# Max-margin structured output learning in $L_1$ norm space

**Sandor Szedmak**[*]
ISIS Group, Electronics and Computer Science
University of Southampton
Southampton, United Kingdom
ss03v@ecs.soton.ac.uk

**Craig J. Saunders**
ISIS Group, Electronics and Computer Science
University of Southampton
Southampton, United Kingdom
cjs@ecs.soton.ac.uk

**Yizhao Ni**
ISIS Group, Electronics and Computer Science
University of Southampton
Southampton, United Kingdom
yn05r@ecs.soton.ac.uk

**Juho Rousu**
Department of Computer Science
University of Helsinki, Finland
juho.rousu@cs.helsinki.fi

## Abstract

We study a structured output learning setting where both the sample size and dimensions of the feature vectors of both the input and output are very large (possibly infinite in the latter case), but the input and output feature representations are nonnegative and very sparse (i.e. the number of nonzero components is finite and their proportion to the dimension is close to zero). Such situations are encountered in real-world problems such as statistical machine translation.

We show that in this setting structured output learning can be efficiently implemented. The solution relies on maximum margin learning of the linear relations between the inputs and outputs in an $L_1$ norm space. This learning problem can be formulated by imposing $L_\infty$ norm regularisation on the linear transformation expressing the relations.

## 1 Introduction

The machine learning researchers have devoted relatively small effort in discovering how the margin based learning methods behave in $L_1$ space. Several papers investigate the case when the $L_1$ norm is applied in the regularisation - e.g. Linear Programming Boosting [DBST01], or under the name “lasso” [Tib96], - or(and) in measuring the loss, e.g the SVM with soft margin, see for example in [CST00]. This paper focuses on the applications where instead of the $L_1$ regularisation, the margin is measured in an $L_1$ sense. It will be shown that this kind of learning displays characteristic properties allowing very large scale problems to be solved with moderate computational effort.

Like in machine learning, also in approximation theory surprisingly little attention to the $L_1$ norm space has been given. A book [Pin89] summaries some results to be good starting points for further research activities. Another valuable source could be [DG85] which deals with density estimation in $L_1$ space. The authors of the latter book emphasise an important fact that measuring the distance between two density functions using the $L_1$ norm is invariant under monotone transformation of the coordinate axes: in other words if only the order of the coordinates are preserved but the scales are changing then the $L_1$ normed base distances of vectors normalised to 1 in the same norm remains the same. This allows us to use a nonparametric approach when the underlying distributions might be irregular, e.g. they have no expected value. Human language is an example application field showing the symptoms of irregularity which has motivated us in formulating the presented approach.

In the following we first formulate the supervised learning problem for structured outputs, then secondly present the optimisation framework. We then show that the base problem can be solved in a very simple way which leads us to an online algorithm. Using a perceptron type interpretation we are able to state Novikoff-style bounds for the new algorithm.

## 2 General Setting

We are given a sample of pairs of input and output objects $\{x_i, y_i\}$, $i = 1, \ldots, m$, taken from the sets $\mathcal{X}$ and $\mathcal{Y}$ independently with respect to an unknown distribution defined on $\mathcal{X} \times \mathcal{Y}$.

Furthermore, there exist two functions $\phi$ and $\psi$ which map the input and output objects into linear vector spaces, namely

$$\phi : \mathcal{X} \Rightarrow \mathcal{L}_\phi,$$
$$\psi : \mathcal{Y} \Rightarrow \mathcal{L}_\psi,$$

where $\mathcal{L}_\phi$ and $\mathcal{L}_\phi$ are linear vector spaces whose elements represent the input and output objects.

The task is to find a linear transformation $\mathbf{W}$ which gives a good predictor of the outputs represented in the corresponding linear vector space by the feature vector of the inputs

$$\boldsymbol{\psi}(y) \Leftarrow \mathbf{W}\boldsymbol{\phi}(x)$$

One concrete example of this style of problem is a machine translation task where both the input and output objects

[*] All authors would like to acknowledge the support of the EU framework six project SMART

are sentences taken from natural languages, e.g. English and French, and are represented occurrences of phrases: ngrams, substrings of words with a special structure, etc. Thus, the input and output feature vectors $\boldsymbol{\phi}(x_i)$ and $\boldsymbol{\psi}(y_i)$ have very high dimensions but they are very sparse. Many other applications studied in the structured prediction literature [TGK03, TLJJ06, TJHA05, RSSST06] however also fit naturally into this framework.

## 3 Optimisation problem

We are going to express the relations between the input and the output via a linear transformation projecting the input feature vectors into the space of the output feature vectors which could be an optimum solution of the following maximum margin problem:

$$\begin{array}{ll} \min & r(\mathbf{W}) \\ \text{w.r.t} & \mathbf{W} : \mathcal{L}_\phi \Rightarrow \mathcal{L}_\psi, \text{ Linear operator} \\ \text{s.t.} & \langle \boldsymbol{\psi}(y_i), \mathbf{W}\boldsymbol{\phi}(x_i)\rangle_{\mathcal{L}_\psi} \geq b, i = 1, \dots, m, \\ & b > 0 \text{ given constant.} \end{array}$$

The objective function $r()$ is assumed to be a regularisation function and its concrete definition is derived later. The constraints force the inner products between the output feature vectors and the images of the input feature vectors with respect to the linear operator $\mathbf{W}$ to be sufficiently and uniformly large.

