Transfer Learning with Adaptive Regularizers

Ulrich Rückert[1] and Marius Kloft[2,*]

[1] University of California, Berkeley, USA
rueckert@eecs.berkeley.edu
[2] Machine Learning Laboratory, Technische Universität Berlin, Germany
mkloft@mail.tu-berlin.de

Abstract. The success of regularized risk minimization approaches to classification with linear models depends crucially on the selection of a regularization term that matches with the learning task at hand. If the necessary domain expertise is rare or hard to formalize, it may be difficult to find a good regularizer. On the other hand, if plenty of related or similar data is available, it is a natural approach to adjust the regularizer for the new learning problem based on the characteristics of the related data. In this paper, we study the problem of obtaining good parameter values for a ℓ_2-style regularizer with feature weights. We analytically investigate a moment-based method to obtain good values and give uniform convergence bounds for the prediction error on the target learning task. An empirical study shows that the approach can improve predictive accuracy considerably in the application domain of text classification.

Keywords: transfer learning, multitask learning, regularization.

1 Introduction

Many approaches to classification optimize the sum of a data-dependent risk functional and a data-independent regularizer. Modern machine learning applications often use such methods on complex data objects, which can be described by large amounts of features. Since one has many more features than training instances in such settings, it is important to choose good regularization. Ideally, one would want to choose a regularizer that matches well with the unknown data-generating distribution. Finding such a good regularizer can either be done based on the available data (which might lead to overfitting) or based on domain expertise or meta knowledge, which is often rare or requires significant amount of work. Modern automated data processing systems, on the other hand, have led to the availability of vast amounts of potentially related data, which might help in selecting a good regularizer.

In this paper we address the problem of automatically adapting the regularizer for a *target learning problem*, if one has access to a (possibly large) number of related *source learning tasks*. To do so, we choose a highly parameterized regularizer for the target learning problem and try to obtain good settings for

[*] A part of the work was fone while MK was with University of California, Berkeley.

D. Gunopulos et al. (Eds.): ECML PKDD 2011, Part III, LNAI 6913, pp. 65–80, 2011.

the parameters from the source data sets. We frame the problem theoretically using a frequentist hierarchical model, similar to the ones by Baxter [4] and Ando and Zhang [1]. However, instead of bounding the average prediction error over all learning tasks (*multitask learning*), we give bounds only for the prediction error on the target task (*inductive transfer*). We also do not make any fixed assumption about how the source and target learning tasks are related, such as transformation-based relatedness [5,12] or preprocessing-based relatedness [11,17]. Instead, we start from a worst-case analysis and only make the assumption that source and target learning tasks are drawn i.i.d. from a fixed but unknown distribution. We then show how one can add additional assumptions to improve those worst case bounds in particular situations. The resulting uniform convergence bounds relate the success of the regularization parameters obtained from the source data sets to the number of source data sets and quantifies the trade-off between estimation and approximation errors.

We evaluate our approach empirically in the application domains of text classification and predicting molecular structure-activity relationships. The results indicate that our approach works at least as well as a regular SVM and in a few cases yields drastic gains in prediction accuracy up to 19% over approaches that do not transfer information from the source tasks.

Our main contributions can be summarized as follows:

- We present a novel approach to transfer learning, for which we show upper bounds on the generalization error on the target task in a hierarchical i.i.d. setup.
- We show how our bound can be further tightened when distribution-dependent information is available. We demonstrate that the so-obtained generalization bounds can be strictly tighter than standard results.
- We show that our approach works well in the domain of text classification, yielding gains in accuracy of up to 19% compared to a regular SVM and the approach of Evgeniou & Pontil [8].

Finally, we would like to mention that our method is easy-to-use since one just needs a regular SVM implementation. We thus believe that our method could be useful to other researchers for exploring new application domains in which transfer learning might be helpful. Our implementation will be made available with the final version of this paper.

2 Regularization Adaptation with Transfer Learning

Let us now describe the setting more formally. We are given a space of data objects $\mathcal{X}$ that are embedded in an Euclidean feature space, i.e. $\mathcal{X} \simeq \mathbb{R}^m$, and a set of binary class labels $\mathcal{Y} = \{-1, +1\}$. We assume that nature poses a sequence of source learning tasks $T^1, \ldots, T^p$ and one target learning task T°. We assume that all these learning tasks are drawn i.i.d. from a fixed but unknown distribution $\mathcal{T}$. The goal is to find a good classifier for the target learning problem T°. For each learning task $T^i = (X^i, Y^i)$ we are given a sample of training data

$X^i = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_{n^i}\}$, and labels $Y^i = \{y_1, ..., y_{n^i}\}$, drawn i.i.d. from some unknown distribution, which in turn is drawn from $\mathcal{T}$. For ease of notation, we assume $n := n^1 = \ldots = n^p$ in the following.

As we are not interested in the actual data-generating distributions for the source tasks, but only the distribution of the observed data, we will not distinguish between "true" and "empirical" source distributions. Instead we simply assume that each source task distribution P^i is defined with regard to the sample (X^i, Y^i). This means we write $\Pr_{P^i}$ to denote the sample probability measure $\Pr_{P^i}(x, y) := \frac{1}{n^i}\sum_{(x^i,y^i)\in(X^i,Y^i)} I[x = x^i \wedge y = y^i]$ and $\mathbb{E}_{(x,y)\sim P^i}$ for the sample expectation $\mathbb{E}_{(x,y)\sim P^i}[f(x, y)] := \frac{1}{n^i}\sum_{(x^i,y^i)\in(X^i,Y^i)} f(x^i, y^i)$. For the target learning task T°, we follow the same convention, but assume that we have seen only a smaller fraction $n^\circ << n$ of all target examples. This means that the "true" probabilities $\Pr_{P^\circ}$ and expectations $\mathbb{E}_{P^\circ}$ are still defined with regard to the sample $X^\circ = \{x_1^\circ, \ldots, x_n^\circ\}$. However, for learning a classifier, we only have access to a smaller subset $\{x_1^\circ, \ldots, x_{n^\circ}^\circ\} \subset X^\circ$ of examples.

We denote by $\mathbb{E}$ and Pr the overall expectation and probability over the choice of both the learning tasks and a particular training sample, unless stated otherwise, while conditional expectations will be marked by a subscript; for example, $\mathbb{E}_{(\boldsymbol{x},y)\sim P^\circ}$ takes the expectation over the target data generating distribution P°, but is still a random variable with regard to the learning task generating distribution $\mathcal{T}$. In the cases where we take the overall probability, the random quantities usually only depend on the drawing of the target task P° from the data set generating distribution $\mathcal{T}$. Therefore, it is usually safe to assume $\mathbb{E} = \mathbb{E}_{\mathcal{T}}$ and $\Pr = \Pr_{\mathcal{T}}$ in the following.

