Discrete Chebyshev Classifiers

Elad Eban* ELADE@CS.HUJI.AC.IL
Elad Mezuman† ELAD.MEZUMAN@MAIL.HUJI.AC.IL
Amir Globerson* GAMIR@CS.HUJI.AC.IL
†Edmond and Lily Safra Center for Brain Sciences. The Hebrew University of Jerusalem
*The Selim and Rachel Benin School of Computer Science and Engineering. The Hebrew University of Jerusalem

Abstract

In large scale learning problems it is often easy to collect simple statistics of the data, but hard or impractical to store all the original data. A key question in this setting is how to construct classifiers based on such partial information. One traditional approach to the problem has been to use maximum entropy arguments to induce a complete distribution on variables from statistics. However, this approach essentially makes conditional independence assumptions about the distribution, and furthermore does not optimize prediction loss. Here we present a framework for discriminative learning given a set of statistics. Specifically, we address the case where all variables are discrete and we have access to various marginals. Our approach minimizes the worst case hinge loss in this case, which upper bounds the generalization error. We show that for certain sets of statistics the problem is tractable, and in the general case can be approximated using MAP LP relaxations. Empirical results show that the method is competitive with other approaches that use the same input.

1. Introduction

Many machine learning algorithms operate on labeled datasets where a set of data points $\mathbf{x}$ and their labels y are provided. However, it is not always realistic to assume such data can be gathered and stored in this form. For example, in medical informatics we often wish to perform diagnostic prediction based on information about the patients (e.g., results of blood tests, personal history etc). Obtaining complete data instances for this case may be impossible due to privacy concerns. However, it may be easier to obtain data such as the probability of a given blood test being abnormal given that the patient has a particular disease. As another example, consider a router of a large internet provider. The number of packets it needs to process is huge, and performing any learning on those would require some sort of aggregation.

Thus, it is of interest to learn classifiers based on partial or aggregated information. Here we focus on the important case where features $x_1, \ldots, x_n$ are discrete and categorical. A natural summary statistic in this case is low order marginals such as $p(x_i, y)$ or $p(x_i, x_j, y)$. These can be estimated reliably given small amounts of data. The question is then how to use these to build a classifier of y for a complete instance $x_1, \ldots, x_n$.

The challenge in the above scenario is that we only have partial information about the true joint distribution $p(x_1, \ldots, x_n, y)$. Namely, its first and second order marginals. A common approach in this case is to assume that the true distribution is the one with maximum entropy subject to these marginal constraints. For first order marginals this results in the popular Naive Bayes classifier, whereas for second order it results in Tree Augmented Naive Bayes (Friedman et al., 1997). However, these approaches do not try to optimize prediction error. They implicitly make conditional independence assumptions about the joint distribution and then use this joint distribution for prediction.

Here we take a strictly discriminative approach to the above problem. Given a set of observed marginals μ we consider the set of distributions $\mathcal{P}(\mu)$ that agree with these marginals. The assumption is that the true distribution is in this set.[1] We want to predict y from $\mathbf{x}$ using some classification function.

Our goal is to find a classifier which has minimal worst case error. The classifier that solves this minimax problem will be robust in the sense that it obtains the best error possible

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

[1] Clearly there are finite sample issues, but these can be addressed as in Dudík et al. (2007), by considering uncertainty around the evaluated statistics.

under our uncertainty about the true distribution.

The above problem is generally hard to solve (Bertsimas and Sethuraman, 2000). The first difficulty is handling the zero-one loss. Here we use the usual approach of replacing it with a surrogate loss, which we choose to be the hinge loss. We show that this replacement results in a 2-approximation of the zero-one loss (see Section 4). However, the problem still seems daunting to optimize due to the maximization over all possible distributions in $\mathcal{P}(\mu)$. Surprisingly, we show that this problem is in fact tractable, as long as the set of pairwise variables in the marginals x_i, x_j correspond to a tree graph. When the graph is not a tree, we show how the commonly used MAP LP relaxation can be employed, resulting in an upper bound on the original minimax problem.

We call our approach Discrete Chebyshev Classifier ($\mathcal{DCC}$), since as in Chebyshev bounds, , it considers worst case behavior under first and second order moment constraints. Empirical comparisons to baselines that use the same statistical information demonstrate that $\mathcal{DCC}$ are competitive on the majority of datasets considered.

2. The DCC Optimization Problem

We begin by defining the minimax optimization problem we set out to solve.

Consider classification problems with n discrete features corresponding to the vector random variable $X = [X_1, \ldots, X_n]$. Assume that each X_i can take d_i values so that $X_i \in \{1, \ldots, d_i\}$. The set of possible values of X will be denoted by $\mathcal{X}$. Similarly, let the discrete variable Y denote the label of X, and denote the domain of Y by $\mathcal{Y}$.

Our focus is on predicting Y from X. Typically one considers a parametric form for such predictors. However, at this point we assume that it can be arbitrary. In Section 2.3 we show that the optimal minimax predictor does in fact have a certain parametric form. For now, we assume that the predictor is defined via a function $f(\mathbf{x}, y) : \mathcal{X}, \mathcal{Y} \to \mathbb{R}$ where the predicted label is given by:

$$\hat{y}(\mathbf{x}; f) = \arg\max_y f(\mathbf{x}, y). \tag{1}$$

Note that any prediction function can be expressed this way, and thus the learner has full expressive power. It will turn out in Section 2.3 that the optimal function has a simple parametric form, due to the fact that it needs to be minimax optimal.

