# Risk Bounds and Learning Algorithms for the Regression Approach to Structured Output Prediction

**Sébastien Giguère** SEBASTIEN.GIGUERE.8@ULAVAL.CA
**François Laviolette** FRANCOIS.LAVIOLETTE@IFT.ULAVAL.CA
**Mario Marchand** MARIO.MARCHAND@IFT.ULAVAL.CA
**Khadidja Sylla** KHADIDJA.SYLLA.1@ULAVAL.CA
Département d'informatique et de génie logiciel, Université Laval, Québec (QC), Canada, G1V-0A6

## Abstract

We provide rigorous guarantees for the regression approach to structured output prediction. We show that the quadratic regression loss is a convex surrogate of the prediction loss when the output kernel satisfies some condition with respect to the prediction loss. We provide two upper bounds of the prediction risk that depend on the empirical quadratic risk of the predictor. The minimizer of the first bound is the predictor proposed by Cortes et al. (2007) while the minimizer of the second bound is a predictor that has never been proposed so far. Both predictors are compared on practical tasks.

## 1. Introduction

Structured output prediction is a supervised learning problem where the goal of the learner is to predict the correct output $y$ associated to some given input $x$. Here, the output $y$ can be a complex structure such as a sequence of symbols, a parse tree, or a graph. The predictor generally consists of a vector $w$ of real-valued weights and each input-output example $(x, y)$ is mapped to a high-dimensional feature vector $Z(x, y)$. The output predicted by $w$ on input $x$ is then the output $y$ that maximizes the inner product $\langle w|Z(x, y)\rangle$. However, as emphasized by Gärtner & Vembu (2009), this pre-image problem is often $\mathcal{NP}$-hard. Consequently, any learning algorithm that needs to solve this pre-image problem, for several weight vectors and every training example, will often take a prohibitive running time. This is probably the most important problem facing several state-of-the-art structured output learning algorithms such as max-margin Markov networks (Taskar et al., 2004) and the structural SVM (Tsochantaridis et al., 2005).

One of the first attempts to design a learning algorithm that avoids the pre-image problem is due to Cortes et al. (2007). They have proposed to find the predictor that minimizes an $\ell_2$-regularized regression objective (which does not depend on the predicted output for a given input) and have obtained empirical results that compare favorably to those of structural SVM and max-margin Markov networks on the word-recognition data set used by Taskar et al. (2004). In this paper, we provide guarantees for such a regression approach by first showing that the quadratic loss function used by Cortes et al. (2007) provides a convex upper bound on the original prediction loss (that depends on the predicted output) provided that the output kernel satisfies some condition with respect to the prediction loss. We also provide two PAC-Bayes upper bounds (McAllester, 2003; Langford, 2005) for the prediction risk that depend on the quadratic empirical loss used by Cortes et al. (2007). The minimizer of the first bound turns out to be the same as the one proposed by Cortes et al. (2007) while the minimizer of the second bound, valid for arbitrary reproducing kernel Hilbert spaces (RKHS), is proposed for the first time. Both predictors are compared on practical tasks.

PAC-Bayes theory has also been applied recently (McAllester, 2007) to structured output prediction for a stochastic predictor that aims at minimizing the expected prediction risk. The resulting learning algorithms need to solve the pre-image problem on each example of the training set for each update of the predictor. In contrast, we present here risk bounds for learning algorithms that avoid solving the pre-image problem and that produce a deterministic predictor instead of a stochastic one.

*Proceedings of the $30^{th}$ International Conference on Machine Learning*, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. 

## 2. From Structured Output Prediction to Regression

In the supervised learning setting, the learner has access to a set $S \stackrel{\text{def}}{=} \{(x_1, y_1), \ldots, (x_m, y_m)\}$ of $m$ training examples where each example consists of an input-output pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$. The input space $\mathcal{X}$ and the output space $\mathcal{Y}$ are both arbitrary but we assume the existence of both an input feature map $X : \mathcal{X} \to \mathcal{H}_\mathcal{X}$ and an output feature map $Y : \mathcal{Y} \to \mathcal{H}_\mathcal{Y}$, where both $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$ are high-dimensional vector spaces and, more generally, reproducing kernel Hilbert spaces (RKHS). In $\mathcal{H}_\mathcal{Y}$, we use $\langle Y(y)|Y(y')\rangle$ to denote the inner product and use $\|Y(y)\|^2 \stackrel{\text{def}}{=} \langle Y(y)|Y(y)\rangle$ for the squared norm. The same notation is used in $\mathcal{H}_\mathcal{X}$.

Given access to a training set $S$, the task of the learner is to construct a structured predictor which is represented by a linear operator $\mathbf{W}$ that transforms vectors of $\mathcal{H}_\mathcal{X}$ into vectors of $\mathcal{H}_\mathcal{Y}$. For any $x \in \mathcal{X}$ and any $\mathbf{W}$, the output $y_\mathbf{W}(x)$ predicted by $\mathbf{W}$ is given by

$$y_\mathbf{W}(x) \stackrel{\text{def}}{=} \operatorname*{argmin}_{y \in \mathcal{Y}} \|Y(y) - \mathbf{W}X(x)\| \,. \qquad (1)$$

Note that $y_\mathbf{W}(x) = \operatorname{argmax}_{y \in \mathcal{Y}} \langle Y(y)|\mathbf{W}X(x)\rangle$ whenever $\|Y(y)\|$ is the same for all $y \in \mathcal{Y}$. In this case, we recover the usual structured output prediction method when the joint feature vectors $Z(x, y)$ are tensor products $X(x) \otimes Y(y)$. Since finding $y_\mathbf{W}(x)$ given $x$ and $\mathbf{W}$ is generally $\mathcal{NP}$-hard (Gärtner & Vembu, 2009), we want to avoid solving this pre-image problem.

We consider feature maps that are defined by kernels such that $K_\mathcal{Y}(y, y') = \langle Y(y)|Y(y')\rangle \; \forall (y, y') \in \mathcal{Y}^2$ and $K_\mathcal{X}(x, x') = \langle X(x)|X(x')\rangle \; \forall (x, x') \in \mathcal{X}^2$. We will see that the proposed solutions for $\mathbf{W}$ will have the property that $\mathbf{W}X(x) = \sum_{i=1}^m \sum_{j=1}^m Y(y_i) A_{i,j} K_\mathcal{X}(x_j, x)$ for some $m \times m$ matrix $\mathbf{A}$. Consequently, the predicted output $y_\mathbf{W}(x)$ only requires the use of the kernels $K_\mathcal{X}$ and $K_\mathcal{Y}$ (instead of the feature maps $X$ and $Y$).

We assume that each example $(x, y)$ is generated independently according to some unknown distribution $D$. Given a function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ that quantifies the loss incurred on $(x, y)$ when the predicted output is $y_\mathbf{W}(x)$, the task of the learner is to find the predictor that minimizes the expected loss (or risk) $\mathbf{E}_{(x,y)\sim D} L(y_\mathbf{W}(x), y)$. We refer to $L$ as the *prediction loss*.