For the source data sets $(X^1, Y^1), \ldots, (X^p, Y^p)$, we would like to find linear classifiers $\boldsymbol{w}^1, \ldots, \boldsymbol{w}^p \in \mathbb{R}^{m^i}$ whose loss $\mathrm{er}^i(\boldsymbol{w}^i)$ is as small as possible, while constraining $\|\boldsymbol{w}\|_2 \leq C$. For ease of notation, we set $C = 1$, but the following results also hold for other choices of C. Define

$$\forall i = 1, ..., p: \quad \boldsymbol{w}^i := \operatorname*{argmin}_{\boldsymbol{w}:\|\boldsymbol{w}\|_2\leq 1} \mathrm{er}^i(\boldsymbol{w}), \tag{1}$$

$$\text{where } \forall \boldsymbol{w}: \ \mathrm{er}^i(\boldsymbol{w}) := \frac{1}{n}\sum_{j=1}^{n} \ell(\boldsymbol{w}^\top \boldsymbol{x}_j^i y_j^i).$$

Here, $\ell : \mathbb{R} \to [0, 1]$ is a loss function measuring the quality of a prediction. Suppose the criterion has a unique solution $\boldsymbol{w}^i$. We can then view the $\boldsymbol{w}^1, \ldots, \boldsymbol{w}^p$ as a random sample of empirical risk minimizers.

However, our goal is to find a vector $\boldsymbol{w}$ that minimizes the expected error $\mathrm{er}^\circ(\boldsymbol{w})$ on the *target* task, while we can only observe the empirical error $\hat{\mathrm{er}}^\circ(\boldsymbol{w})$:

$$\forall \boldsymbol{w}: \quad \mathrm{er}^\circ(\boldsymbol{w}) := \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim P^\circ} \ell(\boldsymbol{w}^\top \boldsymbol{x} y) = \frac{1}{n}\sum_{j=1}^{n} \ell(\boldsymbol{w}^\top \boldsymbol{x}_j^\circ y_j^\circ), \tag{2}$$

$$\hat{\mathrm{er}}^\circ(\boldsymbol{w}) := \frac{1}{n^\circ}\sum_{j=1}^{n^\circ} \ell(\boldsymbol{w}^\top \boldsymbol{x}_j^\circ y_j^\circ).$$

To this aim, we could employ a standard approach such as (1). However, since we know that T° is drawn from the same distribution $\mathcal{T}$ as the $T^1, \ldots, T^p$, it would make sense to re-use some of the information in the source data sets for the selection of the target classifier. In the following we do so by using an adjustable regularization term, which is modeled on base of the observed source tasks:

Proposed Transfer Learning Approach

$$\hat{\boldsymbol{w}}^\circ := \underset{\boldsymbol{w} \in B_b}{\operatorname{argmin}} \ \hat{\text{er}}^\circ(\boldsymbol{w}), \tag{T}$$

where B_b is a regularizer depending on the source tasks. The main idea is that the new regularizer forces the classifier to be from a more restricted set B_b, whose size and form depends on the source learning tasks $T^1, \ldots, T^p$ and a scale parameter $b \in \mathbb{R}$.

More specifically, the B_b is designed to keep the favorable properties of ℓ_2 regularization, but to transfer information about the relevance of individual features or feature groups. A feature, which gets assigned large weights on most source data sets is likely to also be informative on the target data set. Thus, it makes sense to adjust the regularizer for the target learning task so that it encourages the assignment of large weights to informative features and to penalize the assignment of large weights to features which have not received considerable weights on the source learning tasks. Note that the actual sign of a weight is not important, as the assignment of $+1$ and -1 to individual class labels is arbitrary and may change between individual learning tasks. We therefore use the absolute values $|w_j|$, or, more generally, the qth moment $|w_j|^q$ to assess feature relevance. More formally, we define:

$$\hat{\boldsymbol{\mu}} := \frac{1}{p} \sum_{i=1}^{p} |\boldsymbol{w}^i|^q, \qquad \boldsymbol{\mu} := \mathbb{E}\left[|\boldsymbol{w}^1|^q\right] = \ldots = \mathbb{E}\left[|\boldsymbol{w}^p|^q\right], \tag{3}$$

Here, $w^q := (w_1^q, \ldots, w_m^q)$ is meant to be the elementwise power. Note the $\boldsymbol{w}^i$ are i.i.d. and hence the definition of $\boldsymbol{\mu}$ can be made on base of any of the $\boldsymbol{w}^i$. As explained above, the qth moment $\hat{\boldsymbol{\mu}}$ measures how much information each component of the $\boldsymbol{w}^i$s has about the class label on average. For example, large components $\hat{\mu}_j$ correspond to large absolute values $|w_j^i|$, and hence the jth feature is likely to be discriminative—it is thus suggestive to employ a regularizer that promotes features with large μ_j. To promote features that are likely to be discriminative, we employ the following regularizer:

Moment-based Regularizer

$$B_b := B_b(T^1, \ldots, T^M) = \left\{ \boldsymbol{w} \;\middle|\; \|\boldsymbol{w} \circ \hat{\boldsymbol{\mu}}^{-1}\| \leq b \right\}, \quad \text{where } b > 0 \tag{B}$$

Here, $\circ$ denotes the elementwise multiplication of vectors, and we employ the notation $\boldsymbol{w}^{-1} = (1/w_1, \ldots, 1/w_m)$ to denote the elementwise inverse. Note that $\|\boldsymbol{w} \circ \hat{\boldsymbol{\mu}}^{-1}\|$ is only a shorter way to write $\sqrt{\sum_{j=1}^{m} w_j^2/\hat{\mu}_j^2}$. Informally speaking, the regularizer (B) is an ℓ_2-norm regularizer, where the dimensions are scaled according to the moment of the corresponding features in the source data sets. Using this regularizer, we can state the proposed transfer approach as an easy three step procedure: First, obtain good weight vectors $\boldsymbol{w}^i$ on the source data sets, then compute the new regularizer B_b from the moments of the $\boldsymbol{w}^i$, and finally learn a target weight vector $\hat{\boldsymbol{w}}^\circ$ using B_b as regularizer. Note that the restriction to norm constraints is not a limitation since non-centered hypothesis classes are also subsumed by our analysis. This is because translations $\boldsymbol{w} \mapsto \boldsymbol{w}+\boldsymbol{t}$ cannot modify the Rademacher complexity by more than $\|\boldsymbol{t}\|_\infty/\sqrt{n}$ [3].

3 Theoretical Analysis

In this section we analyze the proposed transfer learning method theoretically in terms of upper bounds on the generalization error.

