# Minimax Learning for Remote Prediction

Cheuk Ting Li
University of California, Berkeley
Email: ctli@berkeley.edu

Xiugang Wu, Ayfer Ozgur, Abbas El Gamal
Stanford University
Email: {x23wu, aozgur}@stanford.edu; abbas@ee.stanford.edu

***Abstract*—The classical problem of supervised learning is to infer an accurate predictor of a target variable $Y$ from a measured variable $X$ by using a finite number of labeled training samples. Motivated by the increasingly distributed nature of data and decision making, in this paper we consider a variation of this classical problem in which the prediction is performed remotely based on a rate-constrained description $M$ of $X$. Upon receiving $M$, the remote node computes an estimate $\hat{Y}$ of $Y$. We follow the recent minimax approach to study this learning problem and show that it corresponds to a one-shot minimax noisy source coding problem. We then establish information theoretic bounds on the risk-rate Lagrangian cost and a general method to design a near-optimal descriptor-estimator pair, which can be viewed as a rate-constrained analog to the maximum conditional entropy principle used in the classical minimax learning problem. Our results show that a naive estimate-compress scheme for rate-constrained prediction is not in general optimal.**

## I. Introduction

The classical problem of supervised learning is to infer an accurate predictor of a target variable $Y$ from a measured variable $X$ on the basis of $n$ labeled training samples $\{(X_i, Y_i)\}_{i=1}^n$ independently drawn from an unknown joint distribution $P$. The standard approach for solving this problem in statistical learning theory is empirical risk minimization (ERM). For a given set of allowable predictors and a loss function that quantifies the risk of each predictor, ERM chooses the predictor with minimal risk under the empirical distribution of samples. To avoid overfitting, the set of allowable predictors is restricted to a class with limited complexity.

Recently, an alternative viewpoint has emerged which seeks distributionally robust predictors. Given the labeled training samples, this approach learns a predictor by minimizing its worst-case risk over an ambiguity distribution set centered at the empirical distribution of samples. In other words, instead of restricting the set of allowable predictors, it aims to avoid overfitting by requiring that the learned predictor performs well under any distribution in a chosen neighborhood of the empirical distribution. This minimax approach has been investigated under different assumptions on how the ambiguity set is constructed, e.g., by restricting the moments [1], forming the $f$-divergence balls [2] and Wasserstein balls [3] (see also references therein).

In these previous works, the learning algorithm finds a predictor that acts directly on a fresh (unlabeled) sample $X$ to predict the corresponding target variable $Y$. Often, however the fresh sample $X$ may be only remotely available, and when designing the predictor it is desirable to also take into account the cost of communicating $X$. This is motivated by the fact that bandwidth and energy limitations on communication in networks and within multiprocessor systems often impose significant bottlenecks on the performance of algorithms. There are also an increasing number of applications in which data is generated in a distributed manner and it (or features of it) are communicated over bandwidth-limited links to a central processor to perform inference. For instance, applications such as Google Goggles and Siri process the locally collected data on clouds. It is thus important to study prediction in distributed and rate-constrained settings.

In this paper, we study an extension of the classical learning problem in which given a finite set of training samples, the learning algorithm needs to infer a descriptor-estimator pair with a desired communication rate in between them. This is especially relevant when both $X$ and $Y$ come from a large alphabet or are continuous random variables as in regression problems, so neither the sample $X$ nor its predicted value of $Y$ can be simply communicated in a lossless fashion. We adopt the minimax framework for learning the descriptor-estimator pair. Given a set of labeled training samples, our goal is to find a descriptor-estimator pair by minimizing their resultant worst-case risk over an ambiguity distribution set, where the risk now incorporates both the statistical risk and the communication cost. One of the important conclusions that emerge from the minimax approach to supervised learning in [1] is that the problem of finding the predictor with minimal worst-case risk over an ambiguity set can be broken into two smaller steps: (1) find the worst-case distribution in the ambiguity set that maximizes the (generalized) conditional entropy of $Y$ given $X$, and (2) find the optimal predictor under this worst-case distribution. In this paper, we show that an analogous principle approximately holds for rate-constrained prediction. The descriptor-estimator pair with minimal worst-case risk can be found in two steps: (1) find the worst-case distribution in the ambiguity set that maximizes the risk-information Lagrangian cost, and (2) find the optimal descriptor-estimator pair under this worst-case distribution. We then apply our results to characterize the optimal descriptor-estimator pairs for two applications: rate-constrained linear regression and rate-constrained classification. While a simple scheme whereby we first find the optimal predictor ignoring the rate constraint, then compress and communicate the predictor output, is optimal for the linear regression application, we show via the classification application that such an estimate-compress approach is not optimal in general. We show that when prediction is rate-constrained, the optimal descriptor

978-1-5386-4780-6/18/$31.00©2018 IEEE

aims to send sufficiently (but not necessarily maximally) informative features of the observed variable, which are at the same time easy to communicate. When applied to the case in which the ambiguity distribution set contains only a single distribution (for example, the true or empirical distribution of $X, Y$) and the loss function for the prediction is logarithmic loss, our results provide a new one-shot operational interpretation of the information bottleneck problem. A key technical ingredient in our results is the strong functional representation lemma (SFRL) developed in [4], which we use to design the optimal descriptor-estimator pair for the worst-case distribution.