Note that the output kernel $K_\mathcal{Y}$, being a similarity measure on $\mathcal{Y}^2$, induces a loss function $L_{K_\mathcal{Y}}$ defined as

$$L_{K_\mathcal{Y}}(y_\mathbf{W}(x), y) \stackrel{\text{def}}{=} \frac{1}{2}\|Y(y) - Y(y_\mathbf{W}(x))\|^2 = \frac{K_\mathcal{Y}(y, y) + K_\mathcal{Y}(y_\mathbf{W}(x), y_\mathbf{W}(x))}{2} - K_\mathcal{Y}(y, y_\mathbf{W}(x)) \,. \qquad (2)$$

We refer to $L_{K_\mathcal{Y}}$ as the *output kernel loss*.

Both the prediction loss and the output kernel loss on $(x, y)$ depend on the predicted output $y_\mathbf{W}(x)$. This is in sharp contrast with the *quadratic loss* $\|Y(y) - \mathbf{W}X(x)\|^2$ which does not depend on $y_\mathbf{W}(x)$. However we can show that the quadratic loss provides an upper bound to the output kernel loss.

**Lemma 1.** *For any structured predictor* $\mathbf{W}$ *giving predictions as defined by Equation* (1)*, for any* $(x, y) \in \mathcal{X} \times \mathcal{Y}$*, we have*

$$L_{K_\mathcal{Y}}(y_\mathbf{W}(x), y) \;\leq\; 2\,\|Y(y) - \mathbf{W}X(x)\|^2 \,.$$

*Proof.* From the triangle inequality, we have, for all $\mathbf{W}$ and for all $(x, y)$,

$$\|Y(y) - Y(y_\mathbf{W}(x))\| \leq \|Y(y) - \mathbf{W}X(x)\| + \|Y(y_\mathbf{W}(x)) - \mathbf{W}X(x)\| \,.$$

From Equation (1), we have $\|Y(y_\mathbf{W}(x)) - \mathbf{W}X(x)\| \leq \|Y(y) - \mathbf{W}X(x)\|$ for all $\mathbf{W}$ and for all $(x, y)$. From these two inequalities, we have $\|Y(y) - Y(y_\mathbf{W}(x))\| \leq 2\|Y(y) - \mathbf{W}X(x)\|$, which gives the lemma. □

Lemma 1 has far-reaching consequences whenever we use an output kernel $K_\mathcal{Y}$ such that $L(y, y') \leq L_{K_\mathcal{Y}}(y, y')$ for all $(y, y') \in \mathcal{Y}^2$ because, in that case, we have

$$L(y_\mathbf{W}(x), y) \;\leq\; 2\,\|Y(y) - \mathbf{W}X(x)\|^2 \,,$$

for all predictors $\mathbf{W}$ and all $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

Under these circumstances, a predictor $\mathbf{W}$ having a small quadratic risk $\mathbf{E}_{(x,y)\sim D} \|Y(y) - \mathbf{W}X(x)\|^2$ has also a small prediction risk $\mathbf{E}_{(x,y)\sim D} L(y_\mathbf{W}(x), y)$. To minimize the structured prediction risk, we need to solve the (usually hard) pre-image problem of finding the predicted output $y_\mathbf{W}(x)$ for every example in the training set and for all predictors $\mathbf{W}$ tried by the learning algorithm. Thanks to Lemma 1, we can avoid this computational burden by performing the simpler regression task of minimizing the quadratic risk whenever we use an output kernel $K_\mathcal{Y}$ for which the output kernel loss $L_{K_\mathcal{Y}}$ upper bounds the prediction loss $L$.

Consider the case when the prediction loss $L$ is the zero-one loss. In that case any output kernel $K_\mathcal{Y}$ for which there exists two different outputs $y$ and $y'$ having $K_\mathcal{Y}(y, y) = K_\mathcal{Y}(y', y') = K_\mathcal{Y}(y, y')$ will not give an output kernel loss $L_{K_\mathcal{Y}}$ which upper bounds $L$. But the Dirac kernel for which $K_\mathcal{Y}(y, y') = 1$ if $y = y'$ and 0 otherwise gives an $L_{K_\mathcal{Y}}$ which is identical to $L$.

In the case where the prediction loss $L$ is the Hamming distance, the Hamming kernel provides an output structured loss $L_{K_\mathcal{Y}}$ identical to $L$ but one could

also use any output kernel giving an $L_{K_\mathcal{Y}}$ which upper bounds the Hamming distance at the expense of introducing an additional slackness between the quadratic risk and the prediction risk.

A predictor achieving a small quadratic risk also achieves a small prediction risk when the output kernel $K_\mathcal{Y}$ gives an $L_{K_\mathcal{Y}}$ which upper bounds $L$. However, their exists data-generating distributions where the predictor achieving the smallest possible quadratic risk has a substantially larger prediction risk than the predictor achieving the smallest possible prediction risk. In other words, there is no consistency guarantee for the regression approach to structured output prediction because no such guarantee exists for the particular case of binary classification[1]. However, the regression approach avoids the computational burden of dealing with the pre-image problem and, under some distributions, there might be some kernels for which there exists predictors achieving a small quadratic risk.

Thanks to Lemma 1, any upper bound on the quadratic risk also provides a bound on the prediction risk (provided that there exists an output kernel loss that upper bounds the prediction loss). Consequently, the upper bounds proposed by Caponnetto & De Vito (2007); Baldassarre et al. (2012) also provide bounds on the prediction risk for predictors minimizing the $\ell_2$-regularized least-squares. However, instead of focussing explicitly on such predictors, we provide bounds that hold simultaneously for any predictor $\mathbf{W}$ and that depend on the empirical quadratic risk achieved by $\mathbf{W}$ on the training data.

## 3. PAC-Bayes with Isotropic Gaussians

Values of the prediction loss $L(y_\mathbf{W}(x), y)$ are always between zero an one. However, this is clearly not the case for the quadratic loss $\|Y(y) - \mathbf{W}X(x)\|^2$. Theoretically attainable very large loss values are well known to give very loose concentration inequalities and, unavoidably, very large risk bounds. Therefore, to obtain a tighter risk bound, we use the following lemma which upper bounds the prediction loss in terms of a bounded function of the quadratic loss.

**Lemma 2.** *For any prediction loss $L$ upper-bounded by the output kernel loss $L_{K_\mathcal{Y}}$, for any $(x, y)$, any $\mathbf{W}$, and any $a \geq 1$, we have*

$$L(y_\mathbf{W}(x), y) \;\leq\; \frac{ae}{e-1}\left(1 - e^{-\frac{2}{a}\|Y(y) - \mathbf{W}X(x)\|^2}\right).$$

[1]Indeed, it is easy to find distributions for which the minimizer of the quadratic risk gives a classifier which achieves a much larger 0-1 risk than the optimal classifier. See the supplementary material for a simple example.