We use the inner product in a rather algebraic sense $\langle \mathbf{u}, \mathbf{v}\rangle = \sum_{k=1}^{n} u_k v_k,\ u, v \in \mathbb{R}^n$ instead of the geometric one which assumes a Hilbert space in the background.

The constraints can be rewritten expressing, instead of a regression task, a one-class classification task in the joint feature space for inputs and outputs, namely

$$\begin{array}{l} \big\langle \boldsymbol{\psi}(\mathbf{y}_i), \mathbf{W}\boldsymbol{\phi}(\mathbf{x}_i)\big\rangle_{\mathcal{L}_\psi} \\ = \mathbf{tr}\big(\boldsymbol{\psi}(\mathbf{y}_i)^T \mathbf{W}\boldsymbol{\phi}(\mathbf{x}_i)\big) \\ = \mathbf{tr}\big(\mathbf{W}\boldsymbol{\phi}(\mathbf{x}_i)\boldsymbol{\psi}(\mathbf{y}_i)^T\big) \\ = \Big\langle \mathbf{W}, \big[\boldsymbol{\psi}(\mathbf{y}_i) \otimes \boldsymbol{\phi}(\mathbf{x}_i)\big] \Big\rangle_{\mathcal{L}_\psi \otimes \mathcal{L}_\phi}, \end{array} \tag{1}$$

where $\mathbf{tr}()$ denotes the trace of the matrix in the argument, and operator $\otimes$ marks the tensor product between its operands.

Since the tensor product of two linear vector spaces is a linear vector space too, so we can interpret $\mathbf{z}_i = \boldsymbol{\phi}(x_i) \otimes \boldsymbol{\psi}(y_i)$ as vectors and the linear operator $\mathbf{W}$ becomes a linear functional, thus a vector in the dual space of the space spanned by the vectors $\{\mathbf{z}_i\},\ i = 1, \dots, m$, hence we can use vector notation $\mathbf{w}$ as well.

Based on (1) we arrive at a problem coinciding in form with the standard one-class SVM classification problem:

$$\begin{array}{ll} \min & r(\mathbf{w}) \\ \text{w.r.t} & \mathbf{w} \\ \text{s.t.} & \mathbf{z_i}'\mathbf{w} \geq b, i = 1, \dots, m, \\ & b > 0 \text{ given constant,} \\ & \mathbf{z}_i \in \mathbb{R}^{n_z}. \end{array} \tag{2}$$

The well-known approaches to the regularisation apply $L_2$ norm [Vap98], e.g. Support Vector Machine, or $L_1$ norm, e.g. Linear Programming Boosting [DBST01]. The use of 'structured outputs' or feature spaces in the output space has recently been studied using the standard $L_2$ norm based regularisation, see e.g. [SSTPH05] and [AHP$^+$08]. However investigating the case when the maximum margin is measured in $L_1$ norm have rarely been investigated.

We are going to focus on the regularisation function $r()$ which maximises the $L_1$ norm based distance between the separating hyperplane and the origin in the one-class problem. To this end a subproblem can be formulated computing this distance measured between the origin and the closest point of the hyperplane for which we have

$$\begin{array}{ll} \min & \|\mathbf{u}\|_1 \\ \text{w.r.t} & \mathbf{u} \\ \text{s.t.} & \mathbf{w}'\mathbf{u} = b, \end{array} \tag{3}$$

saying that a vector $\mathbf{u}$ sitting on the hyperplane is looked for with the minimum $L_1$ norm.

The entire problem which maximises the minimum distance takes the following form

$$\begin{array}{ll} \max & f(\mathbf{w}, b) = \left[ \begin{array}{ll} \min & \|\mathbf{u}\|_1 \\ \text{w.r.t} & \mathbf{u} \\ \text{s.t.} & \mathbf{w}'\mathbf{u} = b \end{array} \right] \\ \text{w.r.t} & \mathbf{w} \\ \text{s.t.} & \mathbf{z_i}'\mathbf{w} \geq b, i = 1, \dots, m \\ & b > 0 \text{ given constant.} \end{array} \tag{4}$$

In the sequel we are going to deal with the subcase of (4) when the non-negativity conditions $\mathbf{z}_i \geq \mathbf{0},\ i = 1, \dots, m$ hold.

## 4 Optimum solution

Let us solve first the subproblem given by (3) for a fixed $\mathbf{w}$. Via a simple argument one can show that the optimum value of (3) can be given in a closed form.

**Proposition 1** *The optimum value to the problem (3) is equal to*

$$\|\mathbf{u}^*\|_1 = \frac{b}{\max_j |w_j|} = \frac{b}{\|\mathbf{w}\|_\infty}. \tag{5}$$

**Proof:** First for sake of simplicity, we divide both sides of the equality constraint by $b$, due to it being strictly positive no effect on the problem is caused. Let us denote $\dfrac{\mathbf{w}}{b}$ with $\mathbf{w}_b$. We can assume at least one component of $\mathbf{w}_b$ differs from 0 otherwise no feasible solution exists.