Theoretical performance measure. In order to theoretically measure the success of our approach, we compare its expected test error to one of the theoretically optimal linear classifier on the target data, i.e. we wish to obtain a bound of the form

$$\mathrm{er}^\circ(\hat{\boldsymbol{w}}^\circ) - \mathrm{er}^\circ(\boldsymbol{w}^*) \leq \text{bound},$$

$$\text{where } \boldsymbol{w}^* := \operatorname*{argmin}_{\boldsymbol{w}:\|\boldsymbol{w}\|\leq 1} \mathrm{er}^\circ(\boldsymbol{w}). \tag{4}$$

This bound compares the performance of our method to the one of the theoretically optimal vector $\boldsymbol{w}^*$. Of course, this quantity can not be observed, because the true underlying distribution is unknown. However, one can nevertheless obtain such an inequality by decomposing the above quantities into two terms as follows:

$$\begin{aligned} & \mathrm{er}^\circ(\hat{\boldsymbol{w}}^\circ) - \mathrm{er}^\circ(\boldsymbol{w}^*) \\ \leq\ & \mathrm{er}^\circ(\hat{\boldsymbol{w}}^\circ) - \hat{\mathrm{er}}^\circ(\hat{\boldsymbol{w}}^\circ) + \hat{\mathrm{er}}^\circ(\hat{\boldsymbol{w}}^\circ) \qquad (5) \\ & - \mathrm{er}^\circ(\boldsymbol{w}^\circ) + \mathrm{er}^\circ(\boldsymbol{w}^\circ) - \mathrm{er}^\circ(\boldsymbol{w}^*) \\ \leq\ & \underbrace{2 \sup_{\boldsymbol{w}\in B_b} \Big(|\mathrm{er}^\circ(\boldsymbol{w}) - \hat{\mathrm{er}}^\circ(\boldsymbol{w})|\Big)}_{\text{estimation error} \quad \mathrm{er}_e} + \underbrace{\mathrm{er}^\circ(\boldsymbol{w}^\circ) - \mathrm{er}^\circ(\boldsymbol{w}^*)}_{\text{approximation error} \quad \mathrm{er}_a}, \qquad (6) \end{aligned}$$

where we use the quantity $\boldsymbol{w}^\circ := \mathrm{argmin}_{\boldsymbol{w}\in B_b}\ \mathrm{er}^\circ(\boldsymbol{w})$ (this is the theoretical outcome of our approach (T) if we would optimize with regard to all n target examples instead of just the observed n° examples). To see that inequality (5) holds, note that $\hat{\mathrm{er}}^\circ(\hat{\boldsymbol{w}}^\circ) \leq \hat{\mathrm{er}}^\circ(\boldsymbol{w}^\circ)$. In the following, we address how to bound the estimation and approximation error separately.

3.1 Estimation Error

First, for the estimation error, we assume a fixed target regularizer B_b and only deal with the target data set. As explained in the error decomposition (6), it is sufficient to give a uniform convergence bound on the generalization error to bound the estimation error. To this aim, we give the following result:

Theorem 1. *Let the regularizer be as defined in* (B). *Suppose the loss* $\ell : \mathbb{R} \supset X \to [0,1]$ *is Lipschitz with constant* L*, and the data lies in the unit cube,* $\boldsymbol{x} \in [-1,1]^m$*. Then, the following holds with probability larger than* $1-\delta$*:*

$$\sup_{\boldsymbol{w} \in B_b} \left| \mathrm{er}^\circ(\boldsymbol{w}) - \hat{\mathrm{er}}^\circ(\boldsymbol{w}) \right| \leq \frac{2Lb}{\sqrt{n^\circ}} + 2\sqrt{\frac{2 \ln \frac{2}{\delta}}{n^\circ}} \; .$$

The proof uses the established techniques and is shown in the Appendix. From (6) immediately follows for the estimation error $\mathrm{er}_e \leq 4L\frac{b}{\sqrt{n^\circ}} + 4\sqrt{\frac{2\ln\frac{2}{\delta}}{n^\circ}}$.

3.2 Approximation Error

The following theorems give upper bounds of the approximation error of the proposed approach (T) with B_b defined in (B). We start by giving a worst-case upper bound which does not make any assumption about the dataset-generating distribution $\mathcal{T}$.

Theorem 2. *Let the regularizer be defined as in* (B). *Suppose the loss* $\ell : \mathbb{R} \supset X \to [0,1]$ *is Lipschitz with constant* L*, and the data lies in the unit cube,* $\boldsymbol{x} \in [-1,1]^m$*. Then, with probability greater than* $1-\delta$ *over the choice of the source data sets and with probability greater than* $1-\epsilon$ *over the choice of the target learning task,*[1] *the approximation error* $\mathrm{er}_a = \mathrm{er}^\circ(\boldsymbol{w}^\circ) - \mathrm{er}^\circ(\boldsymbol{w}^*)$ *has*

$$\mathrm{er}_a \leq \frac{Lm}{b^{\frac{1}{q}}} \left[\frac{1}{\epsilon} \left(\frac{q^q}{(q+1)^{q+1}} + b\sqrt{\frac{\ln \frac{m}{\delta}}{2p}} \right) \right]^{\frac{1}{q}} \tag{7}$$

Proof. We start the proof by noting that, since ℓ is Lipschitz with constant L, we know that $\ell(a) - \ell(b) \leq L|a-b|$ for all $a, b \in \mathbb{R}$. Thus we can use the Hölder inequality to bound the difference between the error of $\boldsymbol{w}^\circ$ and the error of $\boldsymbol{w}^*$ as follows:

$$\begin{aligned}
\mathrm{er}_a &= \mathrm{er}^\circ(\boldsymbol{w}^\circ) - \mathrm{er}^\circ(\boldsymbol{w}^*) \\
&= \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim P^\circ} \left[\ell(y\boldsymbol{w}^{\circ\top}\boldsymbol{x}) \right] - \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim P^\circ} \left[\ell(y\boldsymbol{w}^{*\top}\boldsymbol{x}) \right] \\
&\leq L \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim P^\circ} \left[\inf_{\boldsymbol{w}\in B_b} \left| \left(y\left(\boldsymbol{w} - \boldsymbol{w}^*\right)^\top \boldsymbol{x}\right) \right| \right] \\
&\leq L \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim P^\circ} \left[\|y\boldsymbol{x}\|_{\frac{q}{q-1}} \inf_{\boldsymbol{w}\in B_b} \|\boldsymbol{w} - \boldsymbol{w}^*\|_q \right] \\
&\leq Lm^{\frac{q-1}{q}} \inf_{\boldsymbol{w}\in B_b} \|\boldsymbol{w} - \boldsymbol{w}^*\|_q,
\end{aligned} \tag{8}$$

