Minimax Learning for Distributed Inference

Cheuk Ting Li, *Member, IEEE*, Xiugang Wu, *Member, IEEE*, Ayfer Özgür, *Member, IEEE*, and Abbas El Gamal, *Life Fellow, IEEE*

Abstract

The classical problem of supervised learning is to infer an accurate estimate of a target variable Y from a measured variable X 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 inference is distributed between two nodes, e.g., a mobile device and a cloud, with a rate constraint on the communication between them. The mobile device observes X and sends a description M of X to the cloud, which 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 lossy source coding problem. We then establish information theoretic bounds on the risk-rate Lagrangian cost which lead to a general method to designing a near-optimal descriptor-estimator pair. A key ingredient in the proof of our result is a refined version of the strong functional representation lemma previously used to establish several one-shot source coding theorems. Our results show that a naive estimate-compress scheme for rate-constrained inference is not in general optimal. The resulting bounds on the risk-rate Lagrangian cost when the distribution of X, Y is known and logarithmic loss is used provide a new one-shot operational interpretation of the information bottleneck.

Index Terms

Minimax learning, distributionally robust learning, information bottleneck, one-shot source coding, functional representation

I. Introduction

The classical problem of supervised learning is to infer an accurate estimate 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 estimators and a loss function that quantifies the risk of each estimator, ERM chooses the estimator with minimal risk under the empirical distribution of samples. To avoid overfitting, the set of allowable estimators is restricted to a class with limited complexity.

Recently, an alternative viewpoint has emerged which seeks distributionally robust estimators. Given the labeled training samples, this approach learns an estimator 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 estimators, it aims to avoid overfitting by requiring that the learned estimator 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 an estimator 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, for example, at a mobile node, and when designing the estimator it is desirable to also take into account the cost of communicating X. This is motivated by the the bandwidth and energy limitations on communication in networks or within multiprocessor systems, which 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 this data (or features of it) are communicated over rate-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 inference in distributed and rate-constrained settings.

In this paper, which is a more complete version of [4], 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,

This paper was presented in part at the IEEE International Symposium on Information Theory, Vail, Colorado, USA, June 2018.

C. T. Li was with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA. He is now with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA 94720, USA (e-mail: ctli@berkeley.edu).

X. Wu was with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA. He is now with the Department of Electrical and Computer Engineering and the Department of Computer and Information Sciences, University of Delaware, Newark, DE 19716, USA (e-mail: xwu@udel.edu).

A. Özgür and A. El Gamal are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA (e-mail: aozgur@stanford.edu, abbas@ee.stanford.edu).

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 estimator 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 estimator under this worst-case distribution. In this paper, we show that an analogous principle approximately holds for rate-constrained inference. 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. A key technical ingredient that we use to design the close to optimal descriptor-estimator pair for the worst-case distribution is the strong functional representation lemma (SFRL) used in [5] to establish several one-shot coding theorems, including for lossy source coding and multiple description coding. We will need a refined version of this lemma, however, to establish our minimax results.

We 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 estimator ignoring the rate constraint, then compress and communicate the estimator 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 inference is rate-constrained, the optimal descriptor 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 inference is logarithmic loss, our results provide a new one-shot operational interpretation of the information bottleneck problem.

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 estimator 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 estimators $\mathcal{F}$ 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 an estimator 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 estimator $f_n \in \mathcal{F}$ with minimal risk under the empirical distribution P_n of the training samples.

Recently, an alternative approach has emerged which seeks distributionally robust estimators. This approach learns an estimator 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 estimators to a low-complexity class to prevent overfitting, in the minimax approach overfitting is prevented by explicitly requiring that the chosen estimator 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.

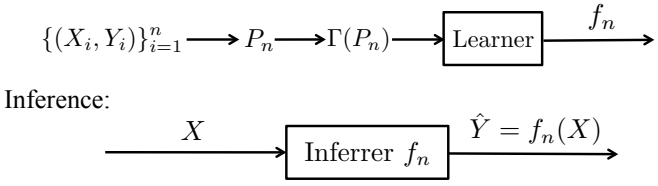

Fig. 1. Minimax approach to supervised learning.

B. Minimax Learning for Distributed Inference

We extend the minimax learning approach to the setting in which the inference 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 inference 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, and the prefix-free codebook for M can depend on W. 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.