*Notation*

We assume that log is base 2 and the entropy $H$ is in bits. The length of a variable-length description $M \in \{0,1\}^*$ is denoted as $|M|$. For random variables $U, V$, denote the joint distribution by $P_{U,V}$ and the conditional distribution of $U$ given $V$ by $P_{U|V}$. For brevity we denote the distribution of $(X, Y)$ as $P$. We write $I_P(X;\hat{Y})$ for $I(X;\hat{Y})$ when $(X, Y) \sim P$, and $P_{\hat{Y}|X}$ is clear from the context.

## II. Problem Formulation

We begin by reviewing the minimax approach to the classical learning problem [1].

### A. Minimax Approach to Supervised Learning

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 a target variable $Y$ from a measured variable $X$ on the basis of a number of independent training samples $\{(X_i, Y_i)\}_{i=1}^n$ drawn from an unknown joint distribution. The standard approach for solving this problem is to use empirical risk minimization (ERM) in which one defines an admissible class of predictors $\mathcal{F}$ that consists of functions $f : \mathcal{X} \to \hat{\mathcal{Y}}$ (where the reconstruction alphabet $\hat{\mathcal{Y}}$ can be in general different from $\mathcal{Y}$) and a loss function $\ell : \hat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$. The risk associated with a predictor $f$ when the underlying joint distribution of $X$ and $Y$ is $P$ is

$$L(f, P) \triangleq \mathsf{E}_P[\ell(f(X), Y)].$$

ERM simply chooses the predictor $f_n \in \mathcal{F}$ with minimal risk under the empirical distribution $P_n$ of the training samples.

Recently, an alternative approach has emerged which seeks distributionally robust predictors. This approach learns a predictor by minimizing its worst-case risk over an ambiguity distribution set $\Gamma(P_n)$, i.e.,

$$f_n = \operatorname*{argmin}_f \max_{P \in \Gamma(P_n)} L(f, P), \tag{1}$$

where $f$ can be any function and $\Gamma(P_n)$ can be constructed in various ways, e.g., by restricting the moments, forming the $f$-divergence balls or Wasserstein balls. While in ERM it is important to restrict the set $\mathcal{F}$ of admissible predictors to a low-complexity class to prevent overfitting, in the minimax approach overfitting is prevented by explicitly requiring that the chosen predictor is distributionally robust. The learned function $f_n$ can be then used for predicting $Y$ when presented with fresh samples of $X$. The learning and inference phases are illustrated in Figure 1.

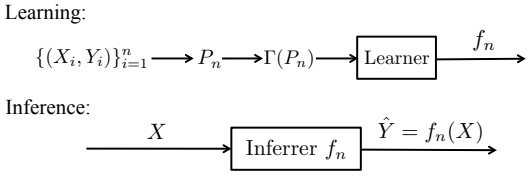


Fig. 1. Minimax approach to supervised learning.

### B. Minimax Learning for Remote Prediction

In this paper, we extend the minimax learning approach to the setting in which the prediction needs to be performed based on a rate-constrained description of $X$. In particular, given a set of finite training samples $\{(X_i, Y_i)\}_{i=1}^n$ independently drawn from an unknown joint distribution $P$, our goal is to learn a pair of functions $(e, f)$, where $e$ is a descriptor used to compress $X$ into $M = e(X) \in \{0,1\}^*$ (a prefix-free code), and $f$ is an estimator that takes the compression $M$ and generates an estimate $\hat{Y}$ of $Y$. See Figure 2.

