ECE 543 Final Project
Principle of Maximum Conditional Entropy

Ravi Kiran Raman

Abstract

This report focuses on gaining a better understanding of the principle of maximum entropy and its relation to supervised learning problems. In particular, we aim to understand the minimax expected loss over a family of distributions and of its relation to the maximum conditional entropy. We then study how this may be leveraged to design decision rules that are robust over the set of distributions. Fundamentally, we aim to obtain a better understanding of the role of entropy and information functionals as a more general definition, in learning problems.

I. Introduction

Conventional machine learning problems deal with the design of algorithms that learn a model distribution using training samples and determine a decision rule that generalizes well on more samples from the model. Several methods have been designed and analyzed both theoretically and empirically under various loss functions and probability distributions.

The empirical risk minimizer (ERM) has emerged as almost a go-to first step solution to most learning problems. Exploiting the convergence of empirical estimates to true estimates, we know we can design learning algorithms that are often stable and consistent under appropriate conditions.

While these are desirable properties, it is to be noted that algorithms such as the ERM are hinged on the distribution they get to learn from the training samples and are primarily meant to cater to samples drawn from the same distribution. Let us consider a mildly different setting. Say we have a model of the underlying problem and distribution $\tilde{P}$ and obtain training samples from this model. However, the true samples for the testing phase may subsequently be drawn from a distribution that is close to, but not necessarily the same as $\tilde{P}$.

One such instance of the problem may be envisioned in crowdsourcing. Often enough we are unaware of the true model governing the responses of the crowd workers. In fact, more often than not, different workers are governed by different underlying stochastic models. In such a context, one might be able to learn of the model from a "golden worker". However, another worker subsequently used is likely to follow a model that is probably close to, but different from this model.

Under such a setting, traditional learning algorithms are bound to suffer as they specifically cater to the distribution as learnt from the training data. For instance, it is highly unlikely that ERM would generalize well under such a context. Thus this report focuses on decision rules that minimize the worst case loss in a given family of distributions.

In particular we focus on a minimax problem on the expected loss over a family of distributions as defined in [1]. We then introduce the notion of generalized entropy and information functionals, and study the problem of identifying the distribution that maximizes the conditional entropy [2]. We then observe the fact that the two problems are related at a fundamental level. Subsequently, restricting the set of distributions, we study dual problems that provide a new insight into the problem. We also study the performance of the derived decision rules on the quadratic loss function and draw insights therein.

The maximum entropy principle has been used to design discriminative learning algorithms in the supervised [3] and semi-supervised settings [4]. Further, the principle also sheds light on the robust feature selection problem and it is fundamentally shown that the subset of features that provide the maximum of the minimum information over the family of distributions are the most useful [3], [1].

II. Problem Setting

Consider the learning problem defined by the observation $X \in \mathcal{X}$, and the target variable $Y \in \mathcal{Y}$ distributed according to the joint distribution P. A decision/action is represented as $a \in \mathcal{A}$, and a decision rule is the function

$f : \mathcal{X} \to \mathcal{A}$. Let the set of all such decision rules (randomized rules included) be given by $\mathcal{F}$. Let the loss function be given by $L : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}^+$.

A decision rule f_B is the Bayes decision rule if for any $f \in \mathcal{F}$,

$$\mathbb{E}\left[L(Y, f_B(X))\right] \leq \mathbb{E}\left[L(Y, f(X))\right].$$

In this work, we will assume that there exists no constraint on the set of decision rules. Thus, a Bayes decision rule also satisfies the following

$$\mathbb{E}\left[L(Y, f_B(X))|X = x\right] \leq \mathbb{E}\left[L(Y, f(X))|X = x\right],$$

for all $x \in \mathcal{X}$ and $f \in \mathcal{F}$. Let the training samples be $(X_1, Y_1), \ldots, (X_n, Y_n)$, and the corresponding empirical distribution be

$$\hat{P}_n(x, y) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left\{X_i = x, Y_i = y\right\}.$$

Consider a set of distributions Γ. The minimax problem of principal interest is given by

$$f^* = \arg\min_{f \in \mathcal{F}} \max_{P \in \Gamma} \mathbb{E}\left[L(Y, f(X))\right], \tag{1}$$

where the solution f^* is referred to as a robust Bayes decision rule. The problem was studied from a Game theoretic perspective originally in [2] and was related to the maximum conditional entropy problem as we shall see later.