Note that the main difference between our minimax learning setup and the minimax noisy source coding problem studied in [6] is that here we are considering the one-shot variable-length setting instead of the asymptotic setting. As such, the proof of our result is quite different from that of the asymptotic setting. In [5], we used the *strong functional representation lemma* (SFRL) to derive an upper bound on the (average) rate-distortion function for the one-shot lossy source coding in terms of the rate distortion function for the asymptotic case. While this proof extends easily to the noisy one-shot lossy compression setting (see Theorem 1 and its proof), we need a refined version of SFRL and several other arguments to establish the corresponding result for the minimax setting in Theorem 2.

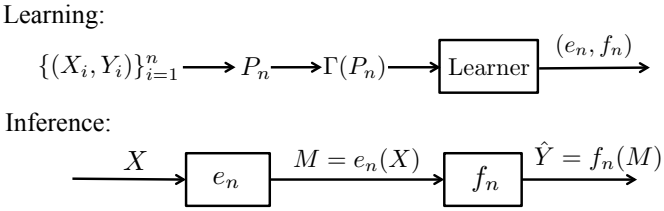

Fig. 2. Minimax learning for distributed inference.

III. Main Results

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

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

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

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

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 the alphabet of $X, Y, \hat{Y}$ are very large, for example, if Y is a large feature set of an image and X is the image itself.

To prove Theorem 1, we use the strong functional representation (also see [7], [8]).

Lemma 1 (see [5])**.** 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 such 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 and the function that generates $\hat{Y}$, we use the Poisson functional representation construction detailed in [5].

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 (Relationship to the information bottleneck). 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

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

which is the information bottleneck function [9]. Therefore the setting of distributed inference provides an approximate one-shot operational interpretation of the information bottleneck (up to a logarithmic gap). In [10], [11] it was shown that the asymptotic lossy source coding from noisy observations problem also provides an operational interpretation of the information bottleneck. Our operational interpretation, however, is more satisfying since the feature extraction problem is by nature one-shot.

We now extend Theorem 1 to the minimax setting.

Theorem 2. Suppose Γ is convex. Then

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

To prove this theorem, we first invoke a minimax result for relative entropy in [12], which generalizes the redundancy-capacity theorem in [13]. We then we apply the following refined version of the strong functional representation lemma which is established in the course of proving Theorem 1 on page 6976 of [5]. We substitute $P_Y \leftarrow \tilde{P}_{\hat{Y}}$ and $P_{Y|X} \leftarrow P_{\hat{Y}|X}$ in the Poisson functional representation (note that while the Poisson functional representation in [5] is stated in terms of P_{XY}, it only depends on P_Y and $P_{Y|X}$). Also see a similar bound in [7].

Lemma 2. 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 for any x, we have $\hat{Y}(k(x, W), W) \sim P_{\hat{Y}|X}(\cdot|x)$, 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. Instead of using the Huffman code as in Theorem 1, we apply a code over positive integers (e.g. Elias delta code [14]) on $k(X,W)$ to produce M.

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:

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

Let

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

Note that g is concave in P for fixed $\tilde{P}_{\hat{Y}}$, since $\mathsf{E}_P[\ell(\hat{Y},Y)]$ 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 [15]. Hence by Fatou's lemma, $\int D\big(P_{\hat{Y}|X=x} \,\big\|\, \tilde{P}_{\hat{Y}}\big)dP(x)$ is lower semicontinuous.

We 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 $\bar{\Gamma_{\hat{Y}}}$ be the closure of $\Gamma_{\hat{Y}}$ in the topology of weak convergence. It can be shown using the same arguments as in [12] (on g instead of relative entropy, and using Sion's minimax thm [16] instead of Lemma 2 in [12]) that if $\Gamma_{\hat{Y}}$ is uniformly tight, then there exists $P^*_{\hat{Y}} \in \bar{\Gamma_{\hat{Y}}}$ 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 [12],

$$\sup_{P\in\Gamma}\inf_{\tilde{P}_{\hat{Y}}} \int D\big(P_{\hat{Y}|X=x} \,\big\|\, \tilde{P}_{\hat{Y}}\big)dP(x) = \infty.$$

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

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

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

We then use Elias delta code [14] for K to produce M. Since the length of the Elias delta codeword for an integer k is upper bounded by $\log k + 2\log\big(\log k + 1\big) + 1$, by Jensen's inequality,

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

Thus,

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

■

Theorem 2 suggests that we can simplify the analysis of the risk-rate cost (2) $L_\lambda = \mathsf{E}_P[\ell(\hat{Y},Y)] + \lambda\,\mathsf{E}_P[|M|]$ by replacing the rate $\mathsf{E}_P[|M|]$ 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^*_λ 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

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

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 [16], which gives the following proposition.