Let $R(e, P) \triangleq \mathsf{E}_P[|e(X)|]$ be the rate of the descriptor $e$ and $L(e, f, P) \triangleq \mathsf{E}_P[\ell(f(e(X)), Y)]$ be the risk associated with the descriptor-estimator pair $(e, f)$, when the underlying distribution of $(X, Y)$ is $P$, and define the risk-rate Lagrangian cost (parametrized by $\lambda > 0$) as

$$L_\lambda(e, f, P) = L(e, f, P) + \lambda R(e, P). \tag{2}$$

Note that this cost function takes into account both the resultant statistical prediction risk of $(e, f)$, as well as the communication rate they require. The task of a minimax learner is to find an $(e_n, f_n)$ pair that minimizes the worst-case $L_\lambda(e, f, P)$ over the ambiguity distribution set $\Gamma(P_n)$, i.e.,

$$(e_n, f_n) = \operatorname*{argmin}_{(e,f)} \max_{P \in \Gamma(P_n)} L_\lambda(e, f, P), \tag{3}$$

for an appropriately chosen $\Gamma(P_n)$ centered at the empirical distribution of samples $P_n$. Note that we allow here all possible $(e, f)$ pairs. We also assume that the descriptor and the estimator can use unlimited common randomness $W$ which is independent of the data, i.e., $e$ and $f$ can be expressed as functions of $(X, W)$ and $(M, W)$, respectively. The availability of such common randomness can be justified by the fact that in practice, although the inference scheme is one-shot, it is used many times (by the same user and by different users), hence the descriptor and the estimator can share a common randomness seed before communication commences without impacting the communication rate.

## III. Main Results

We first consider the case where $\Gamma$ consists of a single distribution $P$, which may be the empirical distribution $P_n$ as in ERM. Define the minimax risk-rate cost as

$$L_\lambda^*(\Gamma) = \inf_{(e,f)} \sup_{P \in \Gamma} L_\lambda(e, f, P). \tag{4}$$

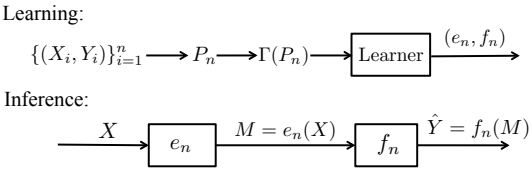


Fig. 2. Minimax learning for remote prediction.

While it is difficult to minimize the risk-rate cost (2) directly, the minimax risk-rate cost can be bounded in terms of the mutual information between $X$ and $\hat{Y}$.

**Theorem 1.** *Let $\Gamma = \{P\}$. Then*

$$L_\lambda^* \geq \inf_{P_{\hat{Y}|X}} \left( \mathsf{E}\left[\ell(\hat{Y}, Y)\right] + \lambda I(X;\hat{Y}) \right),$$

$$L_\lambda^* \leq \inf_{P_{\hat{Y}|X}} \Big( \mathsf{E}\Big[\ell(\hat{Y}, Y)\Big] + \lambda \Big( I(X;\hat{Y}) + \log(I(X;\hat{Y})+1) + 5 \Big) \Big).$$

As in other one-shot compression results (e.g., zero-error compression), there is a gap between the upper and lower bound. While the logarithmic gap in Theorem 1 is not as small as the 1-bit gap in the zero-error compression, it is dominated by the linear term $I(X;\hat{Y})$ when it is large.

To prove Theorem 1, we use the strong functional representation lemma given in [4] (also see [5], [6]): for any random variables $X, \hat{Y}$, there exists random variable $W$ independent of $X$, such that $\hat{Y}$ is a function of $(X, W)$, and

$$H(\hat{Y}|W) \leq I(X;\hat{Y}) + \log(I(X;\hat{Y}) + 1) + 4. \tag{5}$$

Here, $W$ can be intuitively viewed as the part of $\hat{Y}$ which is not contained in $X$. Note that for any $W$ such that $\hat{Y}$ is a function of $(X, W)$ and $W$ is independent of $X$, $H(\hat{Y}|W) \geq I(X;\hat{Y})$. The statement (5) ensures the existence of an $W$, independent of $X$, which comes close to this lower bound, and in this sense it is most informative about $\hat{Y}$. This is critical for the proof of Theorem 1 as we will see next. Identifying the part of $\hat{Y}$ which is not contained in $X$ allows us to generate and share this part between the descriptor and the estimator ahead of time, eliminating the need to communicate it during the course of inference. To find $W$, we use the Poisson functional representation construction detailed in [4].

