Project Report on Minimax Approach for Distributed Inference

Yue Fei 1003944146

***Abstract*—This project is a study on how a minimax-based approach in the classical supervised learning can be extended into a rate-constrained communication problem. The emphasis will be on how a sufficient statistic and an optimal estimator in linear regression tasks exists in distributed learning, with the help of Poisson functional representation at the encoder end. A qualitative simulation is carried out to visually show the performance of the generalized Poisson Functional Representational Lemma (PFR) with a tighter upper bound on the conditional entropy recently introduced.**

I. Introduction

Given the emerging fact that modern mobile devices have access to a rich data-set, that is not only improve individual user's experience on the device, but also suitable for learning models in general. However, this rich data set may only be available remotely, due to privacy sensitivity and extensive sample sizes. Thus, logging between mobile ends and centralized processing servers cannot be communicated cannot be communicated in a lossless approach [1]. The conventional approach is empirical risk minimization (ERM). However, this may lead to over-fitting and not generally lead to a good shared model. Alternative viewpoints towards seeking for distributionally robust estimators are motivated. A minimax game-theoretic setup is suitable in this case, since an estimator in this setup is found by minimizing the worst-case risk over an ambiguity set centered at the empirical distribution of the samples [2].

The original setup for supervised learning in [3] is listed in three main steps (shown in Fig. I): (1) compute the empirical distribution $\hat{P}$ from the data; (2) form a distribution set $\Gamma(\hat{P})$ based on $\hat{P}$; (3) learn a prediction rule ψ^* that minimizes the worst-case expected loss over $\Gamma(\hat{P})$. One major contribution made by [3] is that by using the principle of maximum conditional entropy, the bi-level optimization problem in Step 3 (solid line) can be broken into two disjoint sub problems *3a.* and *3b.* (dashed line). Furthermore, they propose a specific structure for the distribution set $\Gamma(P)$, by moment matching:

1. Match the marginal P X$ofallthejointdistributionsP_X,Yin\Gamma(\hat{P})$ to the empirical marginal P $Matchthecross-momentsbetween$X$andYwiththoseoftheempiricaldistribution$P_X,Y.

This allows us to practically find a sufficient statistics and thus the minimax decision rule ψ^*. As an extension to complete the study on localized supervised learning, Li. et al. proposed a modified framework as in Fig. 2(b). Now the *learner* needs 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 code, such as Huffman code or Elias-delta code), and f is an estimator that takes the compression M and generates an estimate $\hat{Y}$ of Y. A new risk-rate Lagrangian cost with $\lambda > 0$ can be defined as follows:

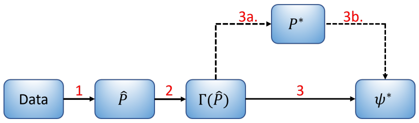


Fig. 1. Minimax Approach for Supervised Learning

$$\mathcal{L}_\lambda(e, f, P) = L(e, f, P) + \lambda R(e, P) \tag{1}$$

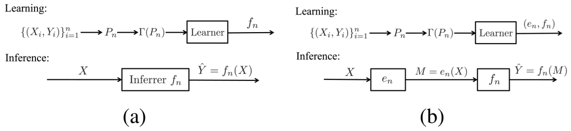
$$R(e, P) \triangleq \mathbf{E}_P[|e(X)|] \tag{2}$$

$$L(e, f, P) \triangleq \mathbf{E}_P[l(f(e(X)), Y)] \tag{3}$$

$$\tag{4}$$

where $R(e, P)$ is the rate of the descriptor e, and can be bounded in terms of the mutual information between X and $\hat{Y}$ using Shannon's entropy. $L(e, f, P)$ is the risk or distortion associated with the descriptor-estimator pair. P is the underlying true distribution of X and Y. The minimax setup is then to minimize the worst-case $L_\lambda(e, f, P)$ over the ambiguity distribution set $\Gamma(P_n)$:

$$(e_n, f_n) = \underset{(e,f)}{\text{argmin}} \max_{P \in \Gamma(P_n)} \mathcal{L}_\lambda(e, f, P). \tag{5}$$



(a)

(b)

Fig. 2. (a) Minimax approach to supervised learning. (b) Minimax learning for distributed inference.