Proposition 1. Suppose $\mathcal{X}$, $\mathcal{Y}$ and $\hat{\mathcal{Y}}$ are finite, Γ is convex and closed, and $\lambda \geq 0$, 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 works 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 Γ, 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 2 on $P_{\hat{Y}|X}$ and the induced distribution $P^*_{\hat{Y}}$ from $P_{\hat{Y}|X}$ and P^*.

IV. Applications

We present two applications of the minimax results discussed in the previous section. The first application shows that with a proper choice of l, Γ, we obtain a rate-constrained linear regression scheme in which the mobile performs linear regression, then communicates a compressed version of it to the cloud. This straightforward estimate-compress scheme is shown not to be optimal in general via a simple classification example.

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 $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$. Let $\boldsymbol{\mu}_{\mathbf{X},n}$, $\mu_{Y,n}$, $\Sigma_{\mathbf{X},n}$, and $C_{\mathbf{X}Y,n}$, respectively, be the empirical means, covariance matrix, and cross covariance matrix estimated from the data. Take Γ to be the set of distributions with these first and second moments, i.e.,

$$\Gamma = \left\{P_{\mathbf{X}Y} : \boldsymbol{\mu}_{\mathbf{X}} = \boldsymbol{\mu}_{\mathbf{X},n},\, \mu_Y = \mu_{Y,n},\, \Sigma_{\mathbf{X}} = \Sigma_{\mathbf{X},n}, \sigma^2_Y = \sigma^2_{Y,n},\, C_{\mathbf{X}Y} = C_{\mathbf{X}Y,n}\right\}, \tag{9}$$

The following proposition shows that for this natural choice of the ambiguity set, the distributions P^* and $P^*_{\hat{Y}|X}$ that achieve the minimax risk-rate cost are both Gaussian.

Proposition 2 (Linear regression with rate constraint)**.** Consider mean-squared loss and define Γ as in (9). Then the distribution that achieves the supremum in Proposition 1, $P^*_{\mathbf{X}Y}$, is Gaussian with its mean and covariance matrix specified in (9), and the optimal estimate and minimax risk-information cost (8) are as follows:

$$\hat{Y} = \begin{cases} a \cdot C^T_{\mathbf{X}Y}\Sigma^{-1}_{\mathbf{X}}(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) + \mu_Y + Z & \text{if } a > 0 \\ \mu_Y & \text{otherwise} \end{cases} \tag{10}$$

$$\tilde{L}^*_\lambda = \begin{cases} \sigma^2_Y - C^T_{\mathbf{X}Y}\Sigma^{-1}_{\mathbf{X}}C_{\mathbf{X}Y} + \frac{\lambda}{2}\log\frac{2eC^T_{\mathbf{X}Y}\Sigma^{-1}_{\mathbf{X}}C_{\mathbf{X}Y}}{\lambda \log e} & \text{if } a > 0 \\ \sigma^2_Y & \text{otherwise,} \end{cases} \tag{11}$$

where

$$a = 1 - \frac{\lambda \log e}{2C^T_{\mathbf{X}Y}\Sigma^{-1}_{\mathbf{X}}C_{\mathbf{X}Y}}, \text{ and } Z \sim \mathrm{N}(0,\, \sigma^2_Z) \text{ is independent of } \mathbf{X} \text{ with } \sigma^2_Z = \lambda a \log e/2.$$

Note that this setting does not satisfy the conditions in Proposition 1. Hence, we analyze (8) directly to obtain the optimal $P^*_{\mathbf{X}Y}$. Given the optimal $P^*_{\mathbf{X}Y}$, Theorem 2 and Lemma 2 can be used to construct the scheme. Operationally, $e_n(x, w)$ is a random quantizer of $a \cdot C^T_{\mathbf{X}Y}\Sigma^{-1}_{\mathbf{X}}(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})$ such that the quantization noise follows $\mathrm{N}(0, \sigma^2_Z)$. Hence, 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 $d = 1$, $\mu_X = \mu_Y = 0$, $\sigma^2_X = \sigma^2_Y = 1$, $\sigma_{XY} = 0.95$ for the scheme constructed using the Poisson functional representation in [5], with the lower bound given by the minimax risk-information cost $\tilde{L}^*_\lambda$, and the upper bound given in Theorem 2. The proof of this proposition is in Appendix A.