*Proof.* For any $0 \leq x \leq 1$, we have $x \leq \frac{e}{e-1}\left(1 - e^{-x}\right)$. Therefore

$$\begin{aligned}\frac{1}{a}L(y_\mathbf{W}(x), y) &\leq \frac{e}{e-1}\left(1 - e^{-\frac{1}{a}L(y_\mathbf{W}(x), y)}\right)\\ &\leq \frac{e}{e-1}\left(1 - e^{-\frac{2}{a}\|Y(y) - \mathbf{W}X(x)\|^2}\right),\end{aligned}$$

where the last equality follows from Lemma 1 and the fact that $L$ is upper-bounded by $L_{K_\mathcal{Y}}$. □

### 3.1. The Risk Bound

We propose here an upper bound on the prediction risk that uses PAC-Bayes theory to upper bound $\mathbf{E}_{(x,y)\sim D}\left(1 - e^{-\frac{2}{a}\|Y(y) - \mathbf{W}X(x)\|^2}\right)$ for some $a \geq 0$. However, PAC-Bayes theory does not directly provide bounds on deterministic predictors such as $\mathbf{W}$. Instead, it provides guarantees for stochastic Gibbs predictors that are described in terms of a posterior distribution $Q$ over deterministic predictors. More precisely, PAC-Bayes theory provide bounds for Gibb's risk defined as the $Q$-average of the risk of deterministic predictors. The following theorem, due to Zhang (2006), provides an example of such a bound[2].

**Theorem 3.** *(from Zhang (2006)) Let $\zeta$ be any loss function, and let $P$ be any prior distribution on $\mathcal{V}$. Then, for any $D$ on $\mathcal{X} \times \mathcal{Y}$, with probability at least $1-\delta$ over all training sets $S$ sampled according to $D^m$, we have, simultaneously for all distributions $Q$ on $\mathcal{V}$,*

$$-\mathop{\mathbf{E}}_{\mathbf{v}\sim Q} \ln \mathop{\mathbf{E}}_{(x,y)\sim D} e^{-\zeta(\mathbf{v}, x, y)} \leq$$
$$\frac{1}{m}\left(\mathop{\mathbf{E}}_{\mathbf{v}\sim Q}\sum_{i=1}^{m}\zeta(\mathbf{v}, x_i, y_i) + \mathrm{KL}(Q, P) + \ln\frac{1}{\delta}\right),$$

*where $\mathrm{KL}(Q, P)$ denotes the Kullback-Leibler divergence between distributions $Q$ and $P$.*

To use Theorem 3, we restrict ourselves (in this section) to the case where both feature spaces $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$ are finite-dimensional vector spaces of dimensions $N_\mathcal{X}$ and $N_\mathcal{Y}$ respectively. The set of predictors thus coincides with the set of $N_\mathcal{Y} \times N_\mathcal{X}$ matrices. Each posterior is chosen to be an isotropic Gaussian of variance $\sigma^2$ and expectation $\mathbf{W}$. If $Q_{\mathbf{W},\sigma}(\mathbf{V})$ denotes the density at matrix $\mathbf{V}$ of this posterior, we have

$$Q_{\mathbf{W},\sigma}(\mathbf{V}) \;=\; \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{N_\mathcal{X} N_\mathcal{Y}} e^{-\frac{1}{2\sigma^2}\|\mathbf{V} - \mathbf{W}\|^2}, \qquad (3)$$

where, for any matrix $\mathbf{V}$, $\|\mathbf{V}\|^2 \stackrel{\text{def}}{=} \sum_{i=1}^{N_\mathcal{Y}}\sum_{j=1}^{N_\mathcal{X}} V_{i,j}^2$ (also called the Frobenius norm).

[2]Unfortunately, there is no dependence on the sample-size $m$ in the theorem sated by Zhang (2006) because the one-example formulation was used. We obtain Theorem 3 if we use $m$ examples instead of one.

For the Prior $P$, we chose the (non-informative) isotropic Gaussian centered at the origin, *i.e.*, $P = Q_{\mathbf{0},\sigma}$. In that case, we have

$$\mathrm{KL}(Q_{\mathbf{W},\sigma}, P) \;=\; \frac{1}{2}\frac{\|\mathbf{W}\|^2}{\sigma^2}\,. \tag{4}$$

The next theorem provides an upper bound on the risk of the (deterministic) predictor $\mathbf{W}$ which depends on its empirical quadratic risk—not on the empirical risk of a stochastic (Gibbs) predictor. This new result was made possible by performing Gaussian integrals over functions of the quadratic loss and by observing that we can choose a value for $\sigma$ such that the noise of the empirical quadratic risk is cancelled by the noise of the true quadratic risk whenever $K_{\mathcal{X}}(x,x)$ is the same for all $x \in \mathcal{X}$.

**Theorem 4.** *Consider any input kernel $K_{\mathcal{X}}$ and any output kernel $K_{\mathcal{Y}}$ inducing finite-dimensional feature spaces. Suppose that $K_{\mathcal{X}}(x,x) = 1$ for all $x \in \mathcal{X}$. Let $D$ be any distribution on $\mathcal{X} \times \mathcal{Y}$. Then, for any prediction loss $L$ upper-bounded by the output kernel loss $L_{K_{\mathcal{Y}}}$, with probability at least $1-\delta$ over all training sets $S$ sampled according to $D^m$, we have, simultaneously for all predictors $\mathbf{W}$,*

$$\mathop{\mathbf{E}}_{(x,y)\sim D} L(y_{\mathbf{w}}(x), y) \leq \frac{5e}{e-1}\left[1 - e^{-\frac{1}{m}\left(2\sum_{i=1}^m \|Y(y_i)-\mathbf{W}X(x_i)\|^2 + \frac{9}{8}\|\mathbf{W}\|^2 + \ln\frac{1}{\delta}\right)}\right].$$

*Proof.* If we use Theorem 3 in the case of the quadratic loss with the proposed posterior $Q_{\mathbf{W},\sigma}$ and prior $P$, and if we use equations (4) and (5) and exploit the convexity of $-\ln x$, we then have that, with probability at least $1-\delta$,

$$-\ln\left(\mathop{\mathbf{E}}_{(x,y)\sim D}\;\mathop{\mathbf{E}}_{\mathbf{V}\sim Q_{\mathbf{W},\sigma}} e^{-2\|Y(y)-\mathbf{V}X(x)\|^2}\right) \leq \frac{1}{m}\left(\mathop{\mathbf{E}}_{\mathbf{V}\sim Q_{\mathbf{W},\sigma}} 2\sum_{i=1}^m \|Y(y_i)-\mathbf{V}X(x_i)\|^2 + \frac{\|\mathbf{W}\|^2}{2\sigma^2} + \ln\frac{1}{\delta}\right).$$

In the supplementary material we provide proofs of the following Gaussian integrals:

$$\mathop{\mathbf{E}}_{\mathbf{V}\sim Q_{\mathbf{W},\sigma}} \|Y(y) - \mathbf{V}X(x)\|^2 = \|Y(y) - \mathbf{W}X(x)\|^2 + \sigma^2 N_{\mathcal{Y}}\|X(x)\|^2\,. \tag{5}$$

$$\mathop{\mathbf{E}}_{\mathbf{V}\sim Q_{\mathbf{W},\sigma}} e^{-2\|Y(y)-\mathbf{V}X(x)\|^2} = \left[\frac{\sigma^{N_{\mathcal{X}}}}{\sqrt{1+4\sigma^2\|X(x)\|^2}}\right]^{N_{\mathcal{Y}}} e^{-\frac{2\|Y(y)-\mathbf{W}X(x)\|^2}{1+4\sigma^2\|X(x)\|^2}}. \tag{6}$$