Now we unfold the norm in the objective by applying the substitution $\mathbf{u} = \mathbf{u}_+ - \mathbf{u}_-,\ \mathbf{u}_+ \geq \mathbf{0},\ \mathbf{u}_- \geq \mathbf{0}$ and write up the dual problem as well.

$$\begin{array}{llll} \min & \mathbf{1}'\big(\mathbf{u}_+ + \mathbf{u}_-\big) & \max & \gamma \\ \text{w.r.t} & \mathbf{u}_+, \mathbf{u}_- & \text{w.r.t} & \gamma \\ \text{s.t.} & \mathbf{w}_b'\big(\mathbf{u}_+ - \mathbf{u}_-\big) = 1, & \text{s.t.} & \gamma\mathbf{w}_b \leq \mathbf{1}, \\ & \mathbf{u}_+ \geq \mathbf{0},\ \mathbf{u}_- \geq \mathbf{0}, & & -\gamma\mathbf{w}_b \leq \mathbf{1}. \end{array} \tag{6}$$

For any strictly positive components of $\mathbf{w}_b$ we have

$$\gamma\mathbf{w}_b \leq \mathbf{1} \Rightarrow \gamma \leq \min_j \frac{1}{(w_b)_j},\ (w_b)_j > 0, \tag{7}$$

and for any strictly negative components of the same vector the following holds

$$-\gamma\mathbf{w}_b \leq \mathbf{1} \Rightarrow \gamma \leq \min_j \frac{1}{-(w_b)_j},\ (w_b)_j < 0, \tag{8}$$

therefore

$$\gamma \leq \min_j \frac{1}{|(w_b)_j|} = \frac{1}{\max_j |(w_b)_j|} = \frac{1}{\|\mathbf{w}_b\|_\infty}. \qquad (9)$$

Since the primal objective has an lower bound, i.e. 0, thus the dual has a feasible bounded optimal solution and the optimum dual value is equal to the optimum primal value, therefore, $\gamma = \|u^*\|_1$ which is the statement of the proposition. ■

Based on Proposition 1 we have

$$\begin{array}{ll} \max & f(\mathbf{w}, b) = \dfrac{b}{\|\mathbf{w}\|_\infty} \\ \text{w.r.t} & \mathbf{w} \\ \text{s.t.} & \mathbf{z_i}'\mathbf{w} \geq b, i = 1, \ldots, m \\ & b > 0 \text{ given constant,} \end{array}$$

and after reformulation it as a minimisation problem we obtain

$$\begin{array}{ll} \min & \dfrac{\|\mathbf{w}\|_\infty}{b} \\ \text{w.r.t} & \mathbf{w} \\ \text{s.t.} & \mathbf{z_i}'\mathbf{w} \geq b, i = 1, \ldots, m \\ & b > 0 \text{ given constant.} \end{array} \qquad (10)$$

**Proposition 2** *If* $\mathbf{z}_i \geq \mathbf{0}$ *for all* $i = 1, \ldots, m$ *then the optimum solution for the Linear Programming Problem given by (10) is equal to*

$$\mathbf{w}^* = \frac{b}{\|\mathbf{z}_{i^*}\|_1}\mathbf{1}, \text{ where } i^* = \arg\min_i \|\mathbf{z}_i\|_1. \qquad (11)$$

**Proof:** First both side of the equality constraints are divided by $b$. Let us use the notation $\mathbf{w}_b$ for $\dfrac{\mathbf{w}}{b}$. So we can obtain

$$\begin{array}{ll} \min & \|\mathbf{w}_b\|_\infty \\ \text{w.r.t} & \mathbf{w}_b \\ \text{s.t.} & \mathbf{z_i}'\mathbf{w}_b \geq 1, i = 1, \ldots, m, \\ & b > 0 \text{ given constant,} \\ & \mathbf{z}_i \geq \mathbf{0},\ i = 1, \ldots, m. \end{array} \qquad (12)$$

We can recognise that $\mathbf{w}^* = \dfrac{1}{\|\mathbf{z}_{i^*}\|_1}\mathbf{1}$ is a feasible solution, since

$$\mathbf{z}_i'\mathbf{w}^* = \frac{1}{\|\mathbf{z}_{i^*}\|_1}\mathbf{z}_i'\mathbf{1} = \frac{\|\mathbf{z}_i\|_1}{\|\mathbf{z}_{i^*}\|_1} = \frac{\|\mathbf{z}_i\|_1}{\min_i \|\mathbf{z}_i\|_1} \geq 1. \qquad (13)$$

Now we need to prove that $\mathbf{w}^*$ is also an optimum solution. If it is not true then we can find a $\hat{\mathbf{w}}$ which is feasible and

$$\|\hat{\mathbf{w}}\|_\infty < \|\mathbf{w}^*\|_\infty.$$

This means that there is a constant $\beta$ such that $\hat{\mathbf{w}}_j \leq \beta < \dfrac{1}{\|\mathbf{z}_{i^*}\|_1}$ for any $j = 1, \ldots, n_z$. From $\mathbf{z}_i \geq \mathbf{0},\ i = 1, \ldots, m$ it follows that $\beta > 0$ otherwise the feasibility assumption is immediately violated.