*Proof of Theorem 1:* Recall that $\hat{Y} = f(e(X, W), W)$. The lower bound follows from the fact that $I_P(X;\hat{Y}) \leq H_P(M) \leq \mathsf{E}[|M|]$. To establish the upper bound, fix any $P_{\hat{Y}|X}$. Let $W$ be obtained from (5). Note that $W$ is independent of $X$ and can be generated from a random seed shared between the descriptor and the estimator ahead of time. For a given $w$, take $m = e(x, w)$ to be the Huffman codeword of $\hat{y}(x, w)$ according to the distribution $P_{\hat{Y}|W}(\cdot|w)$ (recall that $\hat{Y}$ is a function of $(X, W)$), and take $f(m, w)$ to be the decoding function of the Huffman code. The expected codeword length

$$\mathsf{E}[|M|] \leq H(\hat{Y}|W) + 1 \leq I(X;\hat{Y}) + \log(I(X;\hat{Y}) + 1) + 5.$$

Taking an infimum over all $P_{\hat{Y}|X}$ completes the proof. ∎

**Remark 1.** If we consider the logarithmic loss $\ell(\hat{y}, y) = -\log \hat{y}(y)$, where $\hat{y}$ is a distribution over $\mathcal{Y}$, then the lower bound in Theorem 1 reduces to

$$\inf_{P_{U|X}} (H(Y|U) + \lambda I(X;U)) = H(Y) + \inf_{P_{U|X}} (\lambda I(X;U) - I(Y;U)),$$

which is the information bottleneck function [7]. Therefore the setting of remote prediction provides an approximate one-shot operational interpretation of the information bottleneck (up to a logarithmic gap). In [8], [9] it was shown that the asymptotic noisy source coding problem also provides an operational interpretation of the information bottleneck. Our operational interpretation, however, is more satisfying since the feature extraction problem originally considered in [7] is by nature one-shot.

We now extend Theorem 1 to the minimax setting.

**Theorem 2.** *Suppose $\Gamma$ is convex. Then*

$$L_\lambda^* \geq \inf_{P_{\hat{Y}|X}} \sup_{P \in \Gamma} \left( \mathsf{E}_P\left[\ell(\hat{Y}, Y)\right] + \lambda I_P(X;\hat{Y}) \right)$$

$$L_\lambda^* \leq \inf_{P_{\hat{Y}|X}} \sup_{P \in \Gamma} \Big( \mathsf{E}_P\left[\ell(\hat{Y}, Y)\right] + \lambda \left( I_P(X;\hat{Y}) + 2\log(I_P(X;\hat{Y}) + 1) + 6 \right) \Big).$$

This result is related to minimax noisy source coding [10]. The main difference is that we consider the one-shot expected length instead of the asymptotic rate.

To prove this theorem, we first invoke a minimax result for relative entropy in [11] (which generalizes the redundancy-capacity theorem [12]). Then we apply the following refined version of the strong functional representation lemma that is proved in the proof of Theorem 1 in [4] (also see [5]).

**Lemma 1.** *For any $P_{\hat{Y}|X}$ and $\tilde{P}_{\hat{Y}}$, there exists random variable $W$, and functions $k(x, w) \in \{1, 2, \ldots\}$ and $\hat{y}(k, w)$ such that $\hat{y}(k(x, W), W) \sim P_{\hat{Y}|X}(\cdot|x)$, and*

$$\mathsf{E}\left[\log k(x, W)\right] \leq D\big(P_{\hat{Y}|X}(\cdot|x) \,\big\|\, \tilde{P}_{\hat{Y}}\big) + 1.6. \tag{6}$$

We are now ready to prove Theorem 2.

*Proof:* The lower bound follows from $\mathsf{E}_P[|M|] \geq H_P(M) \geq I_P(X;\hat{Y})$. To prove the upper bound, we fix any $P_{\hat{Y}|X}$, and show that the following risk-rate cost is achievable:

$$L' = \sup_{P \in \Gamma} \Big( \mathsf{E}_P\left[\ell(\hat{Y}, Y)\right] + \lambda \left( I_P(X;\hat{Y}) + 2\log(I_P(X;\hat{Y}) + 1) + 6 \right) \Big).$$

Let

$$g(P, \tilde{P}_{\hat{Y}}) = \mathsf{E}_P\left[\ell(\hat{Y}, Y)\right] + \lambda \Bigg( \int D\big(P_{\hat{Y}|X=x} \,\big\|\, \tilde{P}_{\hat{Y}}\big) dP(x) + 2\log\Big( \int D\big(P_{\hat{Y}|X=x} \,\big\|\, \tilde{P}_{\hat{Y}}\big) dP(x) + 1 \Big) + 6 \Bigg).$$

Note that $g$ is concave in $P$ for fixed $\tilde{P}_{\hat{Y}}$ since $\mathsf{E}_P\left[\ell(\hat{Y}, Y)\right]$ and $\int D\big(P_{\hat{Y}|X=x} \,\big\|\, \tilde{P}_{\hat{Y}}\big) dP(x)$ are linear in $P$. Also $g$ is quasiconvex in $\tilde{P}_{\hat{Y}}$ for fixed $P$ since $\int D\big(P_{\hat{Y}|X=x} \,\big\|\, \tilde{P}_{\hat{Y}}\big) dP(x)$ is convex in $\tilde{P}_{\hat{Y}}$, and is lower semicontinuous in $\tilde{P}_{\hat{Y}}$

since $D\big(P_{\hat{Y}|X=x} \,\big\|\, \tilde{P}_{\hat{Y}}\big)$ is lower semicontinuous with respect to the topology of weak convergence [13], and hence $\int D\big(P_{\hat{Y}|X=x} \,\big\|\, \tilde{P}_{\hat{Y}}\big)dP(x)$ is lower semicontinuous by Fatou's lemma.