To invoke the *Poisson functional representation lemma* (PFRL), we suppose that the descriptor and estimator share unlimited common randomness W which is independent of the data X. Therefore, e and f can be expressed as functions of (X, W) and (M, W), respectively. The estimator observes X P_X and sends a prefix code M to the descriptor via a noiseless channel such that the estimator can generate Y (from M and W) according to a prescribed conditional distribution $P_{Y|X}$. In other words, this learning model and thus the descriptor-estimator pair are available at both the mobile end and central processing unit. The common randomness seed can be achieved operationally, since the inference scheme is used many times by the same user and by different users. Some bounds on the common random bits and information transfer in distributed computing are also well-studied in [4, 5, 6], but with a more complicated and rigorous protocol setup. Thus, the beauty about *PFRL* is that through an index coding and sampling criteria, the upper bound of communication cost can be proved much more easily compared to the prior works. Further, its upper bound is even constrained within 3.732 bits, which is much tighter than the $\mathcal{O}(1)$ cost term in [7].

II. Methodology

A. Minimax and Maximin

To be able to apply the *Principle of Maximum Conditional Entropy*, Li et al. first proved that the new definition of risk-rate cost, i.e., risk-information cost can lead to a saddle point.

Proposition 1: Suppose $\mathcal{X}$, $\mathcal{Y}$, and $\hat{\mathcal{Y}}$ are finite Γ is convex and closed, and $\lambda \geq 0$, then

$$\hat{\mathcal{L}_{\geq^*}}(-) = \inf_{P_{\hat{Y}|X}} \sup_{P\in\Gamma} \hat{\mathcal{L}}_{\geq}(P_{\hat{Y}|X,P}) = \sup_{P\in\Gamma} \inf_{P_{\hat{Y}|X}} \hat{\mathcal{L}_{\geq^*}}(-)\hat{\mathcal{L}}_{\geq}(P_{\hat{Y}|X,P}) \tag{6}$$

Proof: 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). The lower bound follows from $\mathbf{E}_P[|M|] \geq H_P(M) \geq I_P\ (X; \hat{Y})$ by using Huffman coding. To construct upper bound, an arbitrary $P_{\hat{Y}|X}$ is fixed. Elias delta code over positive integers on $k(X, W)$ to produce M. Using the updated result from generalized Poisson functional representation lemma in section XIII proposition 4 of [8]. The following risk-information cost is achievable:

$$\mathcal{L}' = \sup_{P\in\Gamma} \Bigg(\mathbf{E}_P[l(\hat{Y}, Y)] \tag{7}$$

$$+ \lambda \Bigg(I_P(X, \hat{Y}) + 2\log(I_P(X; \hat{Y}) + 1) + 3.732 \Bigg) \Bigg). \tag{8}$$

Let

$$g(P, \hat{P}_{\hat{Y}}) = \mathbf{E}_P[l(\hat{Y}, Y)] \tag{9}$$

$$+ \lambda \Bigg(\int D(P_{\hat{Y}|X=x} \parallel \hat{P}_{\hat{Y}}) dP(x) \tag{10}$$

$$+ 2\log(\int D(P_{\hat{Y}|X=x} \parallel \hat{P}_{\hat{Y}}) dP(x) + 1) + 3.732 \Bigg) \tag{11}$$

By showing (1) g is concave in P for fixed $\hat{P}_{\hat{Y}}$, since $\mathbf{E}_P[l(\hat{Y}, Y)]$ and $\int D(P_{\hat{Y}|X=x} \parallel \hat{P}_{\hat{Y}}) dP(x)$ are both linear in P. (2) g is quasiconvex in $\hat{P}_{\hat{Y}}$ for fixed P by showing the lower semi-continuity of $\int D(P_{\hat{Y}|X=x} \parallel \hat{P}_{\hat{Y}}) dP(x)$ in $\hat{P}_{\hat{Y}}$. We can then apply Sion's generalized minimax theorem to show that if $\Gamma_{\hat{Y}}$, i.e. the action space of the opponent player is uniformly tight, then we have

$$\inf_{\tilde{P}_{\hat{Y}} \in \overline{\Gamma_{\hat{Y}}}} \sup_{P\in\Gamma} g(P, \tilde{P}_{\hat{Y}}) \leq \sup_{P\in\Gamma} \inf_{\tilde{P}_{\hat{Y}}} g(P, \tilde{P}_{\hat{Y}}) = L' - 0.1\lambda,$$

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. Next, we will present an application of this principle in the case of *squared loss*.