Let us check the feasibility of $\hat{\mathbf{w}}$

$$(\mathbf{z}_{i^*})'\hat{\mathbf{w}} \leq \beta(\mathbf{z}_{i^*})'\mathbf{1} = \beta\|\mathbf{z}_{i^*}\|_1 < 1, \qquad (14)$$

hence, $\hat{\mathbf{w}}$ violates the constraint belonging to $\mathbf{z}_{i^*}$, with the smallest $L_1$ norm. Thus $\mathbf{w}^*$ is an optimum solution for (10). ■

With a constant, completely flat optimum solution, the predictor to a new $\boldsymbol{\phi}(x) \geq 0$ can be written as

$$(\widehat{\boldsymbol{\psi}(y)})_j = (\mathbf{W}\boldsymbol{\phi}(x))_j = b\frac{\|\boldsymbol{\phi}(x)\|_1}{\mathbf{z}_{i^*}},$$

which might not seem interesting at first sight, however let us now consider the sparse case in the next subsection.

### 4.1 Sparse case

First we define what we understand on sparseness in the problem given by (10). Sparseness means that there is at least an index $j \in \{1, \ldots, n_z\}$ such that for all $(\mathbf{z}_i)_j = 0$ holds. A consequence of this kind of sparseness is that the corresponding components of $\mathbf{w}$ has no influence on the feasibility, thus, this component is not determined except for the upper and lower bound imposed by the objective function, namely $\min \|\mathbf{w}\|_\infty$. Hence, the optimum solution to (10) becomes a set containing the elements obeying the form

$$\begin{array}{ll} (\mathbf{w}^*)_j & = \left\{ \begin{array}{ll} d & \text{if } \exists i = 1, \ldots, m,\ (\mathbf{z}_i)_j > 0, \\ \left[-d, d\right] & \text{otherwise,} \end{array} \right. \\ \text{where} & d = \dfrac{b}{\min_i \|\mathbf{z}_i\|_1} \end{array} \qquad (15)$$

Because $\mathbf{z}_i = \boldsymbol{\psi}(\mathbf{y}_i) \otimes \boldsymbol{\phi}(\mathbf{x}_i)$ then $(\mathbf{z}_i)_j$ is equal to 0 in each of the components where the corresponding components of either $\boldsymbol{\psi}(\mathbf{y}_i)$, or $\boldsymbol{\phi}(\mathbf{x}_i)$ or both, are equal to 0, which has high probability if both terms in the tensor product have lots of zero elements.

In the sparse case we can impose a further optimisation on our base problem, (10), which minimises the number of non-zero components of $\mathbf{w}$:

$$\begin{array}{ll} \min & \|\mathbf{w}\|_0 \\ \text{w.r.t.} & \mathbf{w}, \\ \text{s.t.} & \mathbf{w} \in \mathcal{W}, \\ \text{where} & \mathcal{W} = \left\{ \begin{array}{ll} \arg\min & \dfrac{\|\mathbf{w}\|_\infty}{b} \\ \text{w.r.t} & \mathbf{w} \\ \text{s.t.} & \mathbf{z_i}'\mathbf{w} \geq b, i = 1, \ldots, m \\ & b > 0 \text{ given constant,} \end{array} \right. \end{array} \qquad (16)$$

where the norm $\|\mathbf{w}\|_0$ means the number of non-zero components of the vector in the argument.

This extension gives us the following optimum solution

$$(\mathbf{w}^*)_j = \left\{ \begin{array}{ll} d & \text{if } \exists i = 1, \ldots, m,\ (\mathbf{z}_i)_j > 0, \\ 0 & \text{otherwise.} \end{array} \right. \qquad (17)$$

where $d$ is defined as before.

### 4.2 Scale independency

If the prediction $\boldsymbol{\psi}(\tilde{\mathbf{y}}) = \mathbf{W}\boldsymbol{\phi}(\mathbf{x})$ depends on the "shape" not on the scale we can employ a particular normalised solution

$$(\mathbf{w}^*)_j = \left\{ \begin{array}{ll} 1 & \text{if } \exists i = 1, \ldots, m,\ (\mathbf{z}_i)_j > 0, \\ 0 & \text{otherwise.} \end{array} \right. \qquad (18)$$

instead of the original one given by (15).

## 5 Kernelization

In $L_1$ norm we can not apply the same kernelization that is straightforwardly implemented in the $L_2$ norm case via the dual form of the optimisation problem. A simple approach to this problem is presented here. Let us consider the following embedding $\Phi : \mathcal{Z} \Rightarrow \mathbb{R}^m$ where the components of $\mathbf{\Phi}(\mathbf{z}_i)$ are defined by

$$\begin{aligned} (\mathbf{\Phi}(\mathbf{z}_i))_k &= \langle \mathbf{z}_k, \mathbf{z}_i \rangle \\ &= \langle \boldsymbol{\phi}(x_k) \otimes \boldsymbol{\psi}(y_k), \boldsymbol{\phi}(x_i) \otimes \boldsymbol{\psi}(y_i) \rangle \\ &= \underbrace{\langle \boldsymbol{\phi}(x_k), \boldsymbol{\phi}(x_i) \rangle}_{\kappa^{\phi}(x_k, x_i)} \underbrace{\langle \boldsymbol{\psi}(y_k), \boldsymbol{\psi}(y_i) \rangle}_{\kappa^{\psi}(y_k, y_i)}, \end{aligned} \tag{19}$$