Since, by hypothesis, $\|X(x)\|$ is a constant independent of $x$, with probability at least $1-\delta$,

$$\mathop{\mathbf{E}}_{(x,y)\sim D} 1 - e^{-2\frac{\|Y(y)-\mathbf{W}X(x)\|^2}{1+4\|X(x)\|^2\sigma^2}} \leq 1 - e^{-\xi - \frac{1}{m}\left(2\sum_{i=1}^m \|Y(y_i)-\mathbf{W}X(x_i)\|^2 + \frac{\|\mathbf{W}\|^2}{2\sigma^2} + \ln\frac{1}{\delta}\right)}, \tag{7}$$

where $\xi \stackrel{\text{def}}{=} N_{\mathcal{Y}}\left[2\|X(x)\|^2\sigma^2 + \ln\left(\frac{\sigma^{N_{\mathcal{X}}}}{\sqrt{1+4\|X(x)\|^2\sigma^2}}\right)\right]$.

For $\|X(x)\| = 1$, the value of $\sigma^2$ satisfying $\xi = 0$ is monotonously increasing with $N_{\mathcal{X}}$; going from $\sigma^2 = 0,6752...$ for $N_{\mathcal{X}} = 1$ to $\sigma^2 = 1$ when $N_{\mathcal{X}} \to \infty$. Consider Inequality (7) when $\xi = 0$. In that case $2/3 < \sigma^2 \leq 1$. Then its right-hand side can be upper-bounded by the same quantity but with $\sigma^2$ replaced by $2/3$, and its left-hand side can be lower-bounded by the same quantity but with $\sigma^2$ replaced by 1. The theorem then follows by applying Lemma 2 for $a = 5$. □

### 3.2. The Risk Bound Minimizer

The predictor $\mathbf{W}$ that minimizes the risk bound of Theorem 4 is the one that minimizes the multiple-output ridge regression objective $F_{rr}$, where

$$F_{rr}(\mathbf{W}) \;\stackrel{\text{def}}{=}\; C\sum_{i=1}^m \|Y(y_i) - \mathbf{W}X(x_i)\|^2 + \|\mathbf{W}\|^2\,,$$

for some value of $C > 0$. Note that $F_{rr}$ is exactly the objective to minimize that was proposed by Cortes et al. (2007). At optimality, the gradient of $F_{rr}$ must vanish. As shown by Cortes et al. (2007), the solution $\mathbf{W}^*$ is unique for finite $C$ and is given by

$$\mathbf{W}^* \;=\; \sum_{i=1}^m\sum_{j=1}^m Y(y_i)(\mathbf{K}_{\mathcal{X}} + \frac{1}{C}\mathbf{I})^{-1}_{i,j}X^\intercal(x_j)\,, \tag{8}$$

where $X^\intercal(x)$ denotes the transpose of vector $X(x)$, $\mathbf{K}_{\mathcal{X}}$ denotes the input kernel matrix, and $\mathbf{I}$ denotes the $m \times m$ identity matrix.

Since $\mathbf{W}^*$ is the minimizer of the $\ell_2$-regularized least squares $F_{rr}$, the convergence rates established by Caponnetto & De Vito (2007) also apply to $\mathbf{W}^*$.

## 4. PAC-Bayes with Sample-compression

Note that the predictor minimizing the ridge regression objective is a linear combination of simple predictors $Y(y_i)X^\intercal(x_j)$ that are identified by two training examples. Inspired by some recent work on PAC-Bayes sample-compression (Laviolette & Marchand, 2007; Germain et al., 2011), we want to establish a guarantee on the true risk for arbitrary linear combinations of these simple structured output predictors.

In contrast with Theorem 4, the obtained risk bound will be valid for feature spaces $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$ that are arbitrary RKHS (of possibly infinite dimensionality). For that purpose, let $X^\dagger(x)$ denote the dual of vector $X(x)$. The dual $X^\dagger(x)$ is a map from $\mathcal{H}_\mathcal{X}$ to $\mathbb{R}$ such that $\forall (x,x') \in \mathcal{X}^2$ we have $X^\dagger(x)X(x') = \langle X(x)|X(x')\rangle = K_\mathcal{X}(x,x')$. Thus, given a training set $S$ of $m$ examples, we consider predictors that can be written as

$$\mathbf{W} \;=\; \sum_{i=1}^{m}\sum_{j=1}^{m} Y(y_i) A_{i,j} X^\dagger(x_j)\,, \tag{9}$$

where $A_{i,j} \in \mathbb{R}\ \forall (i,j) \in \{1,\ldots,m\}^2$. In this case, the quadratic loss $\|Y(y) - \mathbf{W}X(x)\|^2$ is now given by

$$\left\| Y(y) - \sum_{i=1}^{m}\sum_{j=1}^{m} A_{i,j} K_\mathcal{X}(x_j,x) Y(y_i) \right\|^2 \stackrel{\text{def}}{=} R(\mathbf{A},x,y)\,. \tag{10}$$

To connect with PAC-Bayes sample-compression, let us write $\mathbf{A}$ in terms of a distribution $\mathbf{q}$ over $2m^2$ predictors. For this purpose, let $q^+_{i,j} \geq 0$ be the weight on predictor $Y(y_i)X^\dagger(x_j)$ and let $q^-_{i,j} \geq 0$ be the weight on the opposite predictor $-Y(y_i)X^\dagger(x_j)$ such that

$$\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{s\in\{-1,+1\}} q^s_{i,j} = 1\,.$$

Now, w.l.o.g., for all $(i,j)$, let $A_{i,j} = \kappa\cdot(q^+_{i,j} - q^-_{i,j})$ for some $\kappa > 0$. For notational brevity, let $R(\mathbf{q},x,y)$ be the quadratic loss obtained from $R(\mathbf{A},x,y)$ when each $A_{i,j}$ is replaced by $\kappa\cdot(q^+_{i,j} - q^-_{i,j})$. In addition, let $\mathcal{I} \stackrel{\text{def}}{=} \{1,\ldots,m\}^2$ denote the set of all pairs of indices and let $\mathcal{W} \stackrel{\text{def}}{=} \{-1,+1\}$. We then have

$$R(\mathbf{q},x,y) \;=\; \sum_{\mathbf{i}\in\mathcal{I}}\sum_{\mathbf{j}\in\mathcal{I}}\sum_{s\in\mathcal{W}}\sum_{t\in\mathcal{W}} q^s_\mathbf{i} q^t_\mathbf{j} \ell^{s,t}_{\mathbf{i},\mathbf{j}}(x,y)\,,$$

where, for $\mathbf{i} = (i,i')$ and $\mathbf{j} = (j,j')$, we have

$$\begin{aligned}\ell^{s,t}_{\mathbf{i},\mathbf{j}}(x,y) \stackrel{\text{def}}{=} \langle Y(y) - \kappa s Y(y_i) K_\mathcal{X}(x_{i'},x)| \\ Y(y) - \kappa t Y(y_j) K_\mathcal{X}(x_{j'},x)\rangle\,.\end{aligned} \tag{11}$$