For a given function f and a pair $\mathbf{x}, y$ the zero-one loss incurred by predicting y from $\mathbf{x}$ using f is:

$$\ell_{zo}(f, \mathbf{x}, y) = \mathcal{I}\left[y \neq \arg\max_{y'} f(\mathbf{x}, y)\right]. \tag{2}$$

Since this loss is not convex, we switch to a surrogate convex loss. Specifically, we use the multiclass hinge loss (e.g., see Crammer and Singer, 2002) defined as:

$$\ell_h(f, \mathbf{x}, y) = \max_{z \in \mathcal{Y}} f(\mathbf{x}, z) - f(\mathbf{x}, y) + \mathcal{I}\left[z \neq y\right]. \tag{3}$$

Given that $\mathbf{x}, y$ are generated via a distribution $p(\mathbf{x}, y)$, the expected loss of f is:

$$\mathbb{E}_p\left[l(f, \mathbf{x}, y)\right] = \sum_{\mathbf{x}, y} p(\mathbf{x}, y) l(f, \mathbf{x}, y). \tag{4}$$

As mentioned earlier, we consider the setting where we are given marginal distributions over pairs of variables X_i, X_j and the label variable Y. Each such distribution will be denoted by $\mu_{ij}(x_i, x_j, y)$ in what follows. Furthermore, we assume we have these for a set of pairs E which form a tree. The tree assumption might seem restrictive, and we will remove it in Section 5. Thus, our input is the set μ of marginals:

$$\mu = \{\, \mu_{ij}(x_i, x_j, y) : (i, j) \in E \,\}. \tag{5}$$

Define $\mathcal{P}(\mu)$ as the set of all probability distributions over (X, Y) that *agree* with the marginals μ. Namely:

$$\mathcal{P}(\mu) = \{p \in \Delta : p(x_i, x_j, y) = \mu_{ij}(x_i, x_j, y)\ \forall (i, j) \in E\} \tag{6}$$

where $p(x_i, x_j, y)$ is the marginal distribution of p on the corresponding variables, and Δ is the set of valid distributions on $X_1, \ldots, X_n$ (i.e., $p(x_1, \ldots, x_n)$ is non-negative and normalizes to one).

2.1. The Minimax Problem

We would like to find the optimal classifier f given that the true distribution is in $\mathcal{P}(\mu)$ (as noted earlier, this can be relaxed by using for instance, box constraints around μ). Since no additional information about the underlying distribution is provided, we consider the **w**orst **c**ase **e**rror of f with respect to *any* distribution in $\mathcal{P}(\mu)$, which is given by:

$$\text{WCE}(f, \mu) = \max_{p \in \mathcal{P}(\mu)} \mathbb{E}_p\left[\ell_{zo}(f, \mathbf{x}, y)\right]. \tag{7}$$

$\text{WCE}(f, \mu)$ is a Chebyshev type bound, since out of all probability distributions over (X, Y) with specific moments μ, it provides the one that maximizes the mass in a particular subset of $\mathcal{X}$ (i.e., the $\mathbf{x}$ where f errs).

It is then natural to seek an f that minimizes $\text{WCE}(f, \mu)$ specifically, to solve:

$$\mathcal{DCC}(\mu) = \min_f \text{WCE}(f, \mu) \tag{8}$$

where $\mathcal{DCC}$ stands for "Discrete Chebyshev Classifier". Due to the hardness of optimizing the zero-one loss we

consider a variant of the above where the zero-one loss in WCE is replaced with the hinge loss. Namely:

$$\text{WCE}_h(f,\mu) = \max_{p\in\mathcal{P}(\mu)} \sum_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}} p(\mathbf{x},y)\ell_h\left(f,\mathbf{x},y\right). \quad (9)$$

And similarly:

$$\mathcal{DCC}_h(\mu) = \min_f \text{WCE}_h(f,\mu). \quad (10)$$

This type of relaxation is common and easily yields an upper bound on the original function. Somewhat surprisingly, in Section 4 we show a tighter connection between the problems – $\mathcal{DCC}_h(\mu)$ is a 2-approximation of $\mathcal{DCC}(\mu)$.

The optimization problem $\mathcal{DCC}_h(\mu)$ still seems daunting due to two key difficulties:

- The function f is over $|\mathcal{X}||\mathcal{Y}|$ variables, which is exponential in n.
- The worst case error, WCE_h, involves maximization over all distributions $p(\mathbf{x},y)$. Again, these would require $|\mathcal{X}||\mathcal{Y}|$ variables to describe.

In what follows we show how these difficulties can be overcome via careful analysis of the optimization problem, use of convex duality, and MAP LP relaxations. The main result is Theorem 2.3 where we present a tractable convex optimization problem equivalent to $\mathcal{DCC}_h(\mu)$. Theorem 3.1 then provides an unconstrained formulation of the problem.

2.2. The Dual of WCE_h

Begin by rewriting WCE_h in its dual form. Since WCE_h is a linear program (LP) in variables $p(\mathbf{x},y)$, it has a dual LP with the same value. The dual variables in this case are $\nu_{ij}(x_i,x_j,y)$ for all $ij\in E$ and x_i,x_j,y. Thus, they can be viewed as local functions on pairs of features and y.