The optimal scheme in the above example corresponds to compressing and communicating the minimax optimal rate-unconstrained estimate $\bar{Y} = C^T_{\mathbf{X}Y}\Sigma^{-1}_{\mathbf{X}}(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) + \mu_Y$, 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.

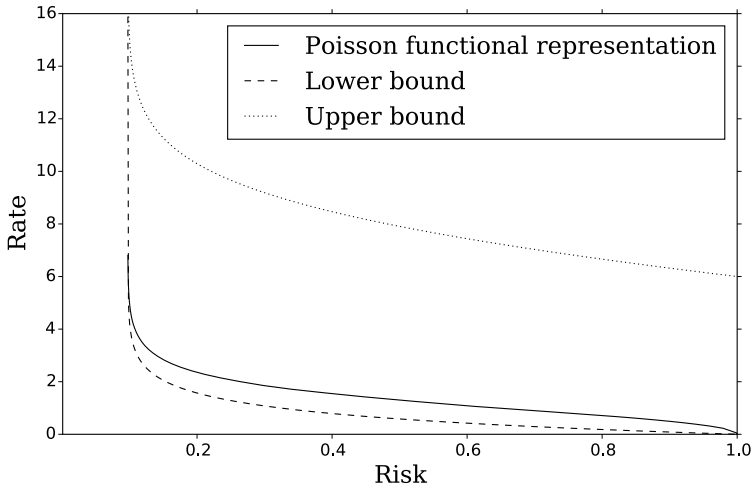

Fig. 3. Tradeoff between the rate and risk in rate-constrained minimax linear regression.

B. Rate-constrained Minimax Classification

Let $\mathcal{Y} = \hat{\mathcal{Y}} = \{1, \ldots, k\}$ and $\mathcal{X}$ be finite, $\ell(\hat{Y}, y) = \mathbf{1}_{\{\hat{y} \neq y\}}$, and Γ be closed and convex. The following gives the minimax risk-information cost and the optimal estimator for this classification setup.

Proposition 3. Consider the setting described above. The minimax risk-information cost is

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

the worst-case distribution P^* is the one attaining the supremum, and the optimal estimator is $P^*_{\hat{Y}|X}(\hat{Y}|x) \propto 2^{\lambda^{-1} P^*_{Y|X}(\hat{Y}|x)} \tilde{P}^*_{\hat{Y}}(\hat{Y})$, where $\tilde{P}^*_{\hat{Y}}$ attains the infimum (when $P = P^*$) and $P^*_{Y|X}$ is obtained from P^*. In particular, if Γ is symmetric for different values of Y (i.e., for any $y_1, y_2 \in \mathcal{Y}$, there exists permutation π of $\mathcal{Y}$, τ of $\mathcal{X}$ such that $\pi(y_1) = y_2$ and $P_{X,Y} \in \Gamma \Leftrightarrow P_{\tau(X),\pi(Y)} \in \Gamma$), then

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

We can see that when $\lambda \to 0$, $P^*_{\hat{Y}|X}$ tends to the MAP estimator (under $\bar{P}^*$, the worst-case distribution when $\lambda = 0$). The proof of this proposition is in Appendix B.

To show that the estimate-compress approach is not always optimal, we consider the following.

Example 1 (Estimate–compress not optimal)**.** Considering the above classification application, Let $\mathcal{Y} = \mathcal{Y}_1 \cup \mathcal{Y}_2$, where $\mathcal{Y}_1 \cap \mathcal{Y}_2 = \emptyset$ and $|\mathcal{Y}_i| = k_i$, $i = 1, 2$. Let $\Gamma = \{P\}$, where P is such that $(X_1, X_2) \sim \mathrm{Unif}(\mathcal{Y}_1 \times \mathcal{Y}_2)$, and $Y = X_i$ with probability q_i for $i = 1, 2$ ($q_2 = 1 - q_1$). By Proposition 3, the optimal risk-information cost is

$$\begin{aligned} &\bar{L}^*_\lambda = 1 - \lambda \log \max \left\{ \frac{a_1}{k_1}, \frac{a_2}{k_2} \right\}, \text{ where} \\ &a_1 = 2^{\lambda^{-1} q_1} + k_1 - 1, \; a_2 = 2^{\lambda^{-1} q_2} + k_2 - 1. \end{aligned} \tag{12}$$

The optimal estimator is as follows. If $a_1/k_1 > a_2/k_2$, then

$$\hat{Y} = \begin{cases} X_1 & \text{w.p. } a_1^{-1} 2^{\lambda^{-1} q_1}, \\ \hat{Y}_1 \sim \mathrm{Unif}(\mathcal{Y}_1 \backslash \{X_1\}) & \text{w.p. } a_1^{-1}, \end{cases} \tag{13}$$

and, if $a_1/k_1 \leq a_2/k_2$, then simply exchange the subscripts 1 and 2 in the above.