An upper bound on $R(\mathbf{q}) \stackrel{\text{def}}{=} \mathbf{E}_{(x,y)\sim D} R(\mathbf{q},x,y)$ also provides an upper bound on the prediction risk $\mathbf{E}_{(x,y)\sim D} L(y_\mathbf{w}(x),y)$ since, by Lemma 1, we have $L(y_\mathbf{w}(x),y) \leq 4R(\mathbf{q},x,y)$ whenever $A_{i,j}$ is replaced by $\kappa\cdot(q^+_{i,j} - q^-_{i,j})$ in Equation (9) and whenever the $L$ is upper-bounded by $L_{K_\mathcal{Y}}$. Our goal is thus to find a tight upper bound on $R(\mathbf{q})$ and then design an algorithm that finds $\mathbf{q}$ (hence, the predictor $\mathbf{W}$) that minimizes this upper bound.

### 4.1. The Risk Bound

The proposed risk bound follows from PAC-Bayes theory and depends on how far is the posterior distribution $\mathbf{q}$ from a prior $\mathbf{p}$. For $\mathbf{p}$, we choose the uniform distribution over $\mathcal{I} \stackrel{\text{def}}{=} \{1,\ldots,2m\}^2$ so that $p^s_\mathbf{i} = 1/(2m^2)\ \forall(\mathbf{i},s) \in \mathcal{I}\times\mathcal{W}$, where $\mathcal{W} \stackrel{\text{def}}{=} \{-1,+1\}$. The posterior $\mathbf{q}$ is chosen to be *quasi-uniform*. By this we mean that for all $\mathbf{i} \in \mathcal{I}$ we have $q^+_\mathbf{i} + q^-_\mathbf{i} = 1/m^2$. In that case, each $q^s_\mathbf{i} \in [0, 1/m^2]$ and, consequently, the KL-divergence $\mathrm{KL}(\mathbf{q},\mathbf{p})$ is always at most $\ln 2$. Such a small upper bound on $\mathrm{KL}(\mathbf{q},\mathbf{p})$ contributes significantly at reducing the risk bound closer to the empirical risk $R(\mathbf{q},S) \stackrel{\text{def}}{=} (1/m)\sum_{i=1}^m R(\mathbf{q},x_i,y_i)$. Moreover, restricting $\mathbf{q}$ to quasi-uniform distributions does not restrict the class of predictors considered by the learner. Indeed, for any predictor $\mathbf{W}$ described by some matrix $\mathbf{A}$ in Equation (9), there exists $\kappa > 0$ and a quasi-uniform $\mathbf{q}$ such that $A_{i,j} = \kappa\cdot(q^+_{i,j} - q^-_{i,j})$.

**Theorem 5.** *Let $a \leq \ell^{s,t}_{\mathbf{i},\mathbf{j}}(x,y) \leq b\ \forall(x,y)\in\mathcal{X}\times\mathcal{Y}$, $\forall(s,t)\in\mathcal{W}^2$, $\forall(\mathbf{i},\mathbf{j})\in\mathcal{I}^2$, and for some interval $[a,b]$. Let $D$ be any distribution on $\mathcal{X}\times\mathcal{Y}$. Let $m \geq 8$. Then, with probability at least $1-\delta$ over all training sets $S$ sampled according to $D^m$, we have, simultaneously for all quasi-uniform distributions $\mathbf{q}$ on $\mathcal{I}$,*

$$R(\mathbf{q}) \;\leq R(\mathbf{q},S) + \sqrt{\frac{b-a}{2(m-4)}\left[20 + \ln\left(\frac{8\sqrt{m}}{\delta}\right)\right]}\,.$$

*Proof.* Given the uniform prior $\mathbf{p}$, consider the Laplace transform

$$\mathcal{L}_\mathbf{p} \stackrel{\text{def}}{=} \sum_{\mathbf{i}\in\mathcal{I}}\sum_{\mathbf{j}\in\mathcal{I}}\sum_{s\in\mathcal{W}}\sum_{t\in\mathcal{W}} p^s_\mathbf{i} p^t_\mathbf{j} e^{2(m-4)\left(\frac{\ell^{s,t}_{\mathbf{i},\mathbf{j}}(S)-a}{b-a} - \frac{\ell^{s,t}_{\mathbf{i},\mathbf{j}}-a}{b-a}\right)^2},$$

where $\ell^{s,t}_{\mathbf{i},\mathbf{j}} \stackrel{\text{def}}{=} \mathbf{E}_{(x,y)\sim D}\ell^{s,t}_{\mathbf{i},\mathbf{j}}(x,y)$ and $\ell^{s,t}_{\mathbf{i},\mathbf{j}}(S) \stackrel{\text{def}}{=} (1/m)\sum_{i=1}^m \ell^{s,t}_{\mathbf{i},\mathbf{j}}(x_i,y_i)$. Note that $\ell^{s,t}_{\mathbf{i},\mathbf{j}}(S)$ is a biased estimate of $\ell^{s,t}_{\mathbf{i},\mathbf{j}}$ since it considers the loss on examples that are used for the predictors described by $(\mathbf{i},s)$ and $(\mathbf{j},t)$. To obtain an unbiased estimator, let $S_{\mathbf{i},\mathbf{j}} \stackrel{\text{def}}{=} \{(x_i,y_i)\in S : i \notin \mathbf{i}\cup\mathbf{j}\}$ and let $m_{\mathbf{i},\mathbf{j}} \stackrel{\text{def}}{=} |S_{\mathbf{i},\mathbf{j}}|$. Then, let[3] $\ell^{s,t}_{\mathbf{i},\mathbf{j}}(S_{\mathbf{i},\mathbf{j}}) \stackrel{\text{def}}{=} \frac{1}{m_{\mathbf{i},\mathbf{j}}}\sum_{k=1}^m I((x_k,y_k)\in S_{\mathbf{i},\mathbf{j}})\ell^{s,t}_{\mathbf{i},\mathbf{j}}(x_k,y_k)$ be our unbiased estimator of $\ell^{s,t}_{\mathbf{i},\mathbf{j}}$. It is then straightforward to show that $\ell^{s,t}_{\mathbf{i},\mathbf{j}}(S_{\mathbf{i},\mathbf{j}}) - 4(b-a)/m \leq \ell^{s,t}_{\mathbf{i},\mathbf{j}}(S) \leq$

[3] Here $I(a) = 1$ if predicate $a$ is true and $I(a) = 0$ otherwise.