Using a standard Lagrangian duality transformation we obtain the following dual:

$$\begin{aligned}\text{WCE}_h(f,\mu) = \min_{\nu}\quad & \nu\cdot\mu \\ \text{s.t.}\quad & \nu(\mathbf{x},y) \geq \ell_h(f,\mathbf{x},y) \quad \forall \mathbf{x}\in\mathcal{X}, y\in\mathcal{Y}\end{aligned} \quad (11)$$

where we use the following notation:

$$\begin{aligned}\nu(\mathbf{x},y) &= \sum_{ij\in E} \nu_{ij}(x_i,x_j,y) \\ \nu\cdot\mu &= \sum_{ij\in E,x_i,x_j} \nu_{ij}(x_i,x_j,y)\mu_{ij}(x_i,x_j,y).\end{aligned}$$

The dual in Eq. (11) has a nice interpretation. The function $\nu(\mathbf{x},y)$ can be thought of as an energy function over $\mathbf{x}$ and y which decomposes according to the set of edges E, and constrained to be pointwise greater than the loss $l_h(f,\mathbf{x},y)$. Using this observation, and since for any distribution $p\in\mathcal{P}(\mu)$, $\mathbb{E}_p\left[\nu(\mathbf{x},y)\right] = \nu\cdot\mu$, it follows that $\nu\cdot\mu$ indeed upper bounds the expected hinge loss for any possible ν.

By switching to the dual we have not made the problem simpler, since now instead of exponentially many variables as in Eq. (7) we have exponentially many constraints. These however can be dealt with efficiently, as we show in Section 2.4. Another advantage is that $\mathcal{DCC}_h$ now becomes a minimization problem (rather than minmax):

$$\begin{aligned}\mathcal{DCC}_h(\mu) = \min_{f,\nu}\quad & \nu\cdot\mu \\ \text{s.t.}\quad & \nu(\mathbf{x},y) \geq \ell_h(f,\mathbf{x},y) \;\; \forall \mathbf{x}\in\mathcal{X}, y\in\mathcal{Y}.\end{aligned} \quad (12)$$

2.3. A Simple Form for the Optimal Classifier

Here we show that there exists an optimal f^* which can be described using much fewer variables. Furthermore, this f^* can be determined via the set of dual parameters ν defined earlier.

Theorem 2.1. *The $\mathcal{DCC}_h$ problem can be expressed as:*

$$\begin{aligned}\min_{\nu}\quad & \nu\cdot\mu \\ \textit{s.t.}\quad & \nu(\mathbf{x},z)+\nu(\mathbf{x},y)-2\cdot\mathcal{I}\left[y\neq z\right]\geq 0 \;\forall_{\mathbf{x},y,z}.\end{aligned} \quad (13)$$

Proof. We start with the $\mathcal{DCC}_h$ problem in Eq. (12). Since f does not appear in the objective (only in the constraints), it can be moved to the constraints in the following way:

$$\begin{aligned}\mathcal{DCC}_h(\mu) = \min_{\nu}\quad & \nu\cdot\mu \\ \text{s.t.}\quad \exists f\;\forall \mathbf{x},y:\quad & \nu(\mathbf{x},y)\geq\ell_h(f,\mathbf{x},y).\end{aligned} \quad (14)$$

If we write the constraints explicitly using the hinge loss definition we get that there should exist a f such that:

$$\forall \mathbf{x},y \quad \nu(\mathbf{x},y)\geq \max_z f(\mathbf{x},z)-f(\mathbf{x},y)+\mathcal{I}\left[y\neq z\right]$$

or equivalently:

$$\forall \mathbf{x},y,z \quad \nu(\mathbf{x},y)\geq f(\mathbf{x},z)-f(\mathbf{x},y)+\mathcal{I}\left[y\neq z\right].$$

Since each pair y,z appears twice with any given x we regroup the inequalities to get:

$$\begin{aligned}\nu(\mathbf{x},y)-\mathcal{I}\left[y\neq z\right] &\geq f(\mathbf{x},z)-f(\mathbf{x},y) \\ &\geq -\nu(\mathbf{x},z)+\mathcal{I}\left[y\neq z\right].\end{aligned} \quad (15)$$

We claim that for a given ν there exists a function f satisfying the above if and only if:

$$\nu(\mathbf{x},y)-\mathcal{I}\left[y\neq z\right]\geq -\nu(\mathbf{x},z)+\mathcal{I}\left[y\neq z\right] \;\; \forall \mathbf{x},y,z. \quad (16)$$

It is immediate that Eq. (15) implies Eq. (16). To see the converse, we define $f(\mathbf{x}, y) = -\frac{1}{2}\nu(\mathbf{x}, y)$ and therefore:

$$\begin{aligned} f(\mathbf{x}, z) - f(\mathbf{x}, y) &= \tfrac{1}{2}\left(\nu(\mathbf{x}, y) - \nu(\mathbf{x}, z)\right) \\ &= \tfrac{\nu(\mathbf{x},y) - \mathcal{I}[y \neq z] - \nu(\mathbf{x},z) + \mathcal{I}[y \neq z]}{2} \end{aligned} \tag{17}$$

which is exactly the mid point between the required lower and upper bounds on $f(\mathbf{x}, z) - f(\mathbf{x}, y)$ from Eq. (15).

This shows the equivalence between the constraints of Eq. (13) and Eq. (14), and hence the equivalence between the problems. □

The above theorem directly implies that the form of the optimal f^* is given by the following corollary.