Write $P_{\hat{Y}|X} \circ P$ for the distribution of $\hat{Y}$ when $(X,Y) \sim P$ and $\hat{Y}|\{X=x\} \sim P_{\hat{Y}|X}(\cdot|x)$. Let $\Gamma_{\hat{Y}} = \{P_{\hat{Y}|X} \circ P : P \in \Gamma\}$ and $\overline{\Gamma_{\hat{Y}}}$ be the closure of $\Gamma_{\hat{Y}}$ in the topology of weak convergence. It can be shown using the same arguments as in [11] (on $g$ instead of relative entropy, and using Sion's minimax theorem [14] instead of Lemma 2 in [11]) that if $\Gamma_{\hat{Y}}$ is uniformly tight, then there exists $P^*_{\hat{Y}} \in \overline{\Gamma_{\hat{Y}}}$ such that

$$\sup_{P\in\Gamma} g(P, \tilde{P}^*_{\hat{Y}}) = \sup_{P\in\Gamma} \inf_{\tilde{P}_{\hat{Y}}} g(P, \tilde{P}_{\hat{Y}}) = L'.$$

If $\Gamma_{\hat{Y}}$ is not uniformly tight, then by Lemma 4 in [11], $\sup_{P\in\Gamma} \inf_{\tilde{P}_{\hat{Y}}} \int D\big(P_{\hat{Y}|X=x} \,\big\|\, \tilde{P}_{\hat{Y}}\big)dP(x) = \infty$, and hence $L' = \sup_{P\in\Gamma} \inf_{\tilde{P}_{\hat{Y}}} g(P, \tilde{P}_{\hat{Y}}) = \infty$.

Applying Lemma 1 to $P_{\hat{Y}|X}$, $P^*_{\hat{Y}}$ we obtain $W$ independent of $X$, random variable $K = k(X,W) \in \{1,2,\ldots\}$, and $\hat{Y} = \hat{y}(K,W)$ following the conditional distribution $P_{\hat{Y}|X}$, and

$$\mathsf{E}\left[\log K \,|\, X = x\right] \le D\big(P_{\hat{Y}|X} \,\big\|\, P^*_{\hat{Y}} \,\big|\, X = x\big) + 1.6$$

for any $x$. Then we use Elias delta code [15] for $K$ to produce $M$. Note that the average length of the Elias delta code is upper bounded by $\log K + 2\log(\log K + 1) + 1$. Hence, we have

$$\begin{aligned}\mathsf{E}_P\left[|M|\right] &\le \mathsf{E}_P\left[\log K\right] + 2\log\left(\mathsf{E}_P\left[\log K\right] + 1\right) + 1 \\ &\le \int D\big(P_{\hat{Y}|X=x} \,\big\|\, P^*_{\hat{Y}}\big)dP(x) \\ &\quad + 2\log\left(\int D\big(P_{\hat{Y}|X=x} \,\big\|\, P^*_{\hat{Y}}\big)dP(x) + 1\right) + 6.\end{aligned}$$

Hence

$$\tilde{L}^*_\lambda \le \sup_{P\in\Gamma}\left(\mathsf{E}_P\left[\ell(\hat{Y},Y) + \lambda|M|\right]\right) \le \sup_{P\in\Gamma} g(P, P^*_{\hat{Y}}) \le L'.$$

■

Theorem 2 suggest that we can simplify the analysis of the risk-rate cost (2) $L_\lambda = \mathsf{E}_P\left[\ell(\hat{Y},Y)\right] + \lambda\,\mathsf{E}_P\left[|M|\right]$ by replacing the rate $\mathsf{E}_P\left[|M|\right]$ with the mutual information $I_P(X;\hat{Y})$. Define the *risk-information cost* as

$$\tilde{L}_\lambda(P_{\hat{Y}|X}, P) = \mathsf{E}_P\left[\ell(\hat{Y},Y)\right] + \lambda I_P(X;\hat{Y}). \tag{7}$$