$\ell^{s,t}_{\mathbf{i},\mathbf{j}}(S_{\mathbf{i},\mathbf{j}}) + 4(b-a)/m$ and, consequently, for $m \geq 8$

$$\left(\frac{\ell^{s,t}_{\mathbf{i},\mathbf{j}}(S) - a}{b-a} - \frac{\ell^{s,t}_{\mathbf{i},\mathbf{j}} - a}{b-a}\right)^2 \leq \left(\frac{\ell^{s,t}_{\mathbf{i},\mathbf{j}}(S_{\mathbf{i},\mathbf{j}}) - a}{b-a} - \frac{\ell^{s,t}_{\mathbf{i},\mathbf{j}} - a}{b-a}\right)^2 + \frac{10}{m}\,. \quad (12)$$

Now, if we use $2(q-p)^2 \leq \mathrm{kl}(q,p) \stackrel{\text{def}}{=} q\ln(q/p) + (1-q)\ln[(1-q)/(1-p)]$, we obtain

$$\begin{aligned}\mathcal{L}_{\mathbf{p}} &\leq \mathop{\mathbf{E}}_{S\sim D^m} \sum_{\mathbf{i}\in\mathcal{I}}\sum_{\mathbf{j}\in\mathcal{I}}\sum_{s\in\mathcal{W}}\sum_{t\in\mathcal{W}} p^s_{\mathbf{i}} p^t_{\mathbf{j}} e^{\frac{20}{m}(m-4)+m_{\mathbf{i},\mathbf{j}}\mathrm{kl}\left(\frac{\ell^{s,t}_{\mathbf{i},\mathbf{j}}(S_{\mathbf{i},\mathbf{j}})-a}{b-a}, \frac{\ell^{s,t}_{\mathbf{i},\mathbf{j}}-a}{b-a}\right)} \\ &= e^{\frac{20}{m}(m-4)} \sum_{\mathbf{i}\in\mathcal{I}}\sum_{\mathbf{j}\in\mathcal{I}}\sum_{s\in\mathcal{W}}\sum_{t\in\mathcal{W}} p^s_{\mathbf{i}} p^t_{\mathbf{j}} \left(\prod_{i\in\mathbf{i}\cup\mathbf{j}} \mathop{\mathbf{E}}_{(x_i,y_i)\sim D}\right) \\ &\qquad \mathop{\mathbf{E}}_{S_{\mathbf{i},\mathbf{j}}\sim D^{m_{\mathbf{i},\mathbf{j}}}} e^{m_{\mathbf{i},\mathbf{j}}\mathrm{kl}\left(\frac{\ell^{s,t}_{\mathbf{i},\mathbf{j}}(S_{\mathbf{i},\mathbf{j}})-a}{b-a}, \frac{\ell^{s,t}_{\mathbf{i},\mathbf{j}}-a}{b-a}\right)}\end{aligned}$$

Since $S_{\mathbf{i},\mathbf{j}}$ is the arithmetic mean of $m_{\mathbf{i},\mathbf{j}}$ iid random variables, the lemma of Maurer (2004) tells us that the last expectation (over $S_{\mathbf{i},\mathbf{j}}$) is at most $2\sqrt{m_{\mathbf{i},\mathbf{j}}}$ and, consequently, $\mathcal{L}_{\mathbf{p}} \leq 2\sqrt{m}\exp(20(m-4)/m)$. Since $\mathcal{L}_{\mathbf{p}}$ is the expectation (over $S$) of a positive random variable, we can use Markov's inequality which states that, with probability of at least $1-\delta$ over the random draws of $S$, we have

$$\begin{aligned}&\ln\left(\sum_{\mathbf{i}\in\mathcal{I}}\sum_{\mathbf{j}\in\mathcal{I}}\sum_{s\in\mathcal{W}}\sum_{t\in\mathcal{W}} p^s_{\mathbf{i}} p^t_{\mathbf{j}} e^{2(m-4)\left(\frac{\ell^{s,t}_{\mathbf{i},\mathbf{j}}(S)-a}{b-a} - \frac{\ell^{s,t}_{\mathbf{i},\mathbf{j}}-a}{b-a}\right)^2}\right) \\ &\leq \frac{20}{m}(m-4) + \ln\left(\frac{2\sqrt{m}}{\delta}\right)\,.\end{aligned}$$

By turning the expectation over $\mathbf{p}^2$ into an expectation over $\mathbf{q}^2$, and by using Jensen's inequality on the concavity of the logarithm, the last inequality implies that we have

$$\begin{aligned}&\sum_{\mathbf{i}\in\mathcal{I}}\sum_{\mathbf{j}\in\mathcal{I}}\sum_{s\in\mathcal{W}}\sum_{t\in\mathcal{W}} q^s_{\mathbf{i}} q^t_{\mathbf{j}} \left(\frac{\ell^{s,t}_{\mathbf{i},\mathbf{j}}(S)-a}{b-a} - \frac{\ell^{s,t}_{\mathbf{i},\mathbf{j}}-a}{b-a}\right)^2 \\ &\leq \frac{1}{2(m-4)}\left(\mathrm{KL}(\mathbf{q}^2,\mathbf{p}^2) + 20 + \ln\frac{2\sqrt{m}}{\delta}\right)\,,\end{aligned}$$

for all $\mathbf{q}$. The theorem then follows by using Jensen's inequality on the convexity of $(q-p)^2$ and by using $\mathrm{KL}(\mathbf{q}^2,\mathbf{p}^2) = 2\mathrm{KL}(\mathbf{q},\mathbf{p}) \leq 2\ln 2$ for quasi-uniform posteriors. □

Hence, for quasi-uniform posteriors $\mathbf{q}$, the upper bound on $R(\mathbf{q})$ is very close to $R(\mathbf{q},S)$ whenever $(b-a) \ll m$. From Equation (11), we can see that $(b-a)$ is at most $2B_{\mathcal{Y}}(1+\kappa B_{\mathcal{X}})^2$ when $|K_{\mathcal{X}}(x,x')| \leq B_{\mathcal{X}}\ \forall (x,x') \in \mathcal{X}^2$ and $|K_{\mathcal{Y}}(y,y')| \leq B_{\mathcal{Y}}\ \forall (y,y') \in \mathcal{Y}^2$.