where $\kappa^{\phi}$ and $\kappa^{\psi}$ are the kernel functions. Changing the margin constraints in (4) from

$$\text{s.t.} \quad \mathbf{z_i}'\mathbf{w} \geq b, i = 1, \ldots, m, \tag{20}$$

into

$$\text{s.t.} \quad \mathbf{\Phi}(\mathbf{z_i})'\mathbf{w}_{\Phi} \geq b, i = 1, \ldots, m, \tag{21}$$

we received the same structure, thus, the solution to the modified problem follows the same pattern. One can recognise that all we did is nothing more than expressing the linear operator projecting the input into the output space by

$$\mathbf{W} = \textstyle\sum_{k=1}^{m} (\mathbf{w}_{\Phi})_k \boldsymbol{\psi}(y_k) \otimes \boldsymbol{\phi}(x_k). \tag{22}$$

A similar kernelization is applied in for example [Man00] to the binary Support Vector Machine.

## 6 Prediction

The prediction to an arbitrary $x$ can be given by

$$y = \boldsymbol{\psi}^{-1}(\mathbf{W}\boldsymbol{\phi}(x)),$$

but to give substance and some interpretation to this formula we need to make some additional remarks. The prediction of $\mathbf{W}\boldsymbol{\phi}(x)$ gives a score vector to every input vector, but this score vector lives in the linear vector space representing the output and not the output object itself. For the complete solution we need to find an inverse image, which is not always straightforward. The inversion can be carried out if we know the structure of the output space and how it is represented. The first step the computation of the scores to $\boldsymbol{\psi}(y)$ is demonstrated in Figure 1.

The prediction can be derived by taking the optimum solution of (18). To find the optimal output $y$ to an input $x$ we need to establish a model First, to invert the function $\psi$ the next optimisation schema can be applied:

$$\begin{array}{lll} \hat{y} = & \arg\max & \Big\langle \mathbf{W}, \big[\boldsymbol{\psi}(\mathbf{y}) \otimes \boldsymbol{\phi}(\mathbf{x})\big] \Big\rangle_{\mathcal{L}_{\psi} \otimes \mathcal{L}_{\phi}} \\ & \text{w.r.t.} & y \\ & \text{s.t.} & y \in \hat{\mathcal{Y}}, \end{array} \tag{23}$$

where $\hat{\mathcal{Y}}$ is the set of possible outputs. The next element of our inversion model is the assumption such that the vectors of $\boldsymbol{\psi}(y),\ y \in \hat{\mathcal{Y}}$ are indicator vectors of the possible patterns should appear in the output, so that we are looking for a vector with a relatively small number of non zeros pointing to the patterns and all other components ought to be set to 0 (i.e. this is explicit in the machine translation example mentioned earlier; vectors are indexed by the entire set of possible phrases/features in a language, but elements are only non-zero for those that actually occur in any given sentence). Our solution approach can exploit the fact that the objective function of (23) is linear in $\boldsymbol{\psi}(y)$ and can be written as $\mathbf{d}'\boldsymbol{\psi}(y)$, where $\mathbf{d} = \mathbf{W}\boldsymbol{\phi}(x)$. However, the maximum of a linear function can be finite only if the components of the vector $\boldsymbol{\psi}(y)$ are bounded components-wise or(and) in a norm. To this end let us consider the following problem

$$\begin{array}{ll} \max & \mathbf{d}'\mathbf{u} \\ \text{w.r.t.} & \mathbf{u}, \\ \text{s.t.} & \mathbf{1}'\mathbf{u} = 1, \\ & 0 \leq \mathbf{u} \leq C, \end{array} \tag{24}$$

which has a simple optimum solution. To derive that first we need to sort the components of vector $\mathbf{d}$ into a decreasing order

$$(d_1, \ldots, d_{n_d}) \Rightarrow (d_{i_1} \geq, d_{i_2}, \ldots, \geq d_{i_{n_d}}), \tag{25}$$

and let $N$ the smallest integer greater than $\frac{1}{C}$, then an optimum is given by

$$u_{i_k} = \begin{cases} C & \text{if } k < K, \\ 1 - C(K-1) & \text{if } k = K, \\ 0 & \text{otherwise} \end{cases} \tag{26}$$

Now the task can be stated as finding an optimal bound $C$ which can approximate the number of nonzero items in the indicator which we look for.

Let us consider the following general two level problem, where the inner part follows (24)

$$\max_{t=1,2,\ldots,} \left\{ \begin{array}{ll} \max & \dfrac{\mathbf{d}'\mathbf{u}_t}{g(t)} \\ \text{w.r.t} & \mathbf{u}_t \\ \text{s.t.} & \mathbf{1}'\mathbf{u}_t = 1, \\ & 0 \leq \mathbf{u}_t \leq g(t) \end{array} \right\}, \tag{27}$$

where $t$ is running on the positive integers and $g$ a real valued monotone increasing function.

What is expressed here is that if $t = 1$ we chose the component of $\mathbf{d}$ with the highest value then the two highest ones and so on.We might stop at the first local minimum of the outer problem. The function $g$ relates to the speed of the decay of decreasingly ordered components of the vector $\mathbf{d}$, thus, in this way it is problem dependent. Possible choices are $g(t) = log(t)$ or $g(t) = \sqrt{t}$. The linear case $g(t) = t$ is obviously wrong, since in case of a decreasing sequence it gives surely the optimum when $t = 1$. Further research activity is needed to estimate a good candidate.