Corollary 2.2. *Denote by ν^* the minimizer of Eq. (13), the $\mathcal{DCC}_h$ classifier f^* (from Eq. (10)) is given by:*

$$f^*(\mathbf{x}, y) = -\frac{1}{2}\nu(\mathbf{x}, y) = -\frac{1}{2}\sum_{ij \in E} \nu^*_{ij}(x_i, x_j, y). \tag{18}$$

2.4. Efficient DCC Optimization

In order to be able to solve the $\mathcal{DCC}_h$ problem as formulated in Eq. (13), one must still deal with the exponential number of constraints. Define:

$$h_\nu(\mathbf{x}, y, z) = \nu(\mathbf{x}, z) + \nu(\mathbf{x}, y) - 2 \cdot \mathcal{I}\left[y \neq z\right].$$

Then the constraint of Eq. (13) can be written as:

$$\min_{\mathbf{x}, y, z} h_\nu(\mathbf{x}, y, z) \geq 0. \tag{19}$$

The key property to note is that the function h decomposes as a sum over factors depending on (x_i, x_j, y, z) and (y, z). In other words, it can be viewed as an energy function of a graphical model with these hyper edges. The complexity of checking the constraints is thus equivalent to the complexity of calculating the MAP assignment of the corresponding model.

If the edges $ij \in E$ are arbitrary, the above problem is as hard as general MAP problems (NP hard). However, since we assumed that E is tree structured, the graphical model corresponding to h has tree width of 3 and can be minimized efficiently (e.g., using the junction tree algorithm. See Koller and Friedman, 2009).

At this point we have shown that for a given ν, the feasibility of the $\mathcal{DCC}_h$ constraints in Eq. (19) can be checked efficiently. Thus, the $\mathcal{DCC}_h$ problem is in fact polynomial time tractable since one can use methods such as ellipsoid or cutting plane (Bertsekas, 1995). While we could theoretically solve the problem this way, these methods do not scale well. We thus turn to further simplify the problem.

A different way to solve Eq. (19) is to realize that it can be expressed as a linear program. There is a rich body of work on LP relaxations for the MAP problem, and their various relaxations (e.g., see Wainwright and Jordan, 2003; Werner, 1993; Globerson and Jaakkola, 2008; Sontag et al., 2011). In our case, such an LP would have as variables the following fourth and second order distributions $\alpha_{ij}(x_i, x_j, y, z)$ and $\tau(y, z)$, and the constraints would be that these distributions agree on the variables in their overlap. It is easy to see that constructing this LP in the standard way will result in the exact MAP since these are the cliques in the junction tree of the model.

We can now take the dual of the above MAP LP. The dual variables in this case will be denoted by $\delta_{ij}(x_j)$ and $\gamma_{ij}(y, z)$. The dual objective $g_\nu(\delta, \gamma)$ is given by:

$$\begin{aligned} g_\nu(\delta, \gamma) = &\max_{\delta, \gamma} \sum_{ij} \min_{\substack{x_i, x_j \\ y, z}} \{\nu_{ij}(x_i, x_j, y) + \nu_{ij}(x_i, x_j, z) \\ &\quad - \delta_{ij}(x_j, y, z) - \delta_{ji}(x_i, y, z) - \gamma_{ij}(y, z)\} \\ &+ \sum_i \min_{x_i, y, z} \left\{ \sum_{j \in N(i)} \delta_{ji}(x_i, y, z) \right\} \\ &+ \min_{y, z} \left\{ -2 \cdot \mathcal{I}\left[y \neq z\right] + \sum_{ij \in E} \gamma_{ij}(y, z) \right\} \end{aligned} \tag{20}$$

From strong duality we then have:

$$\min_{\mathbf{x}, y, z} h_\nu(\mathbf{x}, y, z) = \max_{\delta, \gamma} g_\nu(\delta, \gamma). \tag{21}$$

Finally, by plugging the dual into Eq. (13) we have that $\mathcal{DCC}$ can be expressed as an optimization problem with polynomially many constraints and variables, as stated below.

Theorem 2.3. *The $\mathcal{DCC}_h$ problem is equivalent to:*

$$\begin{array}{ll} \min\limits_{\nu, \delta, \gamma} & \nu \cdot \mu \\ \textit{s.t.} & g_\nu(\delta, \gamma) \geq 0. \end{array} \tag{22}$$

Proof. After substituting Eq. (21) into Eq. (13) we get:

$$\begin{array}{ll} \min_\nu & \nu \cdot \mu \\ \text{s.t.} & \max_{\delta, \gamma} g_\nu(\delta, \gamma) \geq 0. \end{array} \tag{23}$$

Now we only need to notice that δ, γ can be maximized over outside the constraints, since it suffices to find a single assignment to those such that $g_\nu(\delta, \gamma) \geq 0$ for the constraint to hold. The result follows. □

The above problem has polynomially many constraints and variables, and is convex. Thus it can be solved using

generic convex optimization tools. However, solving this problem with off-the-shelf optimization methods does not scale well. Next, we derive an unconstrained version of the problem which can be solved by scalable accelerated methods.

3. Unconstrained Optimization of $\mathcal{DCC}_h$

Here we provide another equivalent form of $\mathcal{DCC}_h$ but one which is unconstrained, and can thus be solved using scalable accelerated methods, such as FISTA (Beck and Teboulle, 2009). The unconstrained problem is presented in the following result.