[1] In other words, with probability $(1-\epsilon)(1-\delta)$ over the combined distribution $\mathcal{T} \times P^\circ$.

where in the last step we exploit that the data lies in the unit cube, $\boldsymbol{x} \in [-1,1]^m$, and the labels are binary, $\boldsymbol{y} \in \{-1,1\}$. Note that $\inf_{\boldsymbol{w}\in B_b} \|\boldsymbol{w} - \boldsymbol{w}^*\|_q$ is a random variable because the optimal $\boldsymbol{w}^*$ depends on the draw of the target task T°. The above result (8) shows that in order to complete the proof, it suffices to show that with high probability (over the draw of the target task) $\inf_{\boldsymbol{w}\in B_b} \|\boldsymbol{w} - \boldsymbol{w}^*\|_q$ is smaller than the rightmost term in (7); i.e., we need: $\Pr\left[\inf_{\boldsymbol{w}\in B_b} \|\boldsymbol{w} - \boldsymbol{w}^*\|_q \geq t\right] \leq$ bound.

Before we proceed with this, we first need an auxiliary result: we show that the moment $\hat{\boldsymbol{\mu}}$ is concentrated around its expected value $\boldsymbol{\mu}$ (see definition in (3)). To this aim, consider the random variable $V_j := \mu_j - \hat{\mu}_j \overset{\text{Def.(3)}}{=} \mathbb{E}[w_j^{1\,q}] - \frac{1}{p}\sum_{i=1}^p w_j^{i\,q}$. Changing one source weight vector $\boldsymbol{w}^i$ changes the value of V_j by at most $\frac{1}{p}$. Thus, we can apply McDiarmid's inequality and obtain that $\Pr[V_j \geq t] \leq e^{-2t^2 p}$. Taking the union bound over all $V_j, 1 \leq j \leq m$ we get $\Pr[\max_j V_j \geq t] \leq me^{-2t^2 p}$. Then, setting $t = \sqrt{\ln(m/\delta)/2p}$ yields

$$\Pr\left[\forall j:\ \mathbb{E}[|w_j^1|^q] - \frac{1}{p}\sum_{i=1}^p |w_j^i|^q \leq \sqrt{\frac{\ln\frac{m}{\delta}}{2p}}\right] \geq 1-\delta. \tag{9}$$

We are now ready to bound $\inf_{\boldsymbol{w}\in B_b} \|\boldsymbol{w} - \boldsymbol{w}^*\|_q$, which is the projection of $\boldsymbol{w}^*$ on B_b. Define $\boldsymbol{w}_B = \boldsymbol{w}^* \circ \min(\mathbf{1}, b\hat{\boldsymbol{\mu}})$, where the min on a vector is understood elementwise. Using $\|\boldsymbol{w}^*\| \leq 1$, it is readily verified that $\boldsymbol{w}_B \in B_b$. Hence, defining $\boldsymbol{x}_+ := \max(\mathbf{0}, \boldsymbol{x})$, we have

$$\begin{aligned}
&\Pr\left[\inf_{\boldsymbol{w}\in B_b} \|\boldsymbol{w} - \boldsymbol{w}^*\|_q \geq t\right] \\
&\leq \Pr\left[\|\boldsymbol{w}_B - \boldsymbol{w}^*\|_q \geq t\right] \\
&= \Pr\left[\|\boldsymbol{w}^* - \boldsymbol{w}_B\|_q \geq t\right] \\
&= \Pr\left[\left\|\boldsymbol{w}^* \circ \left(\mathbf{1} - \frac{b}{p}\sum_{i=1}^p |\boldsymbol{w}^i|^q\right)_+\right\|_q \geq t\right] \\
&= \Pr\left[\sum_{j=1}^m \left|w_j^* \left(1 - \frac{b}{p}\sum_{i=1}^p |w_j^i|^q\right)_+\right|^q \geq t^q\right] \\
&\leq \frac{1}{t^q}\sum_{j=1}^m \underbrace{\mathbb{E}\left[|w_j^*|^q\right] \mathbb{E}\left[\left(1 - \frac{b}{p}\sum_{i=1}^p |w_j^i|^q\right)_+^q\right]}_{=:W_j}.
\end{aligned} \tag{10}$$

The last step is an application of Markov's inequality. We are now left with bounding the right hand side of (10).

We start by distinguishing between the cases $\mathbb{E}[|w_j^*|^q] \leq \sqrt{\ln\frac{m}{\delta} /\ 2p}$ and $\mathbb{E}[|w_j^*|^q] > \sqrt{\ln\frac{m}{\delta}\ /\ 2p}$. In the first case, we can use the bound $W_j \leq \mathbb{E}[|w_j^*|^q] \leq \sqrt{\ln\frac{m}{\delta}\ /\ 2p}$, because $\mathbb{E}\left[\left(1 - \frac{b}{p}\sum_{i=1}^p |w_j^i|^q\right)_+^q\right] \leq 1$. In the second case, we substitute (9) into the definition of W_j, so that it holds with probability larger than

$1-\delta$ that $W_j \le \mathbb{E}\left[|w_j^*|^q\right]\left(1 - b\,\mathbb{E}\left[|w_j^*|^q\right] + b\sqrt{\ln\frac{m}{\delta}/2p}\right)_+^q$. In both cases, W_j is no more than

$$W_j \le \mathbb{E}\left[|w_j^*|^q\right]\left(1 - b\left(\mathbb{E}\left[|w_j^*|^q\right] + \sqrt{\ln\frac{m}{\delta}/2p}\right)_+\right)_+^q.$$

We now proceed by bounding each W_j independently. Setting $a := \sqrt{\frac{\ln\frac{m}{\delta}}{2p}}$ and $x := \mathbb{E}\left[|w_j^*|^q\right]$, the above has the form $f(x) = x[1 - bx + ba]_+^q$, when restricting f to the interval $[a, 1]$. A straightforward calculation shows that f has only one positive maximum at the position $x' = \frac{1+ab}{b+qb}$. If $ab > \frac{1}{q}$, then $x' < a$, so f reaches its maximum at the interval border $x = a$ with maximum value $f(x) = f(a) = a$. On the other hand, if $ab < \frac{1}{q}$, then $\boldsymbol{x}' > a$ and we can use $f(x') = \frac{q^q}{(q+1)^{q+1}}\frac{(1+ab)^{q+1}}{b}$ as an upper bound. In both cases f is not larger than $\frac{q^q}{b(q+1)^{q+1}} + a$. Re-substituting the definitions of a and x' we obtain $W_j \le \frac{q^q}{b(q+1)^{q+1}} + \sqrt{\frac{\ln\frac{m}{\delta}}{2p}}$. Plugging this into (10) yields

$$\begin{aligned}&\Pr\left[\inf_{\boldsymbol{w}_B\in B_b}\|\boldsymbol{w}^* - \boldsymbol{w}_B\| \ge t\right] \\ &\le \frac{m}{t^q}\left(\frac{q^q}{b(q+1)^{q+1}} + \sqrt{\frac{\ln\frac{m}{\delta}}{2p}}\right).\end{aligned} \tag{11}$$

The result follows by setting $t = \left[\frac{m}{\epsilon}\left(\frac{q^q}{b(q+1)^{q+1}} + \sqrt{\frac{\ln\frac{m}{\delta}}{2p}}\right)\right]^{\frac{1}{q}}$.

