

Achievability Results for Statistical 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.

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