### 4.2. The Risk Bound Minimizer

The posterior $\mathbf{q}$ that minimizes the upper bound of Theorem 5 is the posterior minimizing $R(\mathbf{q},S)$ under the constraint that $\mathbf{q}$ is quasi-uniform. In that case, each $q^s_{i,j} \in [0, 1/m^2]$. Since $A_{i,j} = \kappa\cdot(q^+_{i,j} - q^-_{i,j})$, the quasi-uniform constraint on $\mathbf{q}$ implies that $|A_{i,j}| \leq C$ for all $(i,j) \in \mathcal{I}$ and for some $C > 0$. Instead of handling these $m^2$ constraints, it is computationally much cheaper to replace them by the single $\ell_2$ constraint $\sum_{(i,j)\in\mathcal{I}} A^2_{i,j} \leq R^2$ for some $R > 0$. Note that we have $|A_{i,j}| \leq R$ for all $(i,j)$ whenever this $\ell_2$ constraint is satisfied. Hence, given any $R > 0$, let us solve

$$\begin{aligned}&\min_{\mathbf{A}} R(\mathbf{A},S) \stackrel{\text{def}}{=} \frac{1}{m}\sum_{i=1}^m R(\mathbf{A},x_i,y_i) \\ &\text{s.t.} \sum_{i=1}^m\sum_{j=1}^m A_{i,j} \leq R^2 \stackrel{\text{def}}{=} m^2\rho^2\,.\end{aligned} \quad (13)$$

**Theorem 6.** *Let $\mathcal{A}^*$ denote the set of solutions of problem (13). Let $\mathbf{K}_{\mathcal{X}}$ and $\mathbf{K}_{\mathcal{Y}}$ denote, respectively, the input and output kernel matrices. Let $v_1,\ldots,v_m$ and $\lambda_1,\ldots,\lambda_m$ denote, respectively, the eigenvectors and eigenvalues of $\mathbf{K}_{\mathcal{X}}$. Let $u_1,\ldots,u_m$ and $\delta_1,\ldots,\delta_m$ denote, respectively, the eigenvectors and eigenvalues of $\mathbf{K}_{\mathcal{Y}}$. Let $\mathcal{J} \stackrel{\text{def}}{=} \{(i,j) \in \mathcal{I} : \delta_i\lambda_j > 0\}$. Then $\sum_{i=1}^m\sum_{j=1}^m \gamma_{i,j} u_i v_j^\intercal \in \mathcal{A}^*$, where $\gamma_{i,j}$ is given by*

$$\text{if } \sum_{(i,j)\in\mathcal{J}} \frac{(u_i^\intercal v_j)^2}{\lambda_j^2} \leq R^2 \text{ then } \gamma_{i,j} = \begin{cases} 0 & \text{if } \delta_i\lambda_j = 0 \\ \frac{u_i^\intercal v_j}{\lambda_j} & \text{if } \delta_i\lambda_j > 0 \end{cases}$$

$$\text{otherwise } \gamma_{i,j} = \frac{\delta_i\lambda_j(u_i^\intercal v_j)}{\delta_i\lambda_j^2 + m\beta}\,,$$

*where $\beta > 0$ is the solution of $\sum_{i=1}^m\sum_{j=1}^m \frac{\delta_i^2\lambda_j^2(u_i^\intercal v_j)^2}{(\delta_i\lambda_j^2+m\beta)^2} = R^2$.*

*Proof.* Let $L(\mathbf{A},\beta) \stackrel{\text{def}}{=} R(\mathbf{A},S) + \beta\left(\|\mathbf{A}\|^2 - R^2\right)$. Convex optimisation theory tells us that if there exists $\beta \geq 0$ and $\mathbf{A} : \|\mathbf{A}\|^2 \leq R^2$, that satisfies $\partial L/\partial\mathbf{A} = 0$ and $\beta\cdot\left(\|\mathbf{A}\|^2 - R^2\right) = 0$, then $\mathbf{A} \in \mathcal{A}^*$. Here we have

$$\frac{\partial L}{\partial \mathbf{A}} = \frac{2}{m}\mathbf{K}_{\mathcal{Y}}(\mathbf{A}\mathbf{K}_{\mathcal{X}} - \mathbf{I})\mathbf{K}_{\mathcal{X}} + 2\beta\mathbf{A} = 0\,. \quad (14)$$

Since $\mathbf{K}_{\mathcal{X}}$ and $\mathbf{K}_{\mathcal{Y}}$ are symmetric positive semi-definite $m\times m$ matrices, their eigenvalues are all non-negative and their eigenvectors constitute an orthonormal basis

of $\mathbb{R}^m$. Thus, $\{u_i v_j^\intercal\}_{(i,j)\in\mathcal{I}}$ is an orthonormal basis of $\mathbb{R}^{m^2}$. Consequently, w.l.o.g., any $m \times m$ matrix $\mathbf{A}$ can be written as $\mathbf{A} = \sum_{i=1}^m \sum_{j=1}^m \gamma_{i,j} u_i v_j^\intercal$ for some values of $\gamma_{i,j}$. Hence, we have $\|\mathbf{A}\|^2 = \sum_{i=1}^m \sum_{j=1}^m \gamma_{i,j}^2$ and Equation (14) becomes $\gamma_{i,j}(\delta_i \lambda_j^2 + m\beta) = \delta_i \lambda_j (u_i^\intercal v_j)$. When $\beta = 0$, that equation is solved for $\gamma_{i,j} = 0$ when $\delta_i \lambda_j = 0$ and $\gamma_{i,j} = (u_i^\intercal v_j)/\lambda_j$ when $\delta_i \lambda_j > 0$ (a solution where the $\ell_2$ constraint is not active[4]). When $\beta > 0$, that equation is solved for $\gamma_{i,j} = (\delta_i \lambda_j (u_i^\intercal v_j))/(\delta_i \lambda_j^2 + m\beta)$ and, in that case, we have $\|\mathbf{A}\|^2 = R^2$ with a unique non-zero solution for $\beta$. □

Note that the eigenvectors and eigenvalues of $\mathbf{K}_\mathcal{X}$ and $\mathbf{K}_\mathcal{Y}$ can be obtained from their singular value decompositions in $O(m^3)$ time. The solution for $\beta$ can be obtained with Newton's method requiring $\Theta(m^2)$ time for each iteration. Finally, we can obtain $\mathbf{A}$ in $O(m^3)$ by using $\mathbf{A} = \mathbf{u}\boldsymbol{\gamma}\mathbf{v}^\intercal$ where $\mathbf{u}$ and $\mathbf{v}$ are matrices obtained by concatenating the the column eigenvectors of $\mathbf{K}_\mathcal{Y}$ and $\mathbf{K}_\mathcal{X}$ respectively and where $\boldsymbol{\gamma}$ denotes the matrix of $\gamma_{i,j}$ values. Hence, the proposed solution of (13) is reached in $O(m^3)$ time whenever Newton's method requires at most $O(m)$ iterations.

## 5. Empirical Results

We have compared the solution given by Equation (8) (Structured Output by Ridge Regression–SORR) with the one given by Theorem 6 (Structured Output by Sample-Compression–SOSC) on the word recognition task studied by Taskar et al. (2004); Cortes et al. (2007) and the enzyme classification task studied by Rousu et al. (2006). All hyper parameters ($C$, $\rho$, and kernel parameters) were selected with 10-fold cross-validation (CV) on the training sets where we have relied on the pre-images (using Equation (1)) for that purpose only.

The word recognition task consists of predicting the correct word (a sequence of letters) associated to a manuscript picture of the same word. The metrics used for this data set is usually the 0/1-risk (the fraction of errors on words) and the letter risk (the fraction of errors on letters). Hence, following Equation (2), we have used the Dirac kernel ($K_\mathcal{Y}(y, y') = I(y = y')$) and the Hamming kernel (which is given by the length of the largest string minus the Hamming distance between the two strings). The polynomial kernel of degree $d$ was used for the input kernel. All experiments were done using the protocol described in Taskar et al. (2004); Cortes et al. (2007). According to Cortes et al. (2007), SORR achieved better performance than structural SVM and max-margin Markov networks. Our empirical results are shown in Table (1). The error bars are the standard deviation of the corresponding risk over the different CV folds given by Taskar et al. (2004). Clearly, SORR and SOSC achieved very similar generalization performance (with overlapping error bars) on both the 0/1-risk and the letter risk.