Theorem 3.1. *The following unconstrained problem is equivalent to $\mathcal{DCC}_h$:*

$$\min_{\nu,\delta,\gamma} \nu \cdot \mu - \tfrac{1}{2} \max_{\delta,\gamma} g_\nu(\delta,\gamma). \tag{24}$$

Proof. We start by writing the Lagrangian of $\mathcal{DCC}_h$ in Eq. (13) :

$$L(\lambda,\nu) = \nu \cdot \mu - \lambda \left(\min_{\mathbf{x}\in\mathcal{X}_z, y\in\mathcal{Y}} h_\nu(\mathbf{x},y,z) \right). \tag{25}$$

Denote by ν^ω the uniform assignment to the ν variables (i.e., all $\nu_{ij}(x_i,x_j,y)=\omega$). Then some algebra gives

$$L(\lambda,\nu^\omega) = \omega|E|(1-2\lambda) + 2\lambda \tag{26}$$

therefore we have:

$$\mathcal{DCC}(\mu) \le \max_\lambda \min_\omega \omega|E|(1-2\lambda) + 2\lambda \tag{27}$$

which is clearly equal to $-\infty$ for $\lambda \neq \frac{1}{2}$. We conclude that the optimal λ value is $\frac{1}{2}$ and the result follows from the fact we can replace h with g. □

To run FISTA (Beck and Teboulle, 2009) on Eq. (24) we can smooth it using Nesterov's smoothing approach (Nesterov, 2005) and then apply FISTA since the gradients are easy to compute.

4. A 2-Approximation of $\mathcal{DCC}$

The hinge loss is commonly used as a convex upper bound surrogate for the zero-one loss. However, in the general case no further approximation guarantees can be provided. For the $\mathcal{DCC}$ problem it in fact turns out that replacing zero-one loss with hinge loss results in a factor 2 approximation, as stated next.

Lemma 4.1. $\frac{1}{2}\mathcal{DCC}_h(\mu) \le \mathcal{DCC}(\mu) \le \mathcal{DCC}_h(\mu)$.

Proof. The upper bound follows from the fact that hinge loss upper bounds the zero-one loss. To show the lower bound, begin by applying the Sion Minimax Theorem (Sion, 1957) to the definition of $\mathcal{DCC}_h$(using the convexity of the hinge loss wrt f, and linearity wrt to p):

$$\min_f \max_{p\in\mathcal{P}(\mu)} \mathbb{E}_p\left[\ell_h(f,\mathbf{x},y)\right] = \max_{p\in\mathcal{P}(\mu)} \min_f \mathbb{E}_p\left[\ell_h(f,\mathbf{x},y)\right]. \tag{28}$$

For any p, it holds that:[2]

$$\min_f \mathbb{E}_p\left[\ell_h(f,\mathbf{x},y)\right] \le \min_{f:\ f(\mathbf{x},y)\in\{0,1\}} \mathbb{E}_p\left[\ell_h(f,\mathbf{x},y)\right]. \tag{29}$$

For predictor functions f with outputs in $\{0,1\}$ the hinge loss is always exactly twice the error probability. Therefore, as claimed:

$$\mathcal{DCC}_h(\mu) \le 2 \max_{p\in\mathcal{P}(\mu)} \min_f \mathbb{E}_p\left[\ell_{zo}(f,\mathbf{x},y)\right] \le 2\mathcal{DCC}(\mu).$$

Where the second inequality follows from weak min-max bounds. □

5. Relaxing the Tree Assumption

Thus far we assumed that the set of observed marginals correspond to a tree graph. This resulted in a tractable form for the $\mathcal{DCC}_h$ problem. However, the tree assumption may be too restrictive in many cases, both because there is often no natural tree structure, and because learning the optimal tree (as in the Chow Liu procedure used in TAN) seems hard for our objective. Finally, the tree assumption limits the number of statistics we can consider to $n-1$.

Fortunately, the MAP LP relaxation approach used for handling the constraint Eq. (19) in Section 2.4 can easily be generalized to non-tree graphs. In fact, it is one of the most common approximation methods for MAP inference in general graphs (e.g., see Komodakis et al., 2011). Thus, if E is not a tree graph, we basically employ exactly the same procedure as described above. The only difference from the tree case is that Eq. (22) will no longer be equal to the original $\mathcal{DCC}_h$ but instead it will upper bound it.[3] In our experiments we tried both the tree approach and the general graph approach. See Section 8 for further details.

6. Other Applications

The $\mathcal{DCC}$ framework can be applied to other machine learning settings. In what follows we consider four different scenarios of interest.

Conditional $\mathcal{DCC}$: In some settings it is of interest to consider errors conditioned on Y, as in type I and II errors in hypothesis testing. A minimax approach in this

[2] Note that this is in fact an equality.

[3] This is a direct result of the MAP LP relaxation upper bounding the true MAP value.

setting was described in Lanckriet et al. (2003) for second order moments on continuous variables. $\mathcal{DCC}$ can easily be extended to this setting, resulting in similar optimization problems to those derived here.

Semi-Supervised $\mathcal{DCC}$: In many real world applications it is easy to collect a large volume of unlabeled data, but labeled data is harder to obtain. In this context $\mathcal{DCC}$ can naturally be extended by assuming statistics only on the variables $\mathbf{x}$. For example, assume we only have enough labeled data to estimate $\mu_i(x_i, y)$ (but not enough for $\mu_{ij}(x_i, x_j, y)$ marginals). However, we may easily estimate marginals such as $\mu_{ij}(x_i, x_j)$ from unlabeled data. Using our framework it is straightforward to formulate a WCE problem for this set of marginals. It is expected to considerably restrict the set of distributions which we maximize over, and thus make the minimax setting less conservative.