Of course, this inequality is loose in the sense that it gives non-trivial upper bounds only for cases where $b \ge m$. Ideally one would like to have $b < \sqrt{m}$, because this would improve the estimation error over the standard Rademacher bound, which is $O(\sqrt{m/n^\circ})$. The looseness is not surprising, because we have not made any assumption about the dataset-generating distribution $\mathcal{T}$. In the worst case, the distribution might have large variance, so that the source data sets do not contain any useful information about the target weight vector. However, it is easy to see how the result can be adapted to incorporate knowledge about $\mathcal{T}$. In the following, we give two results, where we make additional assumptions on $\mathcal{T}$.

3.3 Approximation Error with Sparse and Concentrated Moment Vectors

In the first case, we assume that only a fraction of the features are informative, so that some of the moment vector $\boldsymbol{\mu}$'s components can be bounded by a small constant. In the second case, we assume that the variance of the moment vectors is bounded.

Theorem 3. *Consider the same setting as in Theorem 2, but assume that there are $m_1 < m$ uninformative features, so that $\boldsymbol{\mu}_j \leq c$ for $1 \leq j \leq m_1$ and some small constant $c > 0$, while the remaining $m_2 := m - m_1$ features are informative, i.e. $\boldsymbol{\mu}_j$ is possibly larger than c for $m_1 < j \leq m$. Then, with probability greater than $1 - \delta$ over the choice of the source data sets and with probability greater than $1 - \epsilon$ over the choice of the target learning task, the approximation error can be upper-bounded by*

$$\mathrm{er}_a \leq \frac{Lm^{\frac{q-1}{q}}}{b^{\frac{1}{q}}} \left[\frac{m_1 c}{\epsilon} + \frac{m_2}{\epsilon} \left(\frac{q^q}{(q+1)^{q+1}} + b\sqrt{\frac{\ln \frac{m}{\delta}}{2p}} \right) \right]^{\frac{1}{q}}$$

Proof. The proof follows the one of Theorem 2, but differs in the bound for the right hand side of (10). For the first $1 \leq j \leq m_1$ summands in (10), we can use $W_j \leq c$, because $\mathbb{E}\left[\left(1 - \frac{b}{p} \sum_{i=1}^{p} |w_j^i|^q \right)_+^q \right] \leq 1$. For the remaining m_2 summands, we use the original bound $W_j \leq \frac{q^q}{b(q+1)^{q+1}} + \sqrt{\frac{\ln \frac{m}{\delta}}{2p}}$. Plugging this into (10) yields

$$\Pr\left[\inf_{\boldsymbol{w}_B \in B_b} \|\boldsymbol{w}^* - \boldsymbol{w}_B\| \geq t \right] \leq \frac{m_1 c}{t^q} + \frac{m_2}{t^q} \left(\frac{q^q}{b(q+1)^{q+1}} + \sqrt{\frac{\ln \frac{m}{\delta}}{2p}} \right).$$

The result follows by setting $t = \left[\frac{m_1 c}{\epsilon} + \frac{m_2}{\epsilon} \left(\frac{q^q}{b(q+1)^{q+1}} + \sqrt{\frac{\ln \frac{m}{\delta}}{2p}} \right) \right]^{\frac{1}{q}}$.

The result leads to particularly tight bounds, if $q = 1$ and m_2 is small. It shows that one can achieve a comparably low generalization error, even if it is not known in advance, which features are informative and which ones are not. The following theorem deals with the case where all features are informative, but the moment vectors for all source and target data sets are concentrated sharply around the mean.

Theorem 4. *Consider the same setting as in Theorem 2, but assume that the variance of the moment vectors for the source and target data sets is bounded, that is, $\mathbb{E}[(\mu_j - \mathbb{E}[\mu_j])^2] \leq v$ is bounded by a constant v for all $1 \leq j \leq m$. Then, with probability greater than $1 - \delta$ over the choice of the source data sets and with probability greater than $1 - \epsilon$ over the choice of the target learning task, the approximation error can be upper-bounded by*

$$\mathrm{er}_a \leq Lm \left[\frac{1}{\epsilon} \left(\frac{q^q}{b(q+1)^{q+1}} + \frac{\ln \frac{m}{\delta}}{3p} + \frac{\sqrt{\frac{4}{9}[\ln \frac{m}{\delta}]^2 + 8pv}}{2p} \right) \right]^{\frac{1}{q}} \tag{12}$$

If $v > 1$, the rightmost summand of the bound scales with $O(\sqrt{1/p})$, just as with the original result in Theorem 2. However, if v is close to zero, the two left summands are bounded by a $O(1/p)$ factor, leading to a significantly tighter bound.

Proof. The result follows by using the following inequality instead of (9) in the proof of Theorem 2:

$$\Pr\left[\forall j:\ \mathbb{E}[|w_j^1|^q] - \frac{1}{p}\sum_{i=1}^{p}|w_j^i|^q \leq \frac{\ln\frac{m}{\delta}}{3p} + \frac{\sqrt{\frac{1}{9}[\ln\frac{m}{\delta}]^2 + 2pv\ln\frac{m}{\delta}}}{p}\right] \geq 1-\delta. \tag{13}$$

To see that (13) holds, consider the random variable $V_j := \mu_j - \hat{\mu}_j \overset{\text{Def.(3)}}{=} \mathbb{E}[|w_j^1|^q] - \frac{1}{p}\sum_{i=1}^{p}|w_j^i|^q$. Since the variance of the moment vector's components is bounded by v, we can use Bernstein's inequality and obtain

$$\Pr[V_j \geq t] \leq \exp\left[\frac{pt^2}{2v + \frac{2t}{3}}\right].$$

Taking the union bound over all V_j, $1 \leq j \leq m$, we get

$$\Pr[\max_j V_j \geq t] \leq m\exp\left[\frac{pt^2}{2v + \frac{2t}{3}}\right].$$

Setting $t = \frac{1}{3p}\ln\frac{m}{\delta} + \frac{1}{p}\sqrt{\frac{1}{9}[\ln\frac{m}{\delta}]^2 + 2pv\ln\frac{m}{\delta}}$ yields (13).