Theorem 2 implies that the minimax risk-rate cost $L^*_\lambda$ can be approximated by the *minimax risk-information cost*

$$\tilde{L}^*_\lambda(\Gamma) = \inf_{P_{\hat{Y}|X}} \sup_{P\in\Gamma} \tilde{L}_\lambda(P_{\hat{Y}|X}, P), \tag{8}$$

within a logarithmic gap. Theorem 2 can also be stated in the following slightly weaker form

$$\tilde{L}^*_\lambda \le L^*_\lambda \le \tilde{L}^*_\lambda + 2\lambda\log(\lambda^{-1}\tilde{L}^*_\lambda + 1) + 7\lambda.$$

The risk-information cost has more desirable properties than the risk-rate cost. For example, it is convex in $P_{\hat{Y}|X}$ for fixed $P$, and concave in $P$ for fixed $P_{\hat{Y}|X}$. This allows us to exchange the infimum and supremum in Theorem 2 by Sion's minimax theorem [14], which gives the following proposition.

**Proposition 1.** *Suppose $\mathcal{X}$, $\mathcal{Y}$ and $\hat{\mathcal{Y}}$ are finite, $\Gamma$ is convex and closed, and $\lambda \ge 0$, then*

$$\tilde{L}^*_\lambda(\Gamma) = \inf_{P_{\hat{Y}|X}} \sup_{P\in\Gamma} \tilde{L}_\lambda(P_{\hat{Y}|X}, P) = \sup_{P\in\Gamma} \inf_{P_{\hat{Y}|X}} \tilde{L}_\lambda(P_{\hat{Y}|X}, P).$$

*Moreover, there exists $P^*_{\hat{Y}|X}$ attaining the infimum in the left hand side, which also attains the infimum on the right hand side when $P$ is fixed to $P^*$, the distribution that attains the supremum on the right hand side.*

Proposition 1 means that in order to design a robust descriptor-estimator pair that work for any $P \in \Gamma$, we only need to design them according to the worst-case distribution $P^*$ as follows.

**Principle of maximum risk-information cost**: Given a convex and closed $\Gamma$, we design the descriptor-estimator pair based on the worst-case distribution

$$P^* = \arg\max_{P\in\Gamma} \inf_{P_{\hat{Y}|X}} \tilde{L}_\lambda(P_{\hat{Y}|X}, P).$$

We then find $P_{\hat{Y}|X}$ that minimizes $\tilde{L}_\lambda(P_{\hat{Y}|X}, P^*)$ and design the descriptor-estimator pair accordingly, e.g. using Lemma 1 on $P_{\hat{Y}|X}$ and the induced distribution $P^*_{\hat{Y}}$ from $P_{\hat{Y}|X}$ and $P^*$.

## IV. Applications

### A. Rate-constrained Minimax Linear Regression

Suppose $\mathbf{X} \in \mathbb{R}^d$, $Y \in \mathbb{R}$, $\ell(\hat{y},y) = (y-\hat{y})^2$ is the mean-squared loss, and we observe the data $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$. Take $\Gamma$ to be the set of distributions with the same first and second moments as given by the empirical distribution, i.e.,

$$\begin{aligned}\Gamma = \big\{P_{\mathbf{X}Y} : \mathsf{E}[\mathbf{X}] = \boldsymbol{\mu}_{\mathbf{X}},\, \mathsf{E}[Y] = \mu_Y,\, \mathrm{Var}[\mathbf{X}] = \Sigma_{\mathbf{X}},\\ \mathrm{Var}[Y] = \sigma^2_Y,\, \mathrm{Cov}[\mathbf{X}, Y] = C_{\mathbf{X}Y}\big\},\end{aligned} \tag{9}$$

where $\boldsymbol{\mu}_{\mathbf{X}}, \mu_Y, \Sigma_{\mathbf{X}}, \sigma^2_Y, C_{\mathbf{X}Y}$ are the corresponding statistics of the empirical distribution. Then the minimax risk-information cost (8) is

$$\tilde{L}^*_\lambda = \sigma^2_Y - C^T_{\mathbf{X}Y}\Sigma^{-1}_{\mathbf{X}}C_{\mathbf{X}Y} + \frac{\lambda}{2}\log\frac{2eC^T_{\mathbf{X}Y}\Sigma^{-1}_{\mathbf{X}}C_{\mathbf{X}Y}}{\lambda\log e}, \tag{10}$$

where the optimal $P^*_{\mathbf{X}Y}$ is Gaussian with its mean and covariance matrix specified in (9), and the optimal estimate

$$\hat{Y} = \begin{cases} aC^T_{\mathbf{X}Y}\Sigma^{-1}_{\mathbf{X}}\mathbf{X} + b + Z & \text{if } \frac{\lambda\log e}{2} < C^T_{\mathbf{X}Y}\Sigma^{-1}_{\mathbf{X}}C_{\mathbf{X}Y} \\ 0 & \text{if } \frac{\lambda\log e}{2} \ge C^T_{\mathbf{X}Y}\Sigma^{-1}_{\mathbf{X}}C_{\mathbf{X}Y}, \end{cases}$$