If $q_1 > q_2$, then the MAP estimator gives $\bar{Y} = X_1$. An estimate-compress approach would either communicate a compressed version of $\bar{Y} = X_1$ as in (13), or randomly select an element in $\mathcal{Y}_2$ (giving a risk of $1 - q_2/k_2$). The risk-information cost achieved by this approach is

$$\bar{L}_\lambda = 1 - \lambda \log \max \left\{ \frac{a_1}{k_1}, 2^{\lambda^{-1} q_2 k_2^{-1}} \right\}. \tag{14}$$

Now, if $k_1 \gg k_2$, the optimal rate-constrained descriptor (13) communicates a compressed version of X_2, and the risk of estimate–compress in (14) is larger than (12). Moreover, the gap between the rates needed by the two approaches for a fixed risk can be unbounded. Let $q_1 = 1 - q_2 = 2/3$, $k_2 = 2$, $k_1 \geq 15$. The minimum rate needed to achieve a risk $2/3$ is 1 (by

$\hat{Y} = X_2$). For the estimate-compress approach, since $\hat{Y} \sim \mathrm{Unif}(\mathcal{Y}_2)$ gives a risk $5/6$, we have to compress X_1 (by passing it through a symmetric channel with $P\{\hat{Y} = X_1\} = 1/2$) to achieve a risk $2/3$, which as k_1 increases, requires an unbounded rate

$$I(X;\hat{Y}) = H(\hat{Y}) - H(\hat{Y}|X_1) = \log k_1 - \frac{1}{2}\log(k_1 - 1) - \frac{1}{2}.$$

Figure 4 compares risk-rate tradeoff for the optimal scheme, the lower bound obtained from the optimal risk-information tradeoff (12), the upper bound in Theorem 1, and the risk-information cost of the estimate-compress approach (14) for $q_1 = 1 - q_2 = 2/3$, $k_1 = 2^{32}$, $k_2 = 2$. Note that the optimal scheme (attaining the optimal risk-rate tradeoff) performs time-sharing between encoding X_1 using 32 bits with risk $1/3$, encoding X_2 using 1 bit with risk $2/3$, and fixing the output at one value of X_2 with zero rate needed and risk $5/6$. The mutual information needed by the estimate-compress approach (which is a lower bound on the actual rate needed by this approach) is strictly greater than the optimal rate (except when the risk is at its minimum $1/3$ or maximum $5/6$).

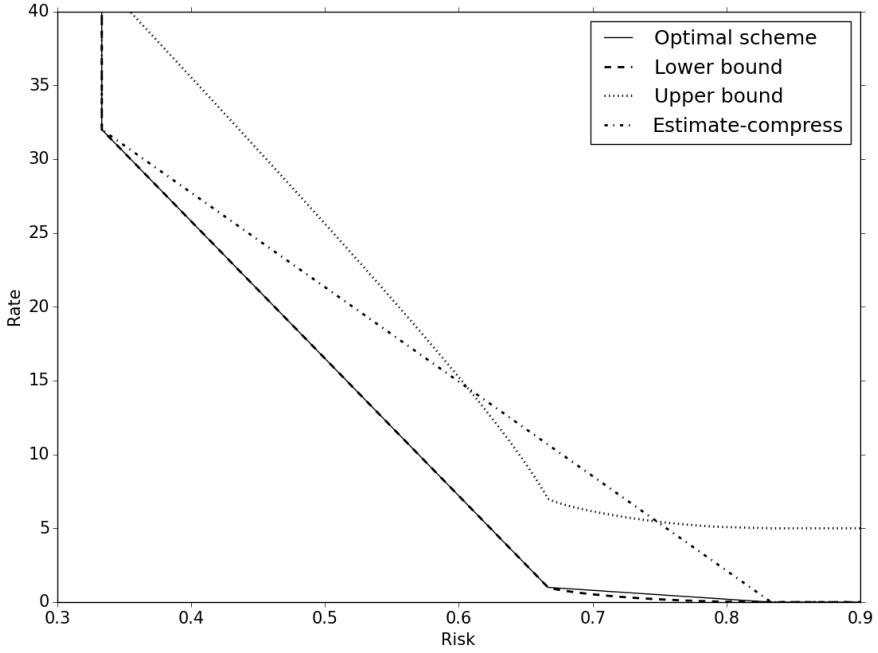

Fig. 4. Tradeoff between the rate and risk in rate-constrained minimax linear classification for the optimal scheme, lower bound (12), upper bound by Theorem 1, and estimate-compress approach (14).