Of course, there are many other possible assumptions about the learning task generating distribution $\mathcal{T}$ which would lead to non-trivial error bounds.

3.4 Discussion

We are now able to combine the bounds for the estimation and approximation error to obtain a bound for the total regret. For instance, setting $q = 1$ and using the setting in Theorem 3, we get the following Corollary:

Corollary 1. *Under the conditions of Theorem 3 it holds for the empirical risk minimizer* $\boldsymbol{w}^\circ \in B_b$ *defined in* (T) *and* (B) *for* $q = 1$

$$\mathrm{er}(\boldsymbol{w}^\circ) - \mathrm{er}(\boldsymbol{w}^*) \leq 4\sqrt{\frac{2\ln\frac{4}{\delta}}{n^\circ}} + L\left(\frac{4b}{\sqrt{n^\circ}} + \frac{m_1 c}{b\epsilon} + \frac{m_2}{b\epsilon}\left(\frac{1}{4} + \sqrt{\frac{\ln\frac{m}{\delta}}{2p}}\right)\right)$$

.

The bound can be expected to improve on standard regret bounds if c and the number of informative features m_2 are small, and the number of uninformative features m_1 is large. This is a very reasonable assumption as there are many problems in practice where only a few features are relevant and the large majority of features gets assigned only small weights (see, for example, [10]). In this scenario, we obtain a bound of the form $O(\frac{b}{\sqrt{n}} + \frac{m_1 c}{b} + \frac{m_2}{b\sqrt{p}})$ (omitting logarithmic factors), which can be considerably smaller than the $O(\sqrt{m/n})$ rate achieved by

Algorithm 1. *Moment-based transfer learning algorithm based on* (B) *and* (T)

```
1: input target data set $T^\circ$ and source data sets $T^i = (X^i, Y^i)$, $i = 1, ..., p$
2: for $i = 1$ to $p$
3:     compute SVM weight vectors $\boldsymbol{w}^i := \mathrm{SVM}(X^i, Y^i)$ for source data sets $(X^i, Y^i)$, $i = 1, ..., p$,
           with SVM parameter $C$ tuned on a validation set
4:     normalize each weight vector to unit norm: $\boldsymbol{w}^i := \boldsymbol{w}^i / \, \|\boldsymbol{w}^i\|_2$ for each $i = 1, \ldots, p$
5: end for
6: for various values of $q$
7:     compute moment vector $\hat{\boldsymbol{\mu}} := \left(\frac{1}{p}\sum_{i=1}^{p} |\boldsymbol{w}^i|^q\right)$
8:     reweight target training data: $\forall i = 1, \ldots, n :\ \boldsymbol{x}_i^\circ := \hat{\boldsymbol{\mu}}\boldsymbol{x}_i^\circ$
9:     train SVM on target data set with parameter $C$ tuned on validation set
10:    denote the so-obtained weight vector by $\boldsymbol{w}_q^\circ$
11: end for
12: output the one SVM weight vector $\boldsymbol{w}_q^\circ$ with $q$ such that the error on the validation
    set is minimal
```

standard Rademacher-style concentration results. As a by-product, our analysis shows that the transfer learning approach is most beneficial when the sample size n is small and the dimensionality m is large. This is in accordance with anecdotal reports indicating that transfer learning is especially beneficial in small sample cases [15].

4 Algorithmic Details

In this section we describe the moment-based transfer learning algorithm based on (B) and (T) that we employed in the experiments in Section 5. To this aim, let us consider (B). It is easy to see that, instead of optimizing the original criterion (T), one can equivalently optimize a regular Support Vector Machine (SVM) [7] with the target data preprocessed by feature reweighting as follows: $\boldsymbol{x}_i^{\text{new}} := \boldsymbol{x}_i^{\text{old}} \circ \hat{\boldsymbol{\mu}}$. To see this, note that by employing a change of variables $\tilde{\boldsymbol{w}} = \boldsymbol{w} \circ \hat{\boldsymbol{\mu}}^{-1}$ it holds for the original criterion

$$\begin{aligned}
\hat{\boldsymbol{w}}^\circ &\overset{(T)}{=} \underset{\boldsymbol{w}:\|\boldsymbol{w}\circ\hat{\boldsymbol{\mu}}^{-1}\|\leq C}{\operatorname{argmin}} \frac{1}{n^\circ}\sum_{i=1}^{n^\circ} \ell(y_i^\circ \boldsymbol{w}^\top \boldsymbol{x}_i^\circ) \\
&= \underset{\tilde{\boldsymbol{w}}:\|\tilde{\boldsymbol{w}}\|\leq C}{\operatorname{argmin}} \frac{1}{n^\circ}\sum_{i=1}^{n^\circ} \ell(y_i^\circ \tilde{\boldsymbol{w}}^\top (\boldsymbol{x}_i^\circ \circ \hat{\boldsymbol{\mu}}))
\end{aligned}$$

The proposed method can now be stated as Algorithm 1. Lines 2–5 compute the optimal SVM weight vectors $\boldsymbol{w}^i$ on the source data sets and lines 6–11 perform the actual transfer learning on the target data set as defined on (B) and (T). In line 7 the transferred moments $\hat{\mu}$ are computed and in line 8–9 the actual transfer learning step is performed—as discussed above this is achieved by reweighting the features (line 8) and subsequently training an SVM on the so-obtained features (line 9). The final weight vector for the target task is output in line 12. The parameters q and C of our algorithm are tuned on a validation set.

5 Experiments

In this section we report on experiments with two application domains, text document classification and structure-activity-relationships. The first application domain, text document classification, is well suited for transfer learning because, even for very specialized topics, the Internet provides a large body of related source learning tasks. We downloaded ten data sets from TechTC, the Technion repository of text categorization data sets[2] [9]. Each data set contains between 142 and 277 text documents from two categories taken from the web directory *Open Directory Project*. In total this results in 1794 documents. The (binary classification) task is to tell for each data set the two categories apart. We employed a (binary) bag of words feature representation as provided by TechTC resulting in total in 142468 features.

To evaluate the predictive accuracy of the induced classifiers, we randomly split each data set into $r = 250$ training/validation/test partitions of size 50/25/25%. Then, we set one data set as target learning problem aside and kept the remaining data sets as source data. This process is repeated for each data set. Subsequently, we run Algorithm 1, a baseline (linear) SVM and the method of Evgeniou & Pontil [8] on the training partitions. For each repetition of the experiment, the optimal parameters were determined by a grid search over $q \in 10^{[-1,-0.8,\ldots,1]}$ and $C \in 10^{[-3,2.5,\ldots,4]}$ on base of the validation data set,[3] and test errors are computed on the test partition for the optimal parameter choices.

