(Y^n,G(Y^n))\\ &\quad\le \mathbb{P}\left((Y^n,G(Y^n))\notin\mathcal{S}_n\right)\\ &\qquad\qquad+\int_{\{y^n:(y^n,G(y^n))\in\mathcal{S}_n\}}\phi_n(y^n,G(y^n))dP_{Y^n}(y^n).\end{aligned}$$

We further have

$$\begin{aligned}&\int_{\{y^n:(y^n,G(y^n))\in\mathcal{S}_n\}}\phi_n(y^n,G(y^n))dP_{Y^n}(y^n)\\ &\quad=\sum_{m=1}^{M}\int_{\{y^n:G(y^n)=u_m^n,(y^n,u_m^n)\in\mathcal{S}_n\}}\phi_n(y^n,u_m^n)dP_{Y^n}(y^n)\\ &\quad\le\sqrt{\delta_n}.\end{aligned}$$

Hence,

$$\mathbb{E}\,\psi_n(X^n,G(Y^n)) \le \mathbb{P}\left((Y^n,G(Y^n))\notin\mathcal{S}_n\right)+\sqrt{\delta_n}.$$

Now we can use Lemma 9.3.1 in [15] to show that, given $\mathcal{S}_n$, M, and an arbitrary $R>0$, there exists a set $\{u_1^n,\ldots,u_M^n\}\subset\mathcal{U}^n$ and a mapping $G_n:\mathcal{Y}^n\to\{u_1^n,\ldots,u_M^n\}$, such that

$$\begin{aligned}&\mathbb{P}\left((Y^n,G_n(Y^n))\notin\mathcal{S}_n\right)\\ &\quad\le \mathbb{P}\left((Y^n,U^n)\notin\mathcal{S}_n\right)+\mathbb{P}\left(i(Y^n,U^n)>nR\right)\\ &\qquad\qquad+\exp\left(-M2^{-Rn}\right),\end{aligned}$$

where

$$i(y^n,u^n)\triangleq\log\frac{dP_{Y^n,U^n}}{dP_{Y^n}\times dP_{U^n}}(y^n,u^n)$$

is the information density. Letting $M=2^{n(I(Y;U)+\epsilon)}$ and $R=I(Y;U)+\epsilon/2$ and using the corresponding mapping G_n, we get

$$\begin{aligned}&\mathbb{E}\,\psi_n(X^n,G_n(Y^n))\\ &\quad\le 2\sqrt{\delta_n}+\exp(-2^{n\epsilon/2})+\mathbb{P}\left(i(Y^n,U^n)>nR\right).\end{aligned}$$

Since $\mathbb{E}\,\psi_n(X^n,U^n)=\delta_n\to 0$ as $n\to\infty$, the first term goes to zero as $n\to\infty$. The second term likewise goes to 0 since $\epsilon>0$. The third term goes to zero owing to the law of large numbers. Choosing n_0 large enough so that the right-hand side of the above inequality is less than ϵ finishes the proof. ■

REFERENCES

[1] M. Raginsky, "Learning from compressed observations," in *Proc. IEEE Inform. Theory Workshop*, Lake Tahoe, CA, September 2007, pp. 420–425.

[2] G. Kramer and S. A. Savari, "Communicating probability distributions," *IEEE Trans. Inform. Theory*, vol. 53, no. 2, pp. 518–525, February 2007.

[3] M. B. Westover and J. A. O'Sullivan, "Achievable rates for pattern recognition," *IEEE Trans. Inform. Theory*, vol. 54, no. 1, pp. 299–320, January 2008.

[4] D. Haussler, "Decision-theoretic generalizations of the PAC model for neural net and other learning applications," *Inform. Comput.*, vol. 100, pp. 78–150, 1992.

[5] M. Vidyasagar, *Learning and Generalization*, 2nd ed. London: Springer-Verlag, 2003.

[6] A. W. van der Waart and J. A. Wellner, *Weak Convergence and Empirical Processes*. New York: Springer-Verlag, 1996.

[7] A. N. Kolmogorov and V. M. Tihomirov, "ϵ-entropy and ϵ-capacity of sets in function spaces," in *Amer. Math. Soc. Transl.*, ser. 2, 1961, vol. 17, pp. 277–364.

[8] R. M. Gray, *Probability, Random Processes, and Ergodic Properties*. New York: Springer-Verlag, 1981.

[9] 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.

[10] 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.

[11] 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.

[12] N. Merhav and J. Ziv, "On the Wyner–Ziv problem for individual sequences," *IEEE Trans. Inform. Theory*, vol. 52, no. 3, pp. 867–873, March 2006.

[13] M. Raginsky, "Universal Wyner–Ziv coding for discrete memoryless sources with known side information statistics," in *Proc. IEEE Int. Symp. on Information Theory*, Toronto, Canada, July 2008, pp. 2167–2171.

[14] 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.

[15] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.