Discrete Chebyshev Bounds: The Chebyshev bound, or Chebyshev inequality is a simple yet extremely useful theorem in probability: For any real random variable, X, with expectation μ and finite non-zero variance σ^2, and for any $\epsilon > 0$.

$$\mathbb{P}\left[(|X - \mu| \geq \epsilon)\right] \leq \frac{\sigma^2}{\epsilon^2}.$$

Many generalizations of these inequalities are known (e.g., Marshall and Olkin, 1960; Bertsimas and Popescu, 2005; Vandenberghe et al., 2007). However, they all deal with continuous distributions.

We define the Discrete Chebyshev inequality in a similar way: Given a set of marginals over $\mathcal{X}$, and a function $f : \mathcal{X} \to \mathbb{R}$ we wish to bound the following probability:

$$\max_{p \in \mathcal{P}(\mu)} \mathbb{P}_p\left[f(\mathbf{x}) \geq 0\right]. \tag{30}$$

Using the same line of reasoning that we used to simplify WCE we can bound the above probability by

$$\min_{\nu} \nu \cdot \mu \text{ s.t. } \sum_{ij} \nu_{ij}(x_i, x_j) \geq \max\{0, 1 + f(\mathbf{x})\} \tag{31}$$

which can be computed if f decomposes.

WCE as a Regularizer: Since $\text{WCE}_h(f, \mu)$ is a bound on generalization performance of the predictor f, it makes sense to use it as a regularizer. Given a dataset with several fully observed examples as well as only statistics μ for the entire data, we may optimize an objective that is a sum of a hinge loss on the labeled examples and $\text{WCE}_h(f, \mu)$. This will combine two different estimates of generalization error, thus effectively leveraging the two data sources.

7. Related Work

The idea of using expected values is common in generative models for prediction, but has seen much fewer applications in discriminative approaches. One exception is the Minimax Probabilistic Machine (*MPM*) (Lanckriet et al., 2003) which solves a robust classification problem for the case of first and second order moments. The derivation in (Lanckriet et al., 2003) relies on Chebyshev type bounds that upper bound the probabilities of certain events subject to constraints (e.g., see Vandenberghe et al., 2007). However, these apply to continuous spaces, as opposed to the discrete case we consider here. Indeed, a key contribution of the current paper is to consider such bounds and their approximation for discrete graphical models.

Information theoretic measures such as entropy and mutual information have also been used in the context of learning with partial information. As mentioned earlier, maximum entropy is a classic approach to the problem. Interestingly, it may also be interpreted as minimax optimal but under a different (non-discriminative) loss function (Grünwald and Dawid, 2004). Another related approach is the minimum mutual information (MinMI) method (Globerson and Tishby, 2004). It minimizes $I(X; Y)$ under marginal constraints as we have here, but is hard to compute even for the case of singleton marginals.

The notion of statistical queries (Kearns, 1998) is also related to our setting. However, in these works the queries used by the learner are chosen from a much larger family than what we use. This is also true for the more restricted correlational queries (Bshouty and Feldman, 2002) where one receives expected values $\mathbb{E}_p\left[f(X_1, \ldots, X_n)Y\right]$ and f can be any function. This setting is thus still considerably less constrained than ours.

There are also minimax approaches which do not consider expected values, but rather seek minimax robustness with respect to perturbations of data points. For example, one may want to minimize prediction error subject to a data point being allowed to move within some prescribed radius. Such settings have been studied for different perturbations, often resulting in SVM like optimization methods (e.g., see Xu et al., 2009; Livni et al., 2012). Note that these approaches however require the full dataset and thus do not operate in our restricted input setting.

Finally, robust optimization approaches are quite common in both optimization (Ben-Tal and Nemirovski, 1998) and statistics (Berger, 1985) but in a context different from what is considered here.

8. Experiments

Here we evaluate our method and other baselines on both synthetic and real world datasets.

8.1. Toy Problem

We first provide a scenario where generative methods fail and $\mathcal{DCC}$ succeeds. We considered two generative (maximum entropy based) baselines: Naive Bayes (*NB*) and Tree Augmented Naive Bayes (*TAN*) (Friedman et al., 1997). The data was generated such that it will violate the conditional independence assumptions of both the *NB* and *TAN* approaches. We used a distribution over a label y and $n = 2k + 11$ binary variables. The distribution corresponds to the Bayesian network shown in Figure 1a, with parameters defined as follows:

- $p(Y = 1) = \frac{1}{2}$
- $p(S = Y) = 0.9$, so that S, Y are strongly correlated.
- $p(W_i = Y) = 0.6$, so that W_i, Y are weakly correlated.
- $C_i = W_1$ i.e. the C_i's are identical to W_1
- $D_i = W_{j_{i1}} \oplus W_{j_{i2}}$ i.e. D_i is determined by its parents. Each D_i has a randomly chosen pair of parents.

The features x are all the $2k+11$ non-label variables. It can be seen that they do not satisfy the conditional assumption of either *NB* or *TAN*. Each synthetic trial contained 5,000 examples divided equally between train and test sets. The results reported are the average over 10 random generations of the data. Figure 1b shows the average classification error of the different algorithms for different numbers of variables. As the number of variables increases, the advantage of $\mathcal{DCC}$ over *NB* and *TAN* is apparent. This experiment illustrates the poor performance of models with implicit assumptions when the assumptions do not hold.