We give the results on the left hand side of Figure 1. The error bars indicate standard errors over the 250 repetitions. One can see that the method of Evgeniou & Pontil achieves an test error that is about 2% lower than the one of the SVM baseline for most data sets. The proposed method is on par with the SVM baseline for seven of the ten data sets and it is never worse than the SVM (as it contains the SVM as a special case for $q \approx 0$). For three data sets our method clearly outperforms the two other approaches with drastic gains in accuracy ranging from 13% to 19%.

In order to investigate why our method performed considerably better on these three particular data sets than on the remaining ones, we performed another experiment. We trained an SVM for each data set, using all feature vectors and all instances. This yields ten linear classifiers, that is, weight vectors $\hat{\boldsymbol{w}}^i$. We then compute the pairwise (absolute) correlation coefficients $\rho_{i,j} := |\mathrm{corr}(\hat{\boldsymbol{w}}^i, \hat{\boldsymbol{w}}^j)|$ for all $i, j = 1, \ldots, 10$. The result is shown in Figure 1 (right). One can see that most tasks are only weakly correlated. This explains that our method did not improve

[2] The data sets are available at `http://techtc.cs.technion.ac.il/`.

[3] Optimal values of C were attained inside the grid. The second regularization parameter δ of the method of Evgeniou & Pontil was also determined by grid search: $\delta \in 10^{[-2,-1,\ldots,2]}$.

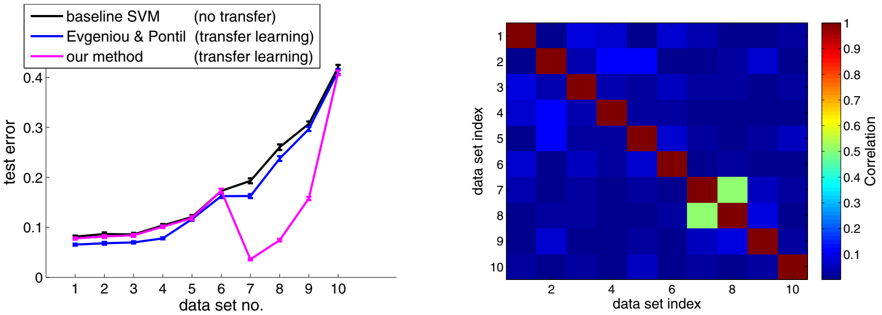

Fig. 1. Empirical results of the text categorization experiment: test errors (left) and correlation coefficients of the SVM weights (right). Vertical bars indicate standard errors. One can see that the correlation is maximized for the data set pair (7,8)—this accordance with the test errors: the gain in accuracy of our method over the baselines is maximal on these data.

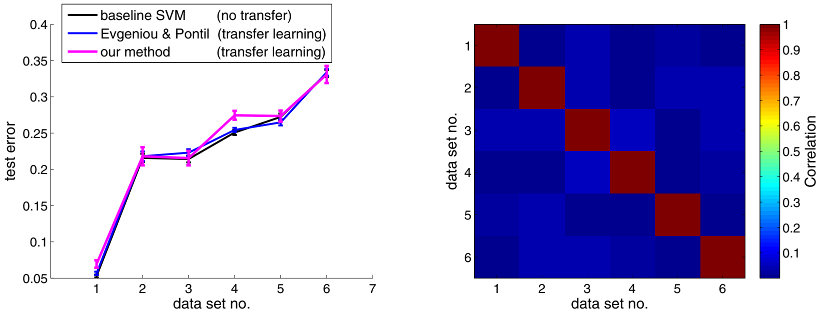

Fig. 2. Empirical results of the structure-activity relationship experiment

over the baselines on most data sets—one might conjecture that the corresponding tasks are only weakly related. However, the correlation is substantially stronger for the data set pair (7,8): it is $\rho_{7,8} = 0.51$ while the second largest coefficient only has $\rho_{2,9} = 0.11$. This observation is in accordance with our empirical results: the gain in accuracy over the baselines was the highest for data sets 7 and 8 (19% and 16%, respectively)—thus, we conclude that those two tasks are very closely related and this is why our method works best in these cases.

We also performed some experiments with learning the biological activity of compounds given their molecular structure as a graph. We obtained the six datasets used in [14]. Each dataset contains a number of molecular graphs and about 1000 features testing for the presence of one frequently occurring substructure. We again randomly split each data set into $r = 100$ training/validation/test partitions of size 50/25/25%, evaluated the proposed approach and compared it to the baseline approaches as done above. It turns out that there is no gain in using the proposed method for these data sets (see Figure 2). A closer investigation indicates that there are too many substructure features, which are distinct between the individual tasks to make feature-level transfer suitable. This is true both for our approach as the one by Evgeniou & Pontil. It is an interesting open question whether one could use feature description data to transfer information between distinct, but similar features.

6 Discussion and Conclusion

In this paper we presented a transfer learning approach for adjusting the regularizer of a target learning problem. This is an important task for many of the modern machine learning applications, where the features often outnumber the training instances. Empirical results have shown that it is often not enough to impose strong standard regularizers (e.g. to encourage sparsity), but that individual learning problems benefit from customized regularization [2,8,13,6,16]. The results in this paper demonstrate that adaptive regularization can be successfully applied to transfer information from source to target data sets. The main idea is to extend an ℓ_2-norm regularizer with feature weights and to transfer good values for these weights from the source data sets. A theoretical analysis showed that the expected prediction error depends critically on the trade-off between estimation and approximation error. If the source classifiers are close to the optimal target classifier, then it is possible to keep the approximation error small simply by choosing a strong regularizer that penalizes weight vectors too far away from the source classifiers. The empirical analysis on real text classification data shows that our approach works well in practice if the dataset share transferable information: for some data sets a gain in accuracy of up to 19% was observed while it never performed worse than the SVM baseline. It is an open question whether the bounds can be improved for special cases, and if other parametrization approaches lead to better theoretical or practical results.

Acknowledgments. UR and MK acknowledge a grant by the German National Science Foundation (DFG Ru 1589/1-1) and a scholarship by the German Academic Exchange Service (DAAD), respectively. This work was supported in part by the European Community under the PASCAL2 Network of Excellence (ICT-216886).