where

$$a = 1 - \frac{\lambda\log e}{2C^T_{\mathbf{X}Y}\Sigma^{-1}_{\mathbf{X}}C_{\mathbf{X}Y}},\ b = \mu_Y - aC^T_{\mathbf{X}Y}\Sigma^{-1}_{\mathbf{X}}\boldsymbol{\mu}_{\mathbf{X}},$$

and $Z \sim \mathrm{N}(0,\, \sigma^2_Z)$ is independent of $\mathbf{X}$ with $\sigma^2_Z = \frac{\lambda a\log e}{2}$.

The proof of this result is in [16]. Note that this setting does not satisfy the conditions in Proposition 1. We directly analyze (8) to obtain the optimal $P^*_{\mathbf{X}Y}$. Given the optimal $P^*_{\mathbf{X}Y}$, Theorem 2 and Lemma 1 can be used to construct

the scheme. Operationally, $e_n(x,w)$ is a random quantizer of $aC_{\mathbf{X}Y}^T\Sigma_{\mathbf{X}}^{-1}\mathbf{X}+b$ such that the quantization noise follows $\mathrm{N}(0,\sigma_Z^2)$. With this natural choice of the ambiguity set, our formulation recovers a compressed version of the familiar MMSE estimator.

Figure 3 plots the tradeoff between the rate and the risk when $d=1$, $\mu_X=\mu_Y=0$, $\sigma_X^2=\sigma_Y^2=1$, $C_{XY}=0.95$ for the scheme constructed using the Poisson functional representation in [4], with the lower bound given by the minimax risk-information cost $\tilde{L}_\lambda^*$, and the upper bound given in Theorem 2.

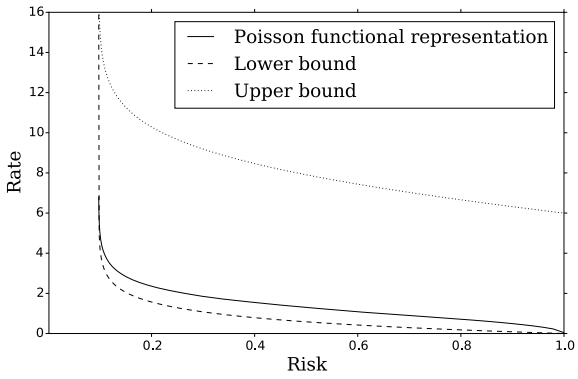


Fig. 3. Tradeoff between the rate and the risk in rate-constrained minimax linear regression.

The optimal scheme in the above example corresponds to compressing and communicating the minimax optimal rate-unconstrained predictor $\bar{Y}=C_{\mathbf{X}Y}^T\Sigma_{\mathbf{X}}^{-1}(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}})+\mu_Y$, since the optimal $\hat{Y}$ can be obtained from $\bar{Y}$ by shifting, scaling and adding noise. This estimate-compress approach can be thought as a *separation* scheme, since we first optimally estimate $\bar{Y}$, then optimally communicate it while satisfying the rate constraint. In the next application, we show that such separation is not optimal in general.

### *B. Rate-constrained Minimax Classification*

Assume $\mathcal{X},\mathcal{Y}$ are finite, $\hat{\mathcal{Y}}=\mathcal{Y}$, $\ell(\hat{y},y)=\mathbf{1}\{\hat{y}\neq y\}$, and $\Gamma$ is closed and convex. It can be shown using Proposition 1 that the minimax risk-information cost is

$$\tilde{L}_\lambda^*=\sup_{P\in\Gamma}\left(1+\lambda\inf_{\tilde{P}_{\hat{Y}}}\mathsf{E}_P\left(-\log\sum_y 2^{\lambda^{-1}P_{Y|X}(y|X)}\tilde{P}_{\hat{Y}}(y)\right)\right), \tag{11}$$

the worst-case distribution $P^*$ is the one attaining the supremum, and the optimal estimator is $P_{\hat{Y}|X}^*(\hat{y}|x)\propto 2^{\lambda^{-1}P_{Y|X}^*(y|x)}\tilde{P}_{\hat{Y}}^*(y)$, where $\tilde{P}_{\hat{Y}}^*$ attains the infimum, and $P_{Y|X}^*$ is obtained from $P^*$. The proof is given in [16]. This result can be simplified if $\Gamma$ is symmetric for different values of $Y$, where we can substitute $\tilde{P}_{\hat{Y}}$ to be the uniform distribution.