The prediction is then derived of the optimum solution $\mathbf{u}^{*}_{t^{*}}$ of (27) interpreting it as an indicator vector. We need to mention the number of non-zero components in this optimum solution will be proportional to the non-zero components of the input but this connection is not a direct one.

## 7 An online framework, a set based perceptron learner

The motivation of the online approach stems of the structure of the solution to the sparse case which, in turn, can be interpreted as a learning method of the possible co-existences of the parts of input and output vectors via an indicator vector.

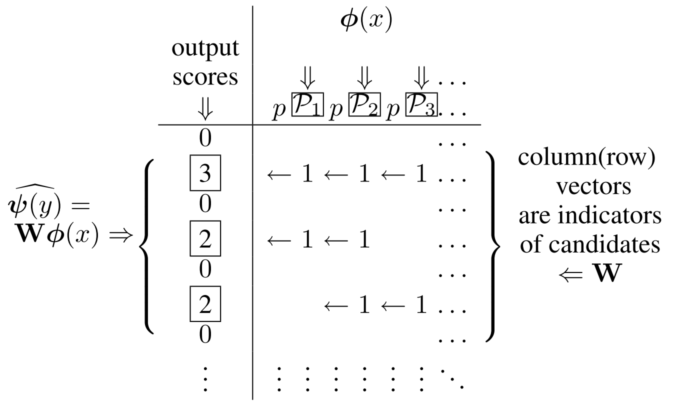
$\boldsymbol{\phi}(x)$
output scores
$\Downarrow$
$\Downarrow \quad \Downarrow \quad \Downarrow \ldots$
$p\,\boxed{\mathcal{P}_1}\,p\,\boxed{\mathcal{P}_2}\,p\,\boxed{\mathcal{P}_3}\ldots$
$\widehat{\boldsymbol{\psi}(y)} = \mathbf{W}\boldsymbol{\phi}(x) \Rightarrow$
0
$\boxed{3}$
0
$\boxed{2}$
0
$\boxed{2}$
0
$\leftarrow 1 \leftarrow 1 \leftarrow 1 \ldots$
$\leftarrow 1 \leftarrow 1$
$\leftarrow 1 \leftarrow 1 \ldots$
column(row) vectors are indicators of candidates $\Leftarrow \mathbf{W}$

Figure 1: The prediction schema

First we outline the problem that we are going to solve. Let us assume that the input and output objects are represented by the following way:

- Given two, supposed to be finite, sets $\Omega_x$ and $\Omega_y$ the collections of the possible pattern characterising the objects $x$ and $y$.
- Every observed $x$ and $y$ are described by function $\mu_x : \Omega_x \rightarrow \mathbb{R}_+$ and $\mu_y : \Omega_y \rightarrow \mathbb{R}_+$, where $\mathbb{R}_+$ denotes the nonnegative real numbers. They may be measures of the importance of the patterns.
- Assume that $\sum_{\omega_x \in \Omega_x} \mu_x(\omega_x) = 1$ and $\sum_{\omega_y \in \Omega_y} \mu_y(\omega_y) = 1$, the measures are normalised in $L_1$ norm. Thus, $\mu_x$ and $\mu_y$ might be interpreted as probabilities.
- Let $\phi(x) = (\mu_x(\omega_x))$ and $\psi(y) = (\mu_y(\omega_y))$ the vectors of the weights of the patterns.

From these conditions we can derive $\Omega_Z = \Omega_x \times \Omega_y$, and $\mathbf{z} = \phi(x) \otimes \psi(y)$ a tensor product which as a consequence is normalised to 1 as well.

Based on the model stated we can formulate the next learning task which is to find $\mathbf{w}$ a linear functional of the space $\{0,1\}^{|\Omega_z|}$ corresponding to a subset $\mathcal{W}$ of $\Omega_z$ such that $\langle \mathbf{w}, \mathbf{z}_i \rangle \geq \lambda$, where $\lambda > 0$ is given constant and $i = 1, \ldots, m$ are the indeces of the sample items. It means by the definition of the inner product

$$\langle \mathbf{w}, \mathbf{z}_i \rangle = \sum_{\omega_{x_i} \in \Omega_{x_i}, \omega_{y_i} \in \Omega_{y_i}} \mu_{x_i}(\omega_{x_i}) \mu_{y_i}(\omega_{y_i}),$$

thus, that we are looking for is a set $\mathcal{W}$ which comprises a sufficiently large number of the most important common patterns of all sample items.

We can associate sets represented by indicators to all the sample items $\mathbf{z}_i = \boldsymbol{\psi}(\mathbf{y}_i) \otimes \boldsymbol{\phi}(\mathbf{x}_i)$ and to the linear functional $\mathbf{w}$ as well by

$$\begin{array}{lcl} \mathbf{z}_i & \rightarrow & [\mathbf{z}_i > 0] = \{z_{ij} > 0\} = \mathcal{Z}_i \\ \mathbf{w} & \rightarrow & [\mathbf{w} > 0] = \{w_j > 0\} = \mathcal{W}. \end{array} \qquad (28)$$

Since all the nonzero components of $\mathbf{w}$ are equal to a constant, thus we can restore $\mathbf{w}$ from its set based representation via a bijective mapping between the vector and set representation.