8.2. Comparing Marginal Based Learners

We tested the $\mathcal{DCC}$ classifier scheme on 12 classification datasets from the UCI repository, nine of them are binary and the other three are multiclass classification tasks. In several datasets there are continuous features, in these cases we used only the discrete features (this follows the setup and datasets used in Globerson and Tishby, 2004). The input to the algorithms was marginals of the form $\mu_{ij}(x_i, x_j, y)$ for all possible pairs of features; the marginals were computed from a train set, and the performance of the resulting classifier was evaluated on a test set. The error rates reported in Table 1 are the average of 5 partitions into train and test sets.

The error rate of $\mathcal{DCC}$ was compared to three baseline algorithms which take marginal distributions as inputs: *NB*, *TAN*, and Minimax Probabilistic Machine (*MPM*) (Lanckriet et al., 2003). *MPM* minimizes an objective function similar to $\mathcal{DCC}$, with the crucial difference that

Dataset	$\mathcal{DCC}$	g$\mathcal{DCC}$	MPM	TAN
adult	**18**	**18**	22	**18**
bcd	20	**18**	26	32
credit	14	**13**	**13**	17
heart-disease	21	20	**18**	23
hepatitis	19	**15**	18	17
hypo	**8**	**8**	N/A	**8**
kr-vs-kp	10	10	**5**	7
lymphography	16	**8**	N/A	18
mushroom	**0**	**0**	**0**	**0**
promoters	5	**3**	6	44
sick	**6**	**6**	23	**6**
votes	**3**	**3**	4	8
Best Performance	5	10	4*	4

Table 1. Performance (in % error) on test data for real-world discrete classification datasets. **MPM* error is missing for the multiclass problems (lymphography, hypo) as *MPM* is a binary classification algorithm.

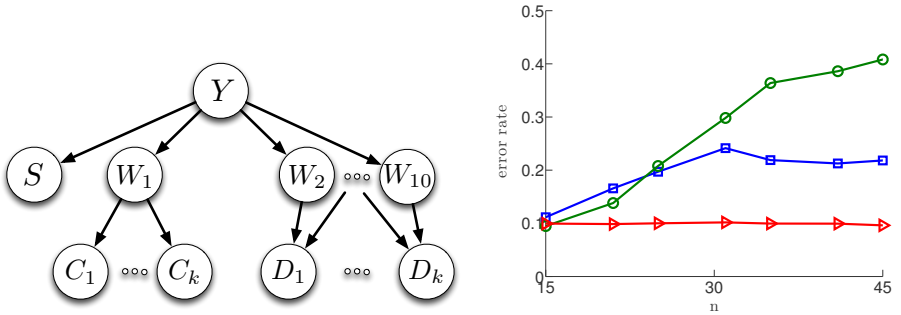
MPM assumes the underlying distribution is continuous. In addition to the vanilla version of $\mathcal{DCC}$ we also examined a greedy algorithm which we denote $g\mathcal{DCC}$. This greedy version starts with an empty set of edges E and at each step adds the edge which minimizes the $\mathcal{DCC}$ error bound until a spanning tree structure is reached.

In Table 1 the error rates of the compared algorithms are presented. The best performing algorithm on each dataset is shown in boldface. As can be seen, our approaches (either $\mathcal{DCC}$ or $g\mathcal{DCC}$) outperform the competing algorithms on 10 out of 12 datasets. To conclude, our approach achieves better or comparable performance on most datasets examined.

8.3. Learning with Missing Features

One of the motivations of $\mathcal{DCC}$ is to learn in scenarios where not all features are available, but pairwise statistics can be estimated. Here we simulate this setting by repeating the experiments in the previous section while "hiding" 25% of the features of each train example. The following baselines are used for comparison:

- Replacing each missing feature with its mean in the training data (i.e., imputation) and running SVM on the resulting full observations.
- Chechik et al. (2008) propose a variant of SVM for learning with missing features, by an appropriate rescaling of the margin. Their algorithms have two versions: avg-w and geom, both of which are evaluated here.
- Running $\mathcal{DCC}$ on the set of single and pairwise statistics collected from the data. Since these may



(a) (b)

Figure 1. Evaluation over synthetic data. (a) The Bayesian network from which the data was drawn. (b) Error rate of competing algorithms as a function of the number of variables. *NB*: green circles, *TAN*: blue squares, $\mathcal{DCC}$: red triangles.

Dataset	$\mathcal{DCC}$	avg-w	geom	SVM
bcd	**27**	29	30	31
credit	**15**	16	19	19
heart-disease	27	25	**20**	**20**
hepatitis	**19**	23	29	22
kr-vs-kp	**10**	13	13	13
lymphography	22	21	26	**18**
promoters	**12**	13	13	**12**
votes	**4**	9	7	7
Best Performance	6	0	1	3

Table 2. Performance (in % error) on noisy data, 25% of training features where erased. Only datasets with noticeable difference between algorithms are presented. avg-w, geom are the algorithms from Chechik et al. (2008), to run SVM we filled the missing values with the mean value of the feature.

not be consistent (e.g., $\mu(x_1|y)$ might be different when marginalizing $\mu(x_1, x_2|y)$ and $\mu(x_1, x_3|y)$) we project those on the pairwise consistency constraints (also known as the local polytope). We use ℓ_1 as the projection metric.[4]

Results are shown in Table 2. It can be seen that $\mathcal{DCC}$ outperforms the other methods on the majority of the datasets.