Before we understand this problem and the solution in more detail, we first introduce the notion of generalized entropy and conditional entropy functionals.

III. Generalized Entropy Functionals

Shannon's classic definitions of entropy and information functionals [5] for a random variables $X, Y \sim P_{X,Y}$ are given by

$$H_S(X) = \mathbb{E}\left[-\log(P_X(X))\right], \quad H_S(Y|X) = \mathbb{E}_X\left[H(Y|X = x)\right], \quad I_S(X; Y) = H(Y) - H(Y|X).$$

The entropy quantifies the extent of randomness while the conditional entropy is quantifies the residual randomness given another random variable. The mutual information is representative of the number of bits of information about one random variable that another provides.

The generalized versions of these quantities were defined in [2]. They are given in reference to an associated loss function L. The generalized entropy (hereafter referred to as entropy) is defined as the minimum loss in predicting Y without any knowledge of X, i.e.,

$$H(Y) = \inf_{a \in \mathcal{A}} \mathbb{E}\left[L(Y, a)\right]. \tag{2}$$

The conditional entropy given $X = x$ is

$$H(Y|X = x) = \inf_{f \in \mathcal{F}} \mathbb{E}\left[L(Y, f(X))|X = x\right], \tag{3}$$

and in turn, the conditional entropy is

$$H(Y|X) = \sum_{x} P_X(x) H(Y|X = x) = \inf_{f \in \mathcal{F}} \mathbb{E}\left[L(Y, f(X))\right]. \tag{4}$$

Note that the sum is appropriately replaced by the corresponding expectation for continuous random variables. It is also worth noting here that the conditional entropy is exactly equal to the Bayes error probability.

The mutual information here is defined as the reduction in expected loss obtained through the knowledge of X and is given as

$$I(X; Y) = H(Y) - H(Y|X). \tag{5}$$

As is the case with Shannon's information, this definition of information is also non-negative owing to Jensen's inequality.

It is worth noting that under logarithmic loss given by $L(y, Q_y) = -\log Q_Y(y)$, the generalized entropy and information functionals reduce to the Shannon definitions. The 0-1 loss, given by $L(y, \hat{y}) = \mathbf{1}\{y \neq \hat{y}\}$, is also quite common in the learning framework. Under this loss function,

$$H_{\text{0-1}}(Y) = 1 - \max_{y \in \mathcal{Y}} P_Y(y), \quad H_{\text{0-1}}(Y|X) = 1 - \sum_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} P_{X,Y}(x, y).$$

We can easily observe that the conditional entropy is the Bayes error corresponding to the MAP estimate.

Another loss function that we shall in particular consider in more detail later is the quadratic loss given by $L(y, \hat{y}) = (y - \hat{y})^2$, and has

$$H_2(Y) = Var(Y), \quad H_2(Y|X) = \mathbb{E}\left[Var(Y|X)\right].$$

Having defined the generalized entropy functionals, we will now study the principle of maximum conditional entropy.

IV. Principle of Maximum Conditional Entropy

For a given set of distributions $\mathcal{P}$, the maximum entropy problem is defined as

$$H^* = \sup_{P \in \mathcal{P}} H(P). \tag{6}$$

Through the definition of the generalized entropy functionals, [2] established the first relation between the problems (6) and (1). In particular, possibly the strongest result therein is that the robust Bayes decision rule is the Bayes decision rule corresponding to the distribution that maximizes the conditional entropy of Y given X.

This result has also been stated elaborately under the learning perspective in [1], and for posterity and continuity, we include these theorems here. This theorem is arguably the fundamental principle and possibly the biggest take back in this project.

Theorem 1:

1) *Weak Version:* If Γ is convex and closed, L is bounded, $\mathcal{X}, \mathcal{Y}$ are finite, and the risk set $S = \{[L(y, a)]_{y \in \mathcal{Y}} : a \in \mathcal{A}\}$ is closed, then there exists a robust Bayes rule f^* which is a Bayes rule for $P^* = \arg\max_{P \in \Gamma} H(Y|X)$.
2) *Strong Version:* Assume that Γ is convex and there exists a Bayes rule for every $P \in \Gamma$, such that they are continuous over Γ. Then, if $P^* = \arg\max_{P \in \Gamma} H(Y|X)$, then any Bayes rule of P^* is a robust Bayes decision rule.