**Algorithm 1** Primal perceprtron for sets

```
**Input of the learner:** The sample $S$,
**Output of the learner:** $\mathbf{W} \in \mathbb{R}^{dim(\mathcal{H}_y)dim(\mathcal{H}_x)}$
**Initialisation:** $\mathcal{W}_t = \emptyset$; $i = 1$;,
$noUpdate = true$
**repeat**
  **for** $i = 1, 2, \ldots, m$ **do**
    read input: $\mathbf{z}_i = \boldsymbol{\psi}(\mathbf{y}_i) \otimes \boldsymbol{\phi}(\mathbf{x}_i)$
    $t = 0$
    **if** $\langle \mathbf{w}_t, \mathbf{z}_i \rangle < \lambda$ **then**
      $\mathbf{w}_t \rightarrow \mathcal{W}_t$
      $\mathbf{z}_i \rightarrow \mathcal{Z}_i$
      $\boxed{\mathcal{W}_{t+1} = \mathcal{W}_t \cup \mathcal{Z}_i}$ {Set update}
      $\mathcal{W}_{t+1} \rightarrow \mathbf{w}_{t+1}$
      $t = t + 1$
      $noUpdate = false$
    **end if**
  **end for**
**until** $noUpdate$ is $true$
```

We can write up a perceptron type algorithm, see details in [CST00], for solving problems following the schema of the $L_1$ norm based maximum margin learner. The algorithm is given by Algorithm 1.

A Novikoff-style bound [Nov62] on this algorithm can be stated. Based on the definition of $\mathbf{w}$ for any two realisations $\mathbf{w}_k$ and $\mathbf{w}_l$ $\langle \mathbf{w}_k, \mathbf{w}_l \rangle = |\mathcal{W}_k \cap \mathcal{W}_l|$, and let $\|\mathbf{w}\|_2 = \langle \mathbf{w}, \mathbf{w} \rangle = |\mathcal{W}|$.

We can define a margin to the perceptron learner by

$$\gamma(\mathbf{w}, \mathcal{S}) = \min_{(y_i, x_i) \in \mathcal{S}} \frac{\langle \mathbf{w}, \mathbf{z}_i \rangle}{\|\mathbf{w}\|_2}, \ \mathbf{z}_i = \boldsymbol{\psi}(y_i) \otimes \boldsymbol{\phi}(x_i). \qquad (29)$$

Let $\delta_i(\lambda, \mathbf{w})$ be such that if $\langle \mathbf{w}, \mathbf{z}_i \rangle \geq \lambda$ then $|\mathcal{W} \cap \mathcal{Z}_i| \geq \delta_i(\lambda, \mathbf{w})$. This kind of $\delta_i(\lambda, \mathbf{w}) > \ell_{min}$ has to exist as a consequence of the definition of the inner product, the nonnegativity and the normalisation of the sample items. Now we consider the minimum of them, i.e.

$$\delta(\lambda, \mathbf{w}) = \min_{i=1,\ldots,m} \delta_i(\lambda, \mathbf{w})$$

. The non-negativity guarantees $\delta(\lambda, \mathbf{w})$ is monotonic, increasing function of its two variables, so, greater $\lambda$ or $|\mathcal{W}|$ implies greater $\delta$.

**Theorem 3** *Let $\mathcal{S} = \{(x_i, y_i)\} \subset (\mathcal{Y} \times \mathcal{X})$, $i = 1, \ldots,$ be a sample independently and identically drawn from an unknown distribution and let $\boldsymbol{\phi} : \mathcal{X} \rightarrow \mathcal{L}_\phi$ and $\boldsymbol{\psi} : \mathcal{Y} \rightarrow \mathcal{L}_\psi$ be mappings into spaces of tuples of indicators.*

*Let $\mathbf{z}_i = \boldsymbol{\psi}(\mathbf{y}_i) \otimes \boldsymbol{\phi}(\mathbf{x}_i)$ be tensor product and $\mathcal{Z}_i$ the indicator set of nonzero items in $\mathbf{z}_i$.*

*Assume that $0 < \lambda < 1$, and $0 < \ell_{min} \leq |\mathcal{Z}_i| \leq \ell_{max}$, $i = 1, \ldots, m$, the support, the number of patterns, of every sample items falls within a given range.*

*Furthermore, there is a $\mathbf{w}_*$ such that Algorithm (1) stops with no more update, $\langle \mathbf{w}_*, \mathbf{z}_i \rangle \geq \lambda$, $i = 1, ..., m$, and then*

1. *the number of updates in Algorithm (1) is bounded by*

$$t \leq \frac{\ell_{max}}{\Delta^2}, \qquad (30)$$

2. *the margin at the solution* $\mathbf{w}_t$ *has a lower bound*

$$\gamma(\mathbf{w}_t, \mathcal{S}) \geq \frac{\lambda\Delta}{\ell_{max}}, \tag{31}$$

*where* $\Delta = \dfrac{\delta(\lambda, \mathbf{w}_*)}{\|\mathbf{w}_*\|_2}$.