III. Applications

Suppose $X \in \mathbb{R}^d$, $Y \in \mathbb{R}$, $\ell(\hat{y}, y) = (y - \hat{y})^2$ is the mean-squared loss, and we observe the data $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$. Let $^-{}_{\mathbf{X},n}$, $\mu_{Y,n}$, $\Sigma_{\mathbf{X},n}$, $C_{\mathbf{X}Y,n}$ respectively, be the empirical means, covariance matrix, and cross covariance matrix estimated from the data. We design Γ to be the set of distributions with these first and second moments, i.e.,

$$\begin{aligned}\Gamma = \{ P_{\mathbf{X}Y} :\ & \boldsymbol{\mu}_{\mathbf{X}} = \boldsymbol{\mu}_{\mathbf{X},n},\ \mu_Y = \mu_{Y,n}, \\ & \Sigma_{\mathbf{X}} = \Sigma_{\mathbf{X},n}, \sigma_Y^2 = \sigma_{Y,n}^2,\ C_{\mathbf{X}Y} = C_{\mathbf{X}Y,n} \},\end{aligned}$$

We wish to show that the candidate distributions of P^* and $P^*_{\hat{Y}|X}$ that achieves the minimax risk-information cost while satisfying the moment constraints are both Gaussian.

*Proposition 2*Given squared loss, the Bayes decision rule for any $P_{X,Y}$ is the well-known minimum mean-square error (MMSE) estimator that is $\psi_{Bayes(x)} = E[Y|X = x]$. The minimax optimal rate-unconstrained estimate $\bar{Y} = C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} (\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) + \mu_Y$. Suppose with rate constraint, the optimal estimate $\hat{Y}$ is a shifting and scaling version of $\bar{Y}$ with additive noise, as shown in Eq.(12)

$$\hat{Y} = \begin{cases} a \cdot C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} (\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) + \mu_Y + Z & \text{if } a > 0 \\ \mu_Y & \text{otherwise} \end{cases} \tag{12}$$

where

$$a = 1 - \frac{\lambda \log e}{2 C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} C_{\mathbf{X}Y}},$$

and $Z \sim \mathrm{N}(0,\, \sigma_Z^2)$ is independent of $\mathbf{X}$ with $\sigma_Z^2 = \lambda a \log e/2$. Thus, the minimax risk-information cost is (13)

$$\tilde{L}_\lambda^* = \begin{cases} \sigma_Y^2 - C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} C_{\mathbf{X}Y} + \frac{\lambda}{2} \log \frac{2e C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} C_{\mathbf{X}Y}}{\lambda \log e} & \text{if } a > 0 \\ \sigma_Y^2 & \text{otherwise,} \end{cases} \tag{13}$$

The cost intuitively makes sense, the mutual information between X and Y is always greater than 0. To show this lower bound on risk-information cost, we need to prove that $\inf_{P_{\hat{Y}|X}} \sup_{P\in\Gamma}$ and $\sup_{P\in\Gamma} \inf_{P_{\hat{Y}|X}}$ converge. First, we find the *least upper bound*, and secondly prove the *max lower bound*.

1. Without loss of generality, assume $\boldsymbol{\mu}_{\mathbf{X}} = \mathbf{0}$ and $\mu_Y = 0$. Fix $P_{\hat{Y}|\mathbf{X}}$ as satisfying the empirical moment constraints on Γ. Consider any $P \in \Gamma$, we have

$$\begin{aligned} \mathsf{E}_P[\ell(\hat{Y}, Y)] &= \mathsf{E}_P[(\hat{Y} - Y)^2] \\ &\le \mathsf{E}_P[\hat{Y}^2] - 2\mathsf{E}_P[\hat{Y}Y] + \mathsf{E}_P[Y^2] \\ &\le \sigma_Y^2 - \mathsf{E}_P[C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1}(\mathbf{X}Y)] + \mathsf{E}_P[Z^2] \\ &\le \sigma_Y^2 + \frac{\lambda \log e}{2} - C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} C_{\mathbf{X}Y}, \\ I_P(\mathbf{X}; \hat{Y}) &= h(\hat{Y}) - h(\hat{Y}|\mathbf{X}) \\ &\le \frac{1}{2} \log \left(\frac{2C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} C_{\mathbf{X}Y}}{\lambda \log e} \right). \end{aligned}$$

where the large variance is achieved when $a = 1$ $\hat{Y} = C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} \mathbf{X} + Z$. When $C_{\mathbf{X}Y}^T$ and $\Sigma_{\mathbf{X}}^{-1}$ are predetermined, Z is independent of Y. The mutual information holds from definition of differential entropy. The variance of $\hat{Y}$ is $C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} C_{\mathbf{X}Y}$. The variance of $\hat{Y}|X$ is the variance of Z.