References

1. Ando, R.K., Zhang, T.: A framework for learning predictive structures from multiple tasks and unlabeled data. Journal of Machine Learning Research 6, 1817–1853 (2005)
2. Argyriou, A., Evgeniou, T., Pontil, M.: Convex multi-task feature learning. Machine Learning 73(3), 243–272 (2008)
3. Bartlett, P.L., Mendelson, S.: Rademacher and gaussian complexities: Risk bounds and structural results. JMLR 3, 463–482 (2002)
4. Baxter, J.: A model of inductive bias learning. Journal of Artificial Intelligence Research 12, 149–198 (2000)

5. Ben-David, S., Schuller, R.: Exploiting task relatedness for mulitple task learning. In: Proceedings of the 16th Annual Conference on Computational Learning Theory, pp. 567–580 (2003)
6. Caruana, R.: Multitask learning. Mach. Learn. 28, 41–75 (1997)
7. Cortes, C., Vapnik, V.N.: Support vector networks. Machine Learning 20, 273–297 (1995)
8. Evgeniou, T., Pontil, M.: Regularized multi–task learning. In: KDD 2004: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 109–117. ACM, New York (2004)
9. Gabrilovich, E., Markovitch, S.: Parameterized generation of labeled datasets for text categorization based on a hierarchical directory. In: Proceedings of The 27th Annual International ACM SIGIR Conference, Sheffield, UK, pp. 250–257. ACM Press, New York (2004)
10. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182 (2003)
11. Maurer, A.: Bounds for linear multi-task learning. J. Mach. Learn. Res. 7, 117–139 (2006)
12. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 99 (2009) (PrePrints)
13. Raina, R., Ng, A.Y., Koller, D.: Constructing informative priors using transfer learning. In: ICML 2006: Proceedings of the 23rd International Conference on Machine Learning, pp. 713–720. ACM, New York (2006)
14. Rückert, U., Kramer, S.: Kernel-based inductive transfer. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part II. LNCS (LNAI), vol. 5212, pp. 220–233. Springer, Heidelberg (2008)
15. Schweikert, G., Widmer, C., Schölkopf, B., Rätsch, G.: An empirical analysis of domain adaptation algorithms for genomic sequence analysis. In: Advances in Neural Information Processing Systems, vol. 21, pp. 1433–1440 (2009)
16. Zhong, E., Fan, W., Peng, J., Verscheure, O., Ren, J.: Universal learning over related distributions and adaptive graph transduction. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009. LNCS, vol. 5782, pp. 678–693. Springer, Heidelberg (2009)
17. Zhong, E., Fan, W., Peng, J., Zhang, K., Ren, J., Turaga, D.S., Verscheure, O.: Cross domain distribution adaptation via kernel mapping. In: Knowledge Discovery and Data Mining, pp. 1027–1036 (2009)

A Proof of Theorem 1

Let $S_n = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ be a set of independent Rademacher variables, which obtain the values -1 or +1 with the same probability 0.5. Then, the *Rademacher complexity* of the class of linear classifiers with regularizer B_b is given by $\mathcal{R}_B := \mathbb{E}_{S_n}\left[\left|\sup_{\boldsymbol{w}\in B_b} \frac{1}{n}\sum_{i=1}^{n} s_i \boldsymbol{w}^\top \boldsymbol{x}_i^\circ\right|\right]$. We will now give an upper bound for the Rademacher complexity of the moment-based approach to transfer learning.

Proposition 1. *Then, the Rademacher complexity of linear classifiers with B_b regularization as defined in* (B) *is upper-bounded by:*

$$\mathcal{R}_B \leq \frac{b}{\sqrt{n}} \; .$$

Proof. By employing variable substitutions of the form $\boldsymbol{v} = b^{-1}\boldsymbol{w} \circ \hat{\boldsymbol{\mu}}^{-1}$ we can use the Cauchy-Schwarz (C-S) inequality to bound the Rademacher complexity $\mathcal{R}_B$ as follows:

$$\begin{aligned}
\mathcal{R}_B &= \mathbb{E}_{S_n}\left[\sup_{\boldsymbol{v}:\|\boldsymbol{v}\|_2\leq 1}\left|\frac{1}{n}\sum_{i=1}^{n} s_i\left(b\boldsymbol{v}\circ\hat{\boldsymbol{\mu}}\right)^\top \boldsymbol{x}_i^\circ\right|\right] \\
&\overset{\mathrm{C-S}}{\leq} \mathop{\mathbb{E}}_{S_n}\left[\sup_{\boldsymbol{v}:\|\boldsymbol{v}\|_2\leq 1}\|\boldsymbol{v}\|\left\|\frac{b}{n}\sum_{i=1}^{n} s_i\left(\hat{\boldsymbol{\mu}}\circ\boldsymbol{x}_i^\circ\right)\right\|\right] \\
&= \frac{b}{n}\mathbb{E}_{S_n}\left[\sqrt{\sum_{i,j=1}^{n} s_i s_j\left(\hat{\boldsymbol{\mu}}\circ\boldsymbol{x}_i\right)^\top\left(\hat{\boldsymbol{\mu}}^q\circ\boldsymbol{x}_j\right)}\right] \\
&\leq \frac{b}{n}\sqrt{\sum_{i,j=1}^{n}\mathbb{E}_{S_n}\left[s_i s_j\left(\hat{\boldsymbol{\mu}}\circ\boldsymbol{x}_i\right)^\top\left(\hat{\boldsymbol{\mu}}\circ\boldsymbol{x}_j\right)\right]} \\
&= \frac{b}{n}\sqrt{\mathbb{E}_{S_n}\sum_{i=1}^{n}\|\hat{\boldsymbol{\mu}}\circ\boldsymbol{x}_i\|^2} \;\leq\; \frac{b}{n}\sqrt{\mathbb{E}_{S_n}\sum_{i=1}^{n}\|\hat{\boldsymbol{\mu}}\|^2}
\end{aligned}$$

where for the third step we use that the Rademacher variables are independent, and in the forth step that the data is in $[-1,1]^m$. Recall that $\|\boldsymbol{w}^i\|_2 \leq 1$ for all i and thus $\||\boldsymbol{w}^i|^q\|_2 \leq 1$. Hence,

$$\|\hat{\boldsymbol{\mu}}\| \leq \|\frac{1}{p}\sum_{i=1}^{p}|\boldsymbol{w}^i|^q\| \leq \frac{1}{p}\sum_{i=1}^{p}\||\boldsymbol{w}^i|^q\| \leq 1.$$

Combining this with the above bound gives the claimed result.

If the Rademacher complexity of a class of classifiers is known, it can be used to bound the generalization error:

Theorem 5 ([3]). *Suppose the loss $\ell : \mathbb{R} \supset X \to [0,1]$ is Lipschitz with constant L. Then, the following holds with probability larger than $1-\delta$:*

$$\sup_{\boldsymbol{w}\in B_b}\left|\mathrm{er}^\circ(\boldsymbol{w}) - \hat{\mathrm{er}}^\circ(\boldsymbol{w})\right| \;\leq\; 2L\mathcal{R}_B + 4\sqrt{\frac{2\ln\frac{4}{\delta}}{n}}\ .$$

Theorem 1 follows now from combining the previous two results.