To show that the estimate-compress approach is not always optimal, let $\ell(\hat{y},y)=\mathbf{1}\{\hat{y}\neq y\}$, $\mathcal{Y}=\mathcal{Y}_1\cup\mathcal{Y}_2$, where $\mathcal{Y}_1\cap\mathcal{Y}_2=\emptyset$ and $|\mathcal{Y}_i|=k_i$ is finite. Let $\Gamma=\{P\}$, where $P$ is the distribution where $(X_1,X_2)\sim\mathrm{Unif}(\mathcal{Y}_1\times\mathcal{Y}_2)$, and $Y=X_i$ with probability $q_i$ for $i=1,2$. By (11), the optimal risk-information cost is

$$1-\lambda\log\max\Big\{\frac{1}{k_1}(2^{\lambda^{-1}q_1}-1)+1,\ \frac{1}{k_2}(2^{\lambda^{-1}q_2}-1)+1\Big\}, \tag{12}$$

and the optimal scheme is to generate $\hat{Y}$ by compressing $X_1$ (by passing it through a symmetric channel) if $\frac{1}{k_1}(2^{\lambda^{-1}q_1}-1)+1>\frac{1}{k_2}(2^{\lambda^{-1}q_2}-1)+1$, and compressing $X_2$ otherwise. Assume $q_1>q_2$, then the optimal MAP estimate is $\bar{Y}=X_1$. An estimate-compress approach would try to communicate a compressed version of $\bar{Y}=X_1$. However, the optimal rate constrained descriptor communicates a lossy version of $X_2$ rather than $X_1$ if $k_1\gg k_2$. The risk-information cost achieved by the estimate-compress approach is

$$1-\lambda\log\max\Big\{\frac{1}{k_1}(2^{\lambda^{-1}q_1}-1)+1,\ 2^{\lambda^{-1}q_2k_2^{-1}}\Big\},$$

which is larger than (12) when $k_1\gg k_2$.

## V. Acknowledgements

This work was partially supported by a gift from Huawei Technologies and by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370.

## References

[1] F. Farnia and D. Tse, "A minimax approach to supervised learning," in *Advances in Neural Information Processing Systems*, 2016, pp. 4240–4248.

[2] H. Namkoong and J. C. Duchi, "Variance-based regularization with convex objectives," in *Advances in Neural Information Processing Systems*, 2017, pp. 2975–2984.

[3] J. Lee and M. Raginsky, "Minimax statistical learning and domain adaptation with Wasserstein distances," *arXiv preprint arXiv:1705.07815*, 2017.

[4] C. T. Li and A. El Gamal, "Strong functional representation lemma and applications to coding theorems," in *Proc. IEEE Int. Symp. Inf. Theory*, June 2017, pp. 589–593.

[5] P. Harsha, R. Jain, D. McAllester, and J. Radhakrishnan, "The communication complexity of correlation," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 438–449, Jan 2010.

[6] M. Braverman and A. Garg, "Public vs private coin in bounded-round information," in *International Colloquium on Automata, Languages, and Programming*. Springer, 2014, pp. 502–513.

[7] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[8] P. Harremoës and N. Tishby, "The information bottleneck revisited or how to choose a good distortion measure," in *Information Theory, 2007. ISIT 2007. IEEE International Symposium on*. IEEE, 2007, pp. 566–570.

[9] T. A. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *IEEE Transactions on Information Theory*, vol. 60, no. 1, pp. 740–761, 2014.

[10] A. Dembo and T. Weissman, "The minimax distortion redundancy in noisy source coding," *IEEE Transactions on Information Theory*, vol. 49, no. 11, pp. 3020–3030, 2003.

[11] D. Haussler, "A general minimax result for relative entropy," *IEEE Transactions on Information Theory*, vol. 43, no. 4, pp. 1276–1280, Jul 1997.

[12] R. G. Gallager, "Source coding with side information and universal coding," *Technical Report LIDS-P-937, MIT Laboratory for Information and Decision Systems*, 1979.

[13] E. Posner, "Random coding strategies for minimum entropy," *IEEE Transactions on Information Theory*, vol. 21, no. 4, pp. 388–391, Jul 1975.

[14] M. Sion, "On general minimax theorems," *Pacific Journal of mathematics*, vol. 8, no. 1, pp. 171–176, 1958.

[15] P. Elias, "Universal codeword sets and representations of the integers," *IEEE Trans. Inf. Theory*, vol. 21, no. 2, pp. 194–203, Mar 1975.

[16] C. T. Li, X. Wu, A. Ozgur, and A. El Gamal, "Minimax learning for remote prediction," *arXiv preprint*, 2018.