2. 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 can rewrite the loss term as

$$\begin{aligned} &\mathsf{E}_P[\ell(\hat{Y}, Y)] \\ &= \mathsf{E}_P[(Y - \hat{Y})^2] \\ &= \underbrace{\sigma_Y^2 - C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} C_{\mathbf{X}Y}}_{\text{Fixed once P* is fixed}} + \mathsf{E}_P[(\hat{Y} - C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} \mathbf{X})^2], \end{aligned}$$

The second term follows from the variance between the current estimate and the optimal estimate given the *worst-case* distribution.

$$\begin{aligned} &I_P(X; \hat{Y}) \\ &= I_P(C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} \mathbf{X}\,;\, \hat{Y}) \\ &= h(C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} \mathbf{X}) - h(C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} \mathbf{X} | \hat{Y}) \\ &\overset{(a)}{\ge} h(C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} \mathbf{X}) - h(C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} \mathbf{X} - \hat{Y}) \\ &\ge \frac{1}{2} \log C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} C_{\mathbf{X}Y} - \frac{1}{2} \log \mathsf{E}_P[(\hat{Y} - C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} \mathbf{X})^2]. \end{aligned}$$

(a) holds since the conditional differential entropy is smaller than the differential entropy of the sum of two random variables, when they are defined on the same interval.

$$h(X_1 + X_2) = \frac{1}{2} \log(2\pi e \cdot (\sigma_1^2 \sigma_2^2))$$

$$h(X_1|X_2) = \frac{1}{2} \log(2\pi e \sigma^2 (1 - \rho^2)),$$

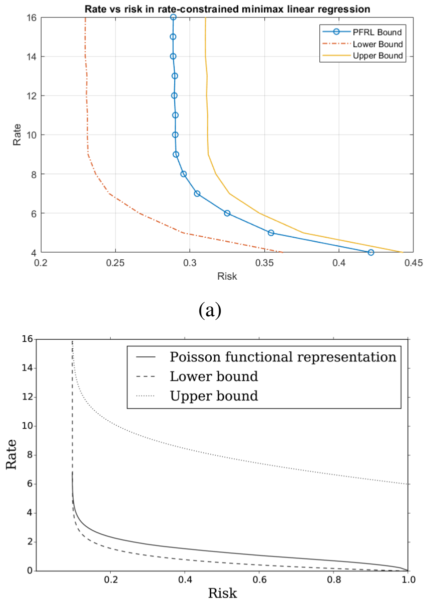
where ρ denotes the correlation coefficient. Letting $\gamma = \mathsf{E}_P[(\hat{Y} - C_{\mathbf{X}Y}^T \Sigma_{\mathbf{X}}^{-1} \mathbf{X})^2]$, and combining $\mathsf{E}_P[\ell(\hat{Y}, Y)]$ and $\lambda I_P(X; \hat{Y})$, we have the maximum lower bound as

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

IV. Experiment

Fig.IV is generated for the case which d=1 , $\mu_X = \mu_Y = 0$, $\sigma_X^2 = \sigma_Y^2 = 1$, $\sigma_{XY} = 0.95$. The lower bound is as obtained in Eq. (13). The upper bound is given in Eq. (8). For PFRL, the bound on conditional entropy $H(Y|W)$ is as given in proposition 4 of [8]. Detailed MATLAB code is attached in Appendix.

In the simulation, $R \in Rate$ denotes the number of bits available for the index source coding. The number of samples of X_i is 2^R. Y_i is generated to be correlated with X_i. Y_i can be considered an optimally estimate $\bar{Y}_i$ at the transmitter end (mobile terminals). $\hat{Y}_i$ represents the target recovered at the receiver end (central node). The calculation for $\mathsf{E}_P[\ell(\hat{Y}, Y)]$ and $I_P(X; \hat{Y})$ follows from the proof of (III).



(a)

(b)

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

V. Conclusion

The application of Poisson Functional Representation Lemma in this rate-constrained linear regression application greatly help me to understand the power of compressive sensing and index coding. More studies on some practical application of such minimax schemes can be a future direction. Since a minimax game-theoretic setup tends to have a generous list of assumptions, which induces significant barriers on the generalization of the solutions derived from such scheme. However, this is not generally the case in this work. Hence, some research on network coding, and the factor graph on computing Bayesian Cramér-Rao's bound [9] can be of the interest.