V. Conclusion

We introduced the minimax learning problem in which the inference is distributed between two nodes (e.g., a mobile device and a cloud) with a constraint on the communication rate between them. We showed that the minimax risk-rate cost can be approximated by the minimax risk-information cost, which is significantly easier to evaluate and leads to a general method for the design a near-optimal descriptor-estimator pair. We showed that the naive estimate-compress scheme for rate-constrained inference is not in general optimal. Our results also provide a new one-shot operational interpretation of the information bottleneck and extends it to the minimax robust setting. Designing efficient algorithms for practical applications is left for future research. Extending the work to the case in which the data is also distributed, hence learning has a communication constraint, would also be of great interest to real world applications such as federated learning [17], [18].

VI. 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.

Appendix

A. Proof of Proposition 2

Without loss of generality, assume $\boldsymbol{\mu}_{\mathbf{X}} = \mathbf{0}$ and $\mu_Y = 0$. We first prove "$\leq$ " in (11). For this, fix $P_{\hat{Y}|\mathbf{X}}$ as given in the Proposition and consider any $P \in \Gamma$. When $\frac{\lambda \log e}{2} < C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} C_{\mathbf{X}Y}$, we have

$$\mathsf{E}_P\left[\ell(\hat{Y}, Y)\right] = \mathsf{E}_P\left[(\hat{Y} - Y)^2\right]$$

$$\le \sigma_Y^2 + \frac{\lambda \log e}{2} - C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} C_{\mathbf{X}Y}, \text{ and}$$

$$I_P(\mathbf{X};\hat{Y}) = h(\hat{Y}) - h(\hat{Y}|\mathbf{X})$$

$$\le \frac{1}{2}\log\left(\frac{2C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} C_{\mathbf{X}Y}}{\lambda \log e}\right).$$

Therefore,

$$\inf_{P_{\hat{Y}|X}} \sup_{P\in\Gamma} \left(\mathsf{E}_P\left[\ell(\hat{Y},Y)\right] + \lambda I_P(X;\hat{Y}) \right) \le \text{R.H.S. of (11)}.$$

It can also be checked that the above relation holds when $\frac{\lambda \log e}{2} \ge C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} C_{\mathbf{X}Y}$, and thus we have proved "$\le$" in (11).

To prove "$\ge$ " in (11), fix a Gaussian $P_{\mathbf{X}Y}$ with its mean and covariance matrix specified in (9) and consider an arbitrary $P_{\hat{Y}|\mathbf{X}}$. We have

$$\mathsf{E}_P\left[\ell(\hat{Y},Y)\right] = \mathsf{E}_P\left[(Y-\hat{Y})^2\right]$$

$$= \sigma_Y^2 - C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} C_{\mathbf{X}Y} + \mathsf{E}_P\left[(\hat{Y} - C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1}\mathbf{X})^2\right], \text{ and}$$

$$I_P(X;\hat{Y}) = I_P\big(C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1}\mathbf{X};\, \hat{Y}\big)$$

$$\ge h\big(C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1}\mathbf{X}\big) - h\big(C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1}\mathbf{X} - \hat{Y}\big)$$

$$\ge \frac{1}{2}\log C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} C_{\mathbf{X}Y} - \frac{1}{2}\log \mathsf{E}_P\left[(\hat{Y} - C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1}\mathbf{X})^2\right].$$

Letting $\gamma = \mathsf{E}_P\left[(\hat{Y} - C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1}\mathbf{X})^2\right]$, we have

$$\mathsf{E}_P\left[\ell(\hat{Y},Y)\right] + \lambda I_P(X;\hat{Y}) \ge \sigma_Y^2 - C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} C_{\mathbf{X}Y} + \frac{\lambda}{2}\log C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} C_{\mathbf{X}Y} + \gamma - \frac{\lambda \log \gamma}{2}$$

$$\ge \text{R.H.S. of (11)},$$

where the second inequality follows by evaluating the minimum value of $\gamma - \frac{\lambda \log \gamma}{2}$. Combing this with the above completes the proof of Proposition 2.