Note that these same results were derived earlier in the Game Theory framework in [2, Theorems 4.1, 5.2] and are fundamentally the same results.

In summation, the above theorem establishes that any decision rule f is a robust Bayes decision rule if and only if it is a Bayes decision rule for P^* under appropriate conditions. Thus it is evident that in order to determine the robust Bayes rule, it would suffice to find the distribution that maximizes the conditional entropy and determine the Bayes rule corresponding to this distribution.

However, this is not easy in practice unless the underlying set of distributions are defined with desirable constraints. We shall now see how duality results for appropriately constrained sets of distributions leads to pleasing optimization problems.

V. Duality under Restricted Set of Distributions

A. Mean Constrained, Linear Loss Models

The problem of determining distributions with maximum entropy (MaxEnt distributions) such that a certain statistic of the random variable satisfies a mean constraint is quite prominent [5]. Thus to provide a simplified setting to better understand the robust Bayes rule, [2] consider the mean constrained models. Specifically, let $\theta(Y) \in \mathbb{R}^t$ be a vector-valued statistic of the target Y. Then, consider the set of distributions constrained as follows:

$$\Gamma(\tau) = \{P : \mathbb{E}\left[\theta(Y)\right] = \tau\}.$$

Under such a restriction, the Lagrangian corresponding to (6) restricted to $\Gamma(tau)$ is given by

$$\mathcal{L}(P, \beta) = H(Y|X) - \beta^T \mathbb{E}\left[\theta(Y)\right],$$

where β is the vector of Lagrange multipliers. The dual framework is thus evident from the above formulation.

A decision rule f is termed "linear" if there exists $\beta_0 \in \mathbb{R}$, $\beta \in \mathbb{R}^t$, such that

$$L(y, f(x)) = \beta_0 + \beta^T \theta(y),$$

for all $x \in \mathcal{X}, y \in \mathcal{Y}$. Then under such restricted models, [2, Theorem 7.1] offers sufficient conditions for robust Bayes decision rules. In particular, a linear decision rule under the mean constraint is robust Bayes. However, such a context is quite narrow and does not cater to more general loss models and distributions.

B. Models with Bounded Deviation Constraints on Mean

A broader set of distributions in comparison to the mean-constrained set is considered in [1] in terms of the deviation of the correlations with some statistic. In particular, let $\theta(Y) \in \mathbb{R}^t$ again be some statistic of Y. Let $\|\cdot\|$ be a given norm. The set of distributions centered at Q is defined as

$$\Gamma(Q) = \{P : P_X = Q_X, \forall i \in [t], \|\mathbb{E}_P[\theta_i(Y)X] - \mathbb{E}_Q[\theta_i(Y)X]\| \leq \epsilon_i\}, \tag{7}$$

where $\epsilon_i \geq 0$ are the slack variables. That is, [1] considers the set of all distributions with the same marginals of X, such that the deviation in the mean of the correlation of X with $\theta_i(Y)$ is bounded up to a given degree of slackness. Consider the special case wherein $\epsilon_i = 0$ for all i. Then we are evidently imposing a mean constraint of a slightly different nature.

A significant result of [1] is the duality of the maximum conditional entropy problem with the maximum likelihood estimation on a generalized linear model. Define $F_\theta : \mathbb{R}^t \to \mathbb{R}$ as

$$F_\theta(z) = \max_{P \in \mathcal{P}_Y} H(Y) + \mathbb{E}\left[\theta(Y)\right]^T z. \tag{8}$$

Then, there exists the following duality result

Theorem 2:

$$\max_{P \in \Gamma(Q)} H(Y|X) = \min_{A \in \mathbb{R}^{t \times d}} \mathbb{E}_Q\left[F_\theta(AX) - \theta(Y)^T AX\right] + \sum_{i=1}^{t} \epsilon_i \|A_i\|_*, \tag{9}$$

where $\|A_i\|_*$ is the dual norm of $\|\cdot\|$ on the ith row of A. Further,

$$\mathbb{E}_{P^*}\left[\theta(Y)|X = x\right] = \nabla F_\theta(A^* X), \tag{10}$$

for all $x \in \mathcal{X}$, where $P^* = \arg\max_{P \in \Gamma(Q)} H(Y|X)$, and A^* is the solution to the minimization problem in (9).