**Proof:** We are going to follow the main thread of the Novikoff's reasoning, but with some extensions.

Because $\mathbf{w}_*$ satisfies all the margin constraints, therefore for any $t$ we have $|\mathcal{W}_*| \geq |\mathcal{W}_t|$, thus, $\delta(\lambda, \mathbf{W}_*) \geq \delta(\lambda, \mathbf{W}_t)$. Let us use the short notation $\delta(\lambda) = \delta(\lambda, \mathbf{W}_*)$.

After $t$ steps of update the squared $L_2$ norm of $\mathbf{w}_t$ can be bounded by

$$\|\mathbf{w}_t\|_2^2 \quad = |\mathcal{W}_t| \leq \ell_{max} t. \tag{32}$$

Now consider the following inner product

$$\begin{aligned} \langle \mathbf{w}_*, \mathbf{w}_t \rangle &= |\mathcal{W}_* \cap \mathcal{W}_t| \\ &= |\mathcal{W}_* \cap (\mathcal{W}_{t-1} \cup \mathcal{Z}_{i_t})| \\ &= |(\mathcal{W}_* \cap \mathcal{W}_{t-1}) \cup (\mathcal{W}_* \cap \mathcal{Z}_{i_t})| \end{aligned} \tag{33}$$

by induction on $t$

$$= |\mathcal{W}_* \cap (\textstyle\bigcup_t \mathcal{Z}_{i_t})| \geq \delta(\lambda) t. \tag{34}$$

since for any $t$ $\mathcal{W}_t \subseteq \mathcal{W}_*$ holds.

Merging the inequalities above we obtain

$$\sqrt{\ell_{max} t}\|\mathbf{w}_*\|_2 \geq \|\mathbf{w}_*\|_2 \|\mathbf{w}_t\|_2 \geq \langle \mathbf{w}_*, \mathbf{w}_t \rangle \geq \delta(\lambda) t. \tag{35}$$

It gives us an upper bound on the number of updates as function of the functional margin

$$t \leq \frac{\ell_{max}}{\Delta^2}. \tag{36}$$

After substituting this inequality into (32) we have

$$\|\mathbf{w}_t\|_2 \leq \frac{\ell_{max}}{\Delta}, \tag{37}$$

and at the end for the functional margin the lower bound can be obtained

$$\gamma(\mathbf{w}_t, \mathcal{S}) \geq \frac{\lambda\Delta}{\ell_{max}}, \tag{38}$$

which statement completes the proof. ■

**Remark 4** *One can recognise behind this scenario a special variant of the weighted set covering problem. The sample items to be found as errors in the perceptron algorithm give a cover to all of the patterns occurring in the sample. It is obviously not an optimum cover, the smallest in cardinality of all possible ones, but a sufficiently good one. It is an open question how to extend the range of the applications of the machine learning algorithms of this kind to produce approximations for hard combinatorial problems. This relationship allow us to find some connections between the* $L_1$ *norm based learning and the Set Covering Machine introduced in [MST02].*

## 8 Discussion

We have shown in this paper that measuring the margin by $L_1$ norm in a maximum margin learning problem gives a simple solution to an otherwise hardly tractable class of structural learning tasks.

## References

[AHP+08] K. Astikainen, L. Holm, E. Pitknen, S. Szedmak, and J. Rousu. Towards structured output prediction of enzyme function. In *BMC Proceedings, 2(Suppl 4):S2*. 2008.

[CST00] N. Cristianini and J. Shawe-Taylor. *An introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[DBST01] A. Demiriz, K. P. Bennett, and J. Shawe-Taylor. Linear programming boosting via column generation. *Machine Learning*, 46:1:225–254, 2001.

[DG85] L. Devroye and L. Györfi. *Nonparametric Density Estimation: The L1 View*. John Wiley, New York, 1985.

[Man00] O. L. Mangasarian. Generalized support vector machines. In *Advances in Large Margin Classifiers*, pages 135–146. MIT Press, 2000.

[MST02] M. Marchand and J. Shawe-Taylor. The set covering machine. *Journal of Machine Learning Research*, 3, 2002.

[Nov62] A. Novikoff. On convergence proofs for perceptrons. In *In Report at the Symposium on Mathematical Theory of Automata*, pages 24–26. Politechnical Institute Brooklyn, 1962.

[Pin89] A.M. Pinkus. *On L1-approximation*. Cambridge University Press, 1989.

[RSSST06] J. Rousu, C.J. Saunders, S. Szedmak, and J. Shawe-Taylor. Kernel-based learning of hierarchical multilabel classification models. *Journal of Machine Learning Research*, Special issue on Machine Learning and Large Scale Optimization, 2006.

[SSTPH05] S. Szedmak, J. Shawe-Taylor, and E. Parado-Hernandez. Learning via linear operators: Maximum margin regression. In *PASCAL Research Reports, http://eprints.pascal-network.org/*. 2005.

[TGK03] B. Taskar, C. Guestrin, and D. Koller. Max margin markov networks. In *NIPS 2003*. 2003.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society, Series B*, 58:267–288, 1996.

[TJHA05] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research (JMLR)*, 6(Sep):1453–1484, 2005.

[TLJJ06] B. Taskar, S. Lacoste-Julien, and M. I. Jordan. Structured prediction, dual extragradient and bregman projections. In *JMLR, Special Topic on Machine Learning and Optimization*, pages 1627–1653. 2006.

[Vap98] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.