B. Proof of Proposition 3

Assume Γ is closed and convex. By Proposition 1, the minimax risk-information cost is $\tilde{L}_\lambda^* = \sup_{P\in\Gamma} \inf_{P_{\hat{Y}|X}} \tilde{L}_\lambda(P_{\hat{Y}|X}, P)$, where

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

$$= \inf_{P_{\hat{Y}|X}} \left(P\{\hat{Y}\neq Y\} + \lambda \inf_{\tilde{P}_{\hat{Y}}} \int D\big(P_{\hat{Y}|X=x} \,\big\|\, \tilde{P}_{\hat{Y}}\big) dP(x) \right)$$

$$= \inf_{\tilde{P}_{\hat{Y}}, P_{\hat{Y}|X}} \left(P\{\hat{Y}\neq Y\} + \lambda \int D\big(P_{\hat{Y}|X=x} \,\big\|\, \tilde{P}_{\hat{Y}}\big) dP(x) \right)$$

$$= 1 + \lambda \inf_{\tilde{P}_{\hat{Y}}, P_{\hat{Y}|X}} \mathsf{E}_P\left(\sum_y P_{\hat{Y}|X}(y|X) \left(\log \frac{P_{\hat{Y}|X}(y|X)}{\tilde{P}_{\hat{Y}}(y)} - \lambda^{-1} P_{Y|X}(y|X) \right) \right)$$

$$= 1 + \lambda \inf_{\tilde{P}_{\hat{Y}}} \inf_{P_{\hat{Y}|X}} \mathsf{E}_P\left(\sum_y P_{\hat{Y}|X}(y|X) \left(\log \frac{P_{\hat{Y}|X}(y|X)}{2^{\lambda^{-1}P_{Y|X}(y|X)} \tilde{P}_{\hat{Y}}(y) / \sum_{y'} 2^{\lambda^{-1}P_{Y|X}(y'|X)} \tilde{P}_{\hat{Y}}(y')} \right) \right.$$

$$\left. - \log \sum_y 2^{\lambda^{-1}P_{Y|X}(y|X)} \tilde{P}_{\hat{Y}}(y) \right)$$

$$\overset{(a)}{=} 1 + \lambda \inf_{\tilde{P}_{\hat{Y}}} \mathsf{E}_P\left(- \log \sum_y 2^{\lambda^{-1}P_{Y|X}(y|X)} \tilde{P}_{\hat{Y}}(y) \right),$$

where (a) is due to the fact that relative entropy is nonnegative and equality is attained when $P_{\hat{Y}|X}(y|x) \propto 2^{\lambda^{-1}P_{Y|X}(y|X)} \tilde{P}_{\hat{Y}}(y)$.

Next we consider the case in which Γ is symmetric. Consider the minimax risk-information cost

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

For any $i,j \in \mathcal{Y} = \{1,\dots,k\}$, let π_{ij} be the permutation over $\mathcal{Y}$ such that $\pi_{ij}(i) = j$ and let τ_{ij} be the corresponding permutation over $\mathcal{X}$ in the symmetry assumption. Since the function

$$P_{\hat{Y}|X} \mapsto \sup_{P\in\Gamma} \tilde{L}_\lambda(P_{\hat{Y}|X}, P)$$

is convex and symmetric about π_{ij} and τ_{ij} (i.e., $\sup_{P\in\Gamma} \tilde{L}_\lambda(P_{\hat{Y}|X}, P) = \sup_{P\in\Gamma} \tilde{L}_\lambda(P_{\pi_{ij}\hat{Y}|\tau_{ij}X}, P)$), to find its infimum, we only need to consider $P_{\hat{Y}|X}$'s satisfying $P_{\hat{Y}|X} = P_{\pi_{ij}\hat{Y}|\tau_{ij}X}$ for all i,j (if not, we can instead consider the average of $P_{\pi_{ij}^a\hat{Y}|\tau_{ij}^a X}$ for a from 1 up to the product of the periods of π_{ij} and τ_{ij}, which gives a value of the function not larger than that of $P_{\hat{Y}|X}$). For brevity we say $P_{\hat{Y}|X}$ is symmetric if it satisfies this condition.