Theorem 2 essentially exploits the linearity in the constraint set to invoke Slater's criterion to establish the duality. In (9), A is the matrix of Lagrange multipliers. The result as is does not provide any added insight into the maximum entropy problem. We now analyze the dual minimization problem.

Consider the exponential family of distributions with sufficient statistic $\theta(Y)$, parameter η, and log-partition function $F_\theta(\eta)$, given by

$$p(y|\eta) = h(y) \exp\left(\eta^T \theta(Y) - F_\theta(\eta)\right). \tag{11}$$

Let the parameter as a function of X be a linear predictor given by $\eta(X) = AX$.

Then, if $\epsilon_i = 0$ for all i, then the minimization problem essentially determines the maximum likelihood linear predictor of the model parameter of (11). In particular, the problem is essentially the same as a regularized version of a maximum likelihood linear fit for the generalized linear model.

In particular, we see that for $\theta(Y) \sim p(Y|\eta(x))$,

$$\mathbb{E}\left[\theta(Y)|X = x\right] = \nabla F_\theta(\eta(x)).$$

From (10), we see that the distribution maximizing the conditional entropy is related to the corresponding exponential family according to (11) in terms of the conditional expectations.

In particular, note that if we choose $\theta(\cdot)$ such that $\mathbb{E}\left[\theta(Y)|X = x\right]$ is representative of the Bayes decision rule, then we also get the robust Bayes rule by computing the gradient of log-partition function.

We elaborate the process of determining the robust Bayes rule for the quadratic loss for better insight. When $L(y, \hat{y}) = (\hat{y} - y)^2$, the Bayes decision rule is the conditional mean estimate. This inspires the choice of $\theta(Y) = Y$. Then the set of distributions is given by

$$\Gamma(Q) = \{P : P_X = Q_X, \|\mathbb{E}_P[YX] - \mathbb{E}_Q[YX]\| \leq \epsilon\}.$$

The Lagrange multipliers can now be chosen to be $a \in \mathbb{R}^d$.

Further, for the quadratic loss, $H_2(Y) = Var(Y)$. Thus, if $\mathcal{P}_Y$, the set of all distributions of Y over which we compute the maximum entropy problem, includes a distribution with infinite second moment, then the objective diverges to ∞. To avoid such contexts, let us include the constraint that $\mathcal{P}_Y = \{P : \mathbb{E}_P[Y^2] \leq \rho^2\}$ for some sufficiently large parameter ρ.

Then, we can solve for $F_\theta(\cdot)$ to obtain

$$F_\theta(z) = \begin{cases} \frac{z^2}{4} + \rho^2 & , \text{ if } |z| \leq 2\rho, \\ \rho|z| & , \text{ if } |z| \geq 2\rho. \end{cases} \tag{12}$$

In particular, if ρ is chosen sufficiently high such that $|a^T X| \leq 2\rho$ almost surely, then it suffices to consider the first case for the expression of $F_\theta(\cdot)$.

The dual minimization problem is then given by

$$\min_{a \in \mathbb{R}^d} a^T \mathbb{E}_Q[XX^T]a - a^T \mathbb{E}_Q[XY] + \epsilon\|a\|_* + \rho^2, \tag{13}$$

which is a regularized quadratic program and hence often relatively easy to solve.

Further, note that

$$\nabla F_\theta(z) = \begin{cases} -\rho & , \text{ if } z \leq -2\rho, \\ \frac{z}{2} & , \text{ if } |z| \leq 2\rho \\ \rho & , \text{ if } z \geq 2\rho. \end{cases} \tag{14}$$

Thus, from (10), the robust Bayes decision rule is given by

$$f^*(x) = \mathbb{E}_{P^*}[Y|X = x] = \frac{\langle a^*, x\rangle}{2}. \tag{15}$$

That is, the robust Bayes estimator is directly obtained to be a linear estimate.

While the duality result is particularly pleasing in examples such as highlighted above, it is not entirely clear what the specific constraint on the set of distributions represent. The linear nature of the constraints on $P_{Y|X}$ enable the use of Slater's criterion and thus the duality. However, if we were to consider families of distribution that are instead of the form

$$\Gamma(Q) = \{P : D_f(P\|Q) \leq \epsilon\},$$

then it is not clear how the problem transforms.