9. Discussion

The $\mathcal{DCC}$ approach is motivated by learning settings where complete observations are unavailable. Instead, access to certain statistics of the data is provided. A minimax approach is very natural in this setting, and yields the $\mathcal{DCC}$ objective. A well known limitation of minimax methods is that they assume a worst case adversary and may thus learn classifiers that are suboptimal for the true distributions that generated the data. To alleviate this shortcoming, one must impose additional constraints on the adversaries. The $\mathcal{DCC}$ approach takes a large step in this direction by limiting the support of the adversarial distribution to only integral assignments. This is in contrast to methods like MPM where the adversarial distribution has unconstrained support on the reals. Indeed, our experiments show that $\mathcal{DCC}$ outperforms MPM in the majority of the cases, presumably because of our less pessimistic approach.

Here we considered a deterministic classifier which always returns the same y for a given $\mathbf{x}$. However, in the minimax setting the optimal strategy is actually stochastic. It would thus be interesting to study the stochastic variant of the $\mathcal{DCC}$ problem. Another advantage of the stochastic case is that the zero one loss is linear in the classifier distribution, possibly leading to efficient algorithms.

The $\mathcal{DCC}$ approach has several natural and interesting extensions. For example, we alluded to the possibility of using it in a semi supervised manner. Another exciting extension is to the structured output prediction setting. Consider for example a part of speech tagging problem, and assume we have access to the statistics of consecutive parts of speech, and those of words and their part of speech. How can these be combined to build discriminative minimax structured output predictors?

Finally, we would like to consider other forms of robust prediction losses. As an example, consider minimizing the *regret* of the classifier (e.g., see Eldar et al., 2004) rather than its worst case loss. This will effectively reduce the strength of the minimax adversary and may result in improved performance. However, it remains to be seen whether it can be solved efficiently as in the minimax case considered here.

Acknowledgments: We thank Shai Shalev-Shwartz, Roi Livni and Yoav Wald for fruitful comments and discussions. This research is funded by the ISF Centers of Excellence grant 1789/11. Elad Eban was partially funded by an IBM PhD fellowship.

[4] See Ravikumar et al. (2010) for other projection schemes.

References

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

A. Ben-Tal and A. Nemirovski. Robust convex optimization. *Math. Oper. Res.*, 23(4):769–805, 1998.

J. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer New York, 1985.

D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 1995.

D. Bertsimas and I. Popescu. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization*, 15(3):780–804, 2005.

D. Bertsimas and J. Sethuraman. Moment problems and semidefinite optimization. In *Handbook of semidefinite programming*, pages 469–509. Springer, 2000.

N. H. Bshouty and V. Feldman. On using extended statistical queries to avoid membership queries. *JMLR*, 2:359–395, 2002.

G. Chechik, G. Heitz, G. Elidan, P. Abbeel, and D. Koller. Max-margin classification of data with absent features. *JMLR*, 9:1–21, 2008.

K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *JMLR*, 2:265–292, 2002.

M. Dudík, S. J. Phillips, and R. E. Schapire. Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *JMLR*, 8:1217–1260, 2007.

Y. C. Eldar, A. Ben-Tal, and A. Nemirovski. Linear minimax regret estimation of deterministic parameters with bounded data uncertainties. *Signal Processing, IEEE Trans. on*, 52(8):2177–2188, 2004.

N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine learning*, 29(2-3):131–163, 1997.

A. Globerson and T. Jaakkola. Fixing max-product: Convergent message passing algorithms for MAP LP-relaxations. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 553–560. MIT Press, Cambridge, MA, 2008.

A. Globerson and N. Tishby. The minimum information principle in discriminative learning. In *Proc. of the 20th conference on Uncertainty in artificial intelligence*, pages 193–200. 2004.

P. D. Grünwald and A. P. Dawid. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *Annals of Statistics*, pages 1367–1433, 2004.

M. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6):983–1006, 1998.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

N. Komodakis, N. Paragios, and G. Tziritas. MRF energy minimization and beyond via dual decomposition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(3):531–552, 2011.

G. R. Lanckriet, L. E. Ghaoui, C. Bhattacharyya, and M. I. Jordan. A robust minimax approach to classification. *JMLR*, 3:555–582, 2003.

R. Livni, K. Crammer, and A. Globerson. A simple geometric interpretation of svm using stochastic adversaries. In *Proc. of the 15th International Conference on Artificial Intelligence and Statistics*, 2012.

A. W. Marshall and I. Olkin. Multivariate Chebyshev inequalities. *The Annals of Mathematical Statistics*, 31 (4):1001–1014, 1960.

Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Prog.*, 103(1):127–152, 2005.

P. Ravikumar, A. Agarwal, and M. J. Wainwright. Message-passing for graph-structured linear programs: Proximal methods and rounding schemes. *JMLR*, 11: 1043–1080, 2010.

M. Sion. *General Minimax Theorems*. United States Air Force, Office of Scientific Research, 1957.

D. Sontag, A. Globerson, and T. Jaakkola. Introduction to dual decomposition for inference. In S. Sra, S. Nowozin, , and S. J. Wright, editors, *Optimization for Machine Learning*, pages 219–254. MIT Press, 2011.

L. Vandenberghe, S. Boyd, and K. Comanor. Generalized chebyshev bounds via semidefinite programming. *SIAM Rev.*, 49(1):52–64, 2007.

M. Wainwright and M. Jordan. Graphical models, exponential families and variational inference. Technical report, UC Berkeley, Dept. of Statistics, 2003.

T. Werner. A linear programming approach to max-sum problem: A review. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29:1165–1179, 1993.

H. Xu, C. Caramanis, and S. Mannor. Robustness and regularization of support vector machines. *JMLR*, 10: 1485–1510, 2009.