Appendix A
MATLAB Code

```
clc;
clear all;
close all;


Rate=4:1:16; %rate
M =length(Rate);
nSamples = 2.^Rate;
var_y = ones(1,M);
e = exp(1);
lambda = 0.01;
rho = 0.95;
Risk_lowerB = zeros(1,M);
Risk_upperB = zeros(1,M);
Risk_PFR = zeros(1,M);

Trials=1000;
for i=1:M
    n = nSamples(i);
    tmp_riskL = 0;
    tmp_riskU = 0;
    tmp_riskPFR = 0;
    for t=1:Trials
        x = genX(n);
        y = genY(x,rho);
        Cxy = 0.95;
        varx = var(x);
        vary = var(y);
        c = 0.95^2;
        a = 1-(lambda*log(e))/(2*c);
        if (lambda*log(e))/(2*c) >1
            disp("error")
        end
        sigmaz = lambda*a*log2(e)/c;
        z_rnd = normrnd(0,sigmaz,size(y))
           ;
        y_hat = Cxy'*x/varx ;
        gamma = mean((y-y_hat).^2);

%          I_XY = 0.5*log2(c) - 0.5*log2(
   gamma);
        I_XY =  0.5*log2(2*c / (lambda*
           log2(e)));
        if n==1
            tmp_riskL = vary - c;
        else
            tmp_riskL = tmp_riskL + vary
               - c + gamma + lambda*I_XY;
            tmp_riskU = tmp_riskU + vary
               - c + gamma + lambda*(I_XY
                + 2*log2(I_XY+1)+3.732);
            tmp_riskPFR = tmp_riskPFR +
               vary - c + gamma + lambda
               *(I_XY + log2(I_XY+1)
               +3.732);
        end
```

```
        Risk_lowerB(i) = tmp_riskL/Trials
           ;
        Risk_upperB(i) = tmp_riskU/Trials
           ;
        Risk_PFR(i) = tmp_riskPFR/Trials;
    end
    disp(Risk_lowerB(i))
end
figure
plot(Risk_PFR,Rate,'-o','LineWidth',1,'
   DisplayName','PFRL Bound')
hold on
plot(Risk_lowerB,Rate,'-.','LineWidth',1,
   'DisplayName','Lower Bound')
hold on
plot(Risk_upperB,Rate,'LineWidth',1,'
   DisplayName','Upper Bound')

legend show
xlabel("Risk")
ylabel("Rate")
title("Rate vs risk in rate-constrained
   minimax linear regression")
grid on



function X=genX(N)
    X = normrnd(0,1,N,1);
end

function Y=genY(X,rho)
    Z = normrnd(0,1,size(X));
    Y = rho.*X + sqrt((1-rho^2)).*Z;
end
```

Published with MATLAB R2020a

References

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," 2023.

[2] C. T. Li, X. Wu, A. Özgür, and A. El Gamal, "Minimax learning for distributed inference," *IEEE Transactions on Information Theory*, vol. 66, no. 12, pp. 7929–7938, 2020.

[3] F. Farnia and D. Tse, "A minimax approach to supervised learning," 2017.

[4] A. C.-C. Yao, "Some complexity questions related to distributive computing(preliminary report)," ser. STOC '79. New York, NY, USA: Association for Computing Machinery, 1979, p. 209–213. [Online]. Available: https://doi.org/10.1145/800135.804414

[5] I. Newman, "Private vs. common random bits in communication complexity," *Information Processing Letters*, vol. 39, no. 2, pp. 67–71, 1991. [Online]. Available: https://www.sciencedirect.com/science/article/pii/002001909190157D

[6] R. Canetti and O. Goldreich, "Bounds on tradeoffs between randomness and communication complexity," in *Proceedings [1990] 31st Annual Symposium on Foundations of Computer Science*, 1990, pp. 766–775 vol.2.

[7] M. Braverman and A. Garg, "Public vs private coin in bounded-round information," in *Automata, Languages, and Programming*, J. Esparza, P. Fraigniaud, T. Husfeldt, and E. Koutsoupias, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 502–513.

[8] C. T. Li and V. Anantharam, "A unified framework for one-shot achievability via the poisson matching lemma," *IEEE Transactions on Information Theory*, vol. 67, no. 5, pp. 2624–2651, 2021.

[9] J. Dauwels, "Computing bayesian cramer-rao bounds," in *Proceedings. International Symposium on Information Theory, 2005. ISIT 2005.*, 2005, pp. 425–429.