In fact, for the simple case of distributions within a ball of radius ϵ in total variation from the center, the resulting maximum entropy problem has a non-trivial dual and no clear simplification of the problem is obtained.

C. Discriminative Learning for Output-controlled Families

Yet another method for addressing a problem of this nature was proposed in [3] through the notion of the minimum mutual information framework. In this paper the authors restrict the focus to Shannon's mutual information and argue that the decision rule that the distribution corresponding to the minimum mutual information caters to the worst loss.

In particular, the authors consider a framework of mean-matched distributions that have the same marginal distribution of the target variable, i.e,

$$\Gamma_I(Q) = \{P : P_Y = Q_Y, \mathbb{E}_{P_{X|Y=y}}[\phi(X)] = \mathbb{E}_{Q_{X|Y=y}}[\phi(X)], \forall y \in \mathcal{Y}\}. \tag{16}$$

Under such a system of distributions, since $H(Y)$ is a constant, the minimum mutual information corresponds to the maximum conditional entropy principle of the logarithmic loss.

The minimum information problem is given by

$$P_I^* = \arg\min_{P \in \Gamma_I(Q)} I(P(X;Y)), \tag{17}$$

where $I(p)$ is the mutual information corresponding to the joint distribution p. The solution to the problem using Lagrange multipliers is given by

$$P_I^*(x|y) = P_I^*(x) \exp\left(\psi(y)^T \phi(x) + \gamma(y)\right), \tag{18}$$

where $\psi(y)$ is the vector of Lagrange multipliers and $\gamma(y)$ is the normalization term.

In particular,

$$I^* = I(P_I^*) = \mathbb{E}_{P_Y}[\gamma(Y) + \psi(Y)^T \mathbb{E}_{Q_{X|Y}}[\phi(X)]], \tag{19}$$

and

$$P_I^*(y|x) = P_Y \exp\left(\psi(y)^T \phi(x) + \gamma(y)\right). \tag{20}$$

Using this formulation, the authors go on to construct the dual problem and design disciminative learning algorithms using the structure therein.

VI. Generalization Bound for Worst-case Risk

In [1], for the set of distributions as defined in (7), the authors also establish a generalization bound on the worst-case risk. In particular, let $\hat{f}_n$ be the robust Bayes decision rule of $\Gamma(\hat{P}_n)$ obtained by solving the dual problem in (9). Similarly, let $\tilde{f}$ be the robust Bayes decision rule of $\Gamma(\tilde{P})$, where $\tilde{P}$ is the true distribution of the training samples.

Theorem 3: Let $M = \max_{P \in \mathcal{P}_Y} H(Y)$, $\|\cdot\|$ be the ℓ_p norm, and $\|X\|_p \leq B$, $\|\theta(Y)\|_2 \leq L$ almost surely. Then, for any $\delta > 0$,

$$\max_{P \in \Gamma(\tilde{P})} \mathbb{E}\left[L(Y, \hat{f}_n(X))\right] - \max_{P \in \Gamma(\tilde{P})} \mathbb{E}\left[L(Y, \tilde{f}(X))\right] \leq \frac{4BLM}{\sqrt{n}} \left(\sqrt{2(p-1)} + \sqrt{\frac{9\log(4/\delta)}{8}}\right) \sum_{i=1}^{t} \frac{1}{\epsilon_i} \tag{21}$$

holds with probability at least $1 - \delta$.

The above theorem indicates that the worst-case performance of the empirical robust Bayes estimator is asymptotically as good as that of the optimal robust Bayes estimator. Thus learning the model through the empirical distribution is validated in the minimax sense.

However, we note that the algorithms obtained do not necessarily generalize well as the training distribution is not necessarily the worst distribution in the family. To see this, consider the binary hypothesis test governed by $\mathcal{X} = \mathcal{Y} = \{-1, 1\}$ under the 0-1 loss. Let $X \sim \text{Bern}(1/2)$ and $Y = X + Z$, where $Z \sim \text{Bern}(\epsilon)$ for some $\epsilon < \delta$. Let $\theta(Y) = Y$. Let the distribution of the training samples be drawn for the case that $\epsilon = 0$. Then, the family of distributions is given by

$$\Gamma(\tilde{P}) = \{P : P_X = \text{Bern}(1/2), 2\epsilon \leq \tilde{\epsilon}\}.$$

Now clearly for any decision rule, there exists a distribution in $\Gamma(\tilde{P})$ such that the difference in training and test error is strictly bounded away from 0 irrespective of the size of the training data set. This in turn highlights the fact that the robust Bayes decision rules are not always generalizable/stable.