Fix any symmetric $P_{\hat{Y}|X}$. Since the function

$$P \mapsto \tilde{L}_\lambda(P_{\hat{Y}|X}, P)$$

is concave and symmetric about π_{ij} and τ_{ij} (i.e., $\tilde{L}_\lambda(P_{\hat{Y}|X}, P_{X,Y}) = \tilde{L}_\lambda(P_{\hat{Y}|X}, P_{\tau_{ij}X,\pi_{ij}Y})$), to find its supremum, we only need to consider symmetric P's. Hence,

$$\begin{aligned}
\tilde{L}_\lambda^* &= \inf_{P_{\hat{Y}|X}\,\mathrm{symm}} \sup_{P\in\Gamma\,\mathrm{symm}} \tilde{L}_\lambda(P_{\hat{Y}|X}, P) \\
&= \inf_{P_{\hat{Y}|X}\,\mathrm{symm}} \sup_{P\in\Gamma\,\mathrm{symm}} \left(\mathsf{E}_P\left[\ell(\hat{Y},Y)\right] + \lambda I_P(X;\hat{Y})\right) \\
&= \inf_{P_{\hat{Y}|X}\,\mathrm{symm}} \sup_{P\in\Gamma\,\mathrm{symm}} \left(P\{\hat{Y}\neq Y\} + \lambda(\log k - H_P(\hat{Y}|X))\right) \\
&= 1 + \lambda\log k + \lambda \inf_{P_{\hat{Y}|X}\,\mathrm{symm}} \sup_{P\in\Gamma\,\mathrm{symm}} \mathsf{E}_P\left(\sum_y P_{\hat{Y}|X}(y|X)\left(\log P_{\hat{Y}|X}(y|X) - \lambda^{-1} P_{Y|X}(y|X)\right)\right) \\
&= 1 + \lambda\log k + \lambda \inf_{P_{\hat{Y}|X}\,\mathrm{symm}} \sup_{P\in\Gamma\,\mathrm{symm}} \mathsf{E}_P\left(\sum_y P_{\hat{Y}|X}(y|X) \log \frac{P_{\hat{Y}|X}(y|X)}{2^{\lambda^{-1}P_{Y|X}(y|X)}/\sum_{y'} 2^{\lambda^{-1}P_{Y|X}(y'|X)}}\right. \\
&\qquad \left. - \log\sum_y 2^{\lambda^{-1}P_{Y|X}(y|X)}\right) \\
&\geq 1 + \lambda\log k + \lambda \inf_{P_{\hat{Y}|X}\,\mathrm{symm}} \sup_{P\in\Gamma\,\mathrm{symm}} \mathsf{E}_P\left(-\log\sum_y 2^{\lambda^{-1}P_{Y|X}(y|X)}\right) \\
&= \sup_{P\in\Gamma\,\mathrm{symm}} \left(1 + \lambda\log k - \lambda\,\mathsf{E}_P \log\sum_y 2^{\lambda^{-1}P_{Y|X}(y|X)}\right),
\end{aligned}$$

where the inequality is because relative entropy is nonnegative (and equality is attained when $P_{\hat{Y}|X}(y|x) \propto 2^{\lambda^{-1}P_{Y|X}(y|x)}$). Note that

$$1 + \lambda\log k - \lambda\,\mathsf{E}_P \log\sum_y 2^{\lambda^{-1}P_{Y|X}(y|X)} = \inf_{P_{\hat{Y}|X}} \left(P\{\hat{Y}\neq Y\} + \lambda(\log k - H_P(\hat{Y}|X))\right)$$

is an infimum of affine functions of P, hence it is concave in P. Also it is symmetric about π and τ, hence

$$\begin{aligned}
\tilde{L}_\lambda^* &\geq \sup_{P\in\Gamma\,\mathrm{symm.}} \left(1 + \lambda\log k - \lambda\,\mathsf{E}_P \log\sum_y 2^{\lambda^{-1}P_{Y|X}(y|X)}\right) \\
&= \sup_{P\in\Gamma} \left(1 + \lambda\log k - \lambda\,\mathsf{E}_P \log\sum_y 2^{\lambda^{-1}P_{Y|X}(y|X)}\right).
\end{aligned}$$

The other direction follows from setting $\tilde{P}_{\hat{Y}}(y) = 1/k$.

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, X. Wu, A. Özgür, and A. El Gamal, "Minimax learning for remote prediction," in *Proc. IEEE Int. Symp. Inf. Theory*, June 2018, pp. 541–545.

[5] C. T. Li and A. El Gamal, "Strong functional representation lemma and applications to coding theorems," *IEEE Transactions on Information Theory*, vol. 64, no. 11, pp. 6967–6978, Nov 2018.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

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

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

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

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

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

[17] J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," *arXiv preprint arXiv:1610.02527*, 2016.

[18] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. of the International Conference on Artificial Intelligence and Statistics*, 2017, pp. 1273–1282.