The enzymes hierarchical classification task consists of predicting a path in a enzyme classification scheme used by biologist to classify amino acid sequences of enzymatic proteins. As in Rousu et al. (2006), the 4-gram kernel was used in the input space. Focussing on the hierarchical risk (the length of the incorrect sub-path from the root to the enzyme leaf) as the most natural metric for this data set, we have used the hierarchical kernel of (Jacob et al., 2008) (given by the length of the common sub-path between two paths) on the output space. All experiments were done using the protocol described in Rousu et al. (2006) and our empirical results are shown in Table (2). We have also included the results obtained by Rousu et al. (2006) for $H$–$M^3$–$\ell_{\tilde{H}}$ and $H$–$M^3$–$\ell_\Delta$, which are variants of the max-margin Markov networks. In the case of the 0/1 risk (the fraction of misclassification errors), we have computed the 90% confidence intervals from the binomial tail inversion method of Langford (2005). From Table (2), we see that the 0/1 risk differences between all algorithms are significant (at 0.9 confidence level), with SORR being the best algorithm. For the hierarchical risk, note that from the central limit theorem, the standard deviation of this metric is given by $\sigma/\sqrt{n}$ for a testing set of $n = 1755$ examples when the hierarchical loss variance is $\sigma^2$. Since $\sigma \leq 3$ for the hierarchy of 4 levels, the hierarchical risk differences between all algorithms appear to be significant, with $H$–$M^3$–$\ell_\Delta$ being the best algorithm.

## 6. Conclusion

We have shown that the quadratic regression loss is a convex surrogate of the prediction loss when the prediction loss is upper-bounded by the output kernel loss. We have provided two PAC-Bayes upper bounds of the structured prediction risk that depend on the empirical quadratic risk of the deterministic predictor. The second bound, based on the PAC-Bayes sample-compression approach, is more general than the first bound as it holds for feature spaces that are arbitrary RKHS. The minimizer of the first bound, SORR, turns out to be the predictor proposed by Cortes et al. (2007) while the minimizer of the second bound, SOSC, is a

[4] This is the smallest $\ell_2$ norm solution of the unconstrained problem. The Moore-Penrose pseudo-inverse of $\mathbf{K}_\mathcal{X}$ is also a solution of the unconstrained problem since, in that case, it suffice for $\mathbf{A}$ to satisfy $\mathbf{A}\mathbf{K}_\mathcal{X}^2 = \mathbf{K}_\mathcal{X}$.

*Table 1.* Empirical results on the word recognition task

| | Dirac kernel | | Hamming kernel | |
|---|---|---|---|---|
| | SORR | SOSC | SORR | SOSC |
| 0/1 risk | 0.0539 $\pm$.0087 | 0.0525 $\pm$.0085 | 0.0871 $\pm$.0078 | 0.0871 $\pm$.0078 |
| Letter risk | 0.0294 $\pm$.0067 | 0.0285 $\pm$.0062 | 0.0370 $\pm$.0047 | 0.0367 $\pm$.0049 |

*Table 2.* Empirical results on the enzyme hierarchical classification task

| | $H$–$M^3$–$\ell_\Delta$ | $H$–$M^3$–$\ell_{\tilde{H}}$ | SORR | SOSC |
|---|---|---|---|---|
| 0/1 risk | 0.957 [0.949, 0.965] | 0.855 [0.840, 0.869] | 0.640 [0.621, 0.659] | 0.684 [0.666, 0.702] |
| Hierarchical risk | 1.2 | 2.50 | 1.71 | 1.84 |
| F1 Score | 0.6330 | 0.5340 | 0.5813 | 0.5569 |

predictor that has never been proposed so far. Both predictors have been compared on practical tasks. Finally, although it would be time-consuming, it would be interesting to see if we can improve SOSC by using the full set of $m^2$ constraints instead of the single $\ell_2$ constraint used in optimization problem (13).

## References


Baldassarre, Luca, Rosasco, Lorenzo, Barla, Annalisa, and Verri, Alessandro. Multi-output learning via spectral filtering. *Machine Learning*, 87:259–301, 2012.

Caponnetto, A. and De Vito, E. Optimal rates for regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Cortes, Corinna, Mohri, Mehryar, and Weston, Jason. A general regression framework for learning string-to-string mappings. In Bakır, Gökhan, Hofmann, Thomas, Schölkopf, Bernhard, Smola, Alexander J., Taskar, Ben, and Vishwanathan, S. V. N. (eds.), *Predicting Structured Data*, chapter 8, pp. 143–168. MIT Press, Cambridge, MA, 2007.

Gärtner, Thomas and Vembu, Shankar. On structured output training: hard cases and an efficient alternative. *Machine Learning*, 79:227–242, 2009.

Germain, Pascal, Lacoste, Alexandre, Laviolette, François, Marchand, Mario, and Shanian, Sara. A PAC-Bayes sample-compression approach to kernel methods. In Getoor, Lise and Scheffer, Tobias (eds.), *Proceedings of the 28th International Conference on Machine Learning*, ICML '11, pp. 297–304, New York, NY, USA, June 2011. ACM.

Jacob, Laurent, Hoffmann, Brice, Stoven, Veronique, and Vert, Jean-Philippe. Virtual screening of gpcrs: An in silico chemogenomics approach. *BMC Bioinformatics*, 9(1):363, 2008.

Langford, John. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6:273–306, 2005.

Laviolette, François and Marchand, Mario. PAC-Bayes risk bounds for stochastic averages and majority votes of sample-compressed classifiers. *Journal of Machine Learning Research*, 8:1461–1487, 2007.

Maurer, Andreas. A note on the PAC Bayesian theorem. *CoRR*, cs.LG/0411099, 2004.

McAllester, David. PAC-Bayesian stochastic model selection. *Machine Learning*, 51:5–21, 2003.

McAllester, David. Generalization bounds and consistency for structured labeling. In Bakır, Gökhan, Hofmann, Thomas, Schölkopf, Bernhard, Smola, Alexander J., Taskar, Ben, and Vishwanathan, S. V. N. (eds.), *Predicting Structured Data*, chapter 11, pp. 247–261. MIT Press, Cambridge, MA, 2007.

Rousu, Juho, Saunders, Craig, Szedmak, Sandor, and Shawe-Taylor, John. Kernel-based learning of hierarchical multilabel classification models. *J. Mach. Learn. Res.*, 7:1601–1626, December 2006.

Taskar, Ben, Guestrin, Carlos, and Koller, Daphne. Max-margin markov networks. In Thrun, Sebastian, Saul, Lawrence, and Schölkopf, Bernhard (eds.), *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

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

Zhang, Tong. Information theoretical upper and lower bounds for statistical estimation. *IEEE Transaction on Information Theory*, 52:1307–1321, 2006.