Further, it is not outrightly clear if the robust Bayes decision rules are consistent under some additional constraints. Note that Theorem 3 only bounds the difference in the worst-case loss values. Intuitively we do expect that for well-behaved loss functions and probability distributions, the robust Bayes decision rules will in fact be consistent. However this question is open and warrants further consideration.

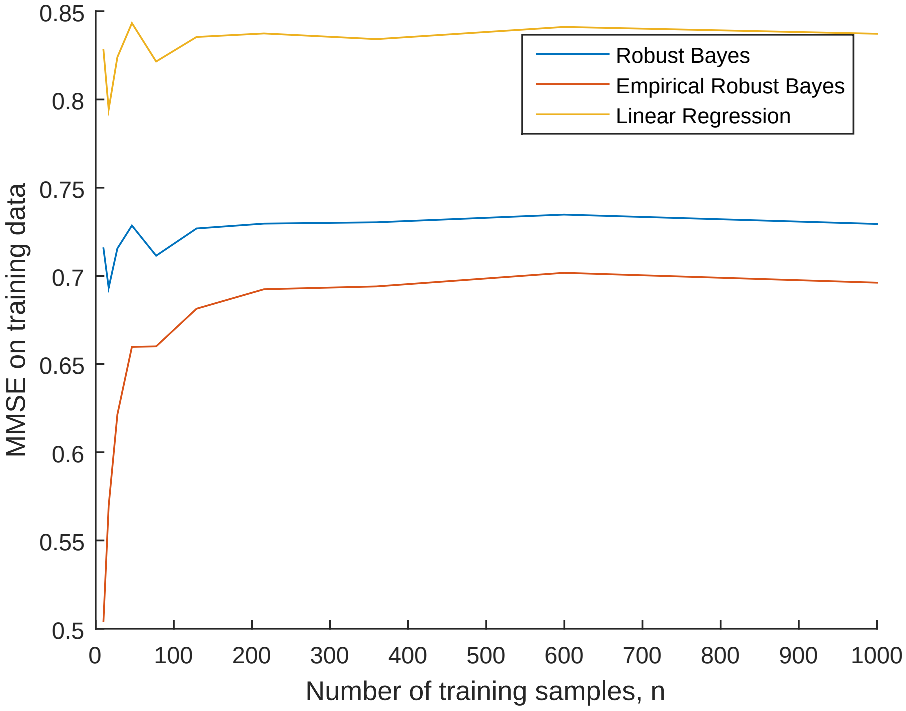

Fig. 1. Error on the training data: The empirical estimator converges to the optimal robust Bayes estimate as the number of samples increases.

VII. Simulations

Again consider the quadratic loss function. Let $Y \sim \mathcal{N}(0,1)$ be a message, and $\mu \in \mathbb{R}^d$ be some known signal. Consider a system that transmits Y as an amplitude modulated (AM) signal over μ. Let $Z \sim \mathcal{N}(0, Sigma)$ be additive noise in the system. Then the receiver receives $X = Y\mu + Z$.

Then, under $\tilde{P}$, the dual minimization problem for the quadratic loss function is obtained from (13) as

$$\min_{a\in\mathbb{R}^d} a^T(\mu\mu^T + \Sigma)a - a^T\mu + \epsilon\|a\|_* + \rho^2,$$

and can be appropriately solved and thus the robust Bayes decision rule is obtained. On the other hand, the empirical estimates of the correlation and cross-correlation are obtained to determine the corresponding empirical robust Bayes decision rule.

While it is of interest to determine the worst-case loss under each estimator, determining the distribution corresponding to this is still analytically tedious. Thus, we consider a corrupted model wherein samples are given by $(X, \tilde{Y})$, where $\tilde{Y} = Y + W$ for some independent noise W. However, instead of considering a single model for the noise W, we consider a random collection of distributions that satisfy the slackness conditions in $\Gamma(\hat{P}_n)$. By averaging over several such noise models, we obtain a lower bound on the worst-case loss for each case. For comparison, we also consider the linear regression estimate for comparison of the performance of the robust Bayes estimates.

Simulating this system model for different sizes of the training set, we determine the loss on the training data and that on the noisy test model. First, in Fig. 1 we plot the performance of the estimators on the training data. As can be observed, the empirical estimator converges to the optimal robust Bayes estimator. Further, it is to be noted that the error for the empirical estimate increases with the number of training samples. This is fundamentally owing to the fact that the we get to learn the model better with samples and thus the optimize for the family of distributions better. That is, it generalizes (becomes more robust over $\Gamma(\tilde{P})$) as n increases.

Similarly, Fig. 2 computes the error on the noisy test model described above. It can be seen again that the empirical estimate generalizes better with more training samples to eventually converge to the optimal robust Bayes estimate. Both the plots here seem to reemphasize the belief that the robust Bayes estimator could in fact be consistent.

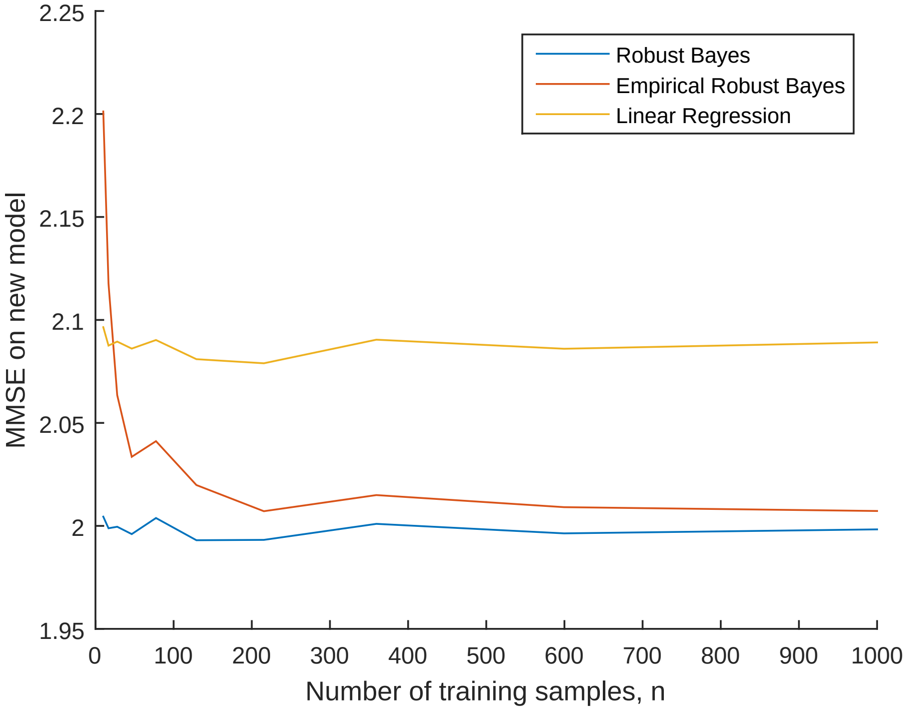

Fig. 2. Error on the noisy test data: The empirical estimator converges to the optimal robust Bayes estimate as the number of samples increases.

VIII. Conclusion

We thus obtained a glimpse into the problems of minimax supervised learning and maximum conditional entropy, and their inherent similarity. We also saw the formulations of the dual of the maximization problem that yield added insight and perspective to the nature of the solution.

The study however does leave a lot of open questions. In particular, it is certainly worth exploring if the robust Bayes estimates are consistent under added constraints. Further, the dual of the maximum entropy problem derived here is sensitive to the set of constraints imposed here and does not generalize to other families of distributions as easily. It would be of interest to determine if such results can be obtained for a broader class of distributions.

References

[1] F. Farnia and D. Tse, "A minimax approach to supervised learning," in *Advances In Neural Information Processing Systems*, 2016, pp. 4233–4241.

[2] P. D. Grünwald and P. Dawid, "Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory," *Annals of Statistics*, pp. 1367–1433, 2004.

[3] A. Globerson and N. Tishby, "The minimum information principle for discriminative learning," in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*. AUAI Press, 2004, pp. 193–200.

[4] A. Erkan and Y. Altun, "Semi-supervised learning via generalized maximum entropy." in *AISTATS*, 2010, pp. 209–216.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley & Sons, 1991.