Stability of Transductive Regression Algorithms

Corinna Cortes CORINNA@GOOGLE.COM
Google Research, 76 Ninth Avenue, New York, NY 10011.

Mehryar Mohri MOHRI@CIMS.NYU.EDU
Courant Institute of Mathematical Sciences and Google Research, 251 Mercer Street, New York, NY 10012.

Dmitry Pechyony PECHYONY@CS.TECHNION.AC.IL
Technion - Israel Institute of Technology, Haifa 32000, Israel.

Ashish Rastogi RASTOGI@CS.NYU.EDU
Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012.

Abstract

This paper uses the notion of algorithmic stability to derive novel generalization bounds for several families of transductive regression algorithms, both by using convexity and closed-form solutions. Our analysis helps compare the stability of these algorithms. It suggests that several existing algorithms might not be stable but prescribes a technique to make them stable. It also reports the results of experiments with local transductive regression demonstrating the benefit of our stability bounds for model selection, in particular for determining the radius of the local neighborhood used by the algorithm.

1. Introduction

Many learning problems in information extraction, computational biology, natural language processing and other domains can be formulated as *transductive inference* problems (Vapnik, 1982). In the transductive setting, the learning algorithm receives both a labeled training set, as in the standard induction setting, and a set of unlabeled test points. The objective is to predict the labels of the test points. No other test points will ever be considered. This setting arises in a variety of applications. Often, the points to label are known but they have not been assigned a label due to the prohibitive cost of labeling. This motivates the use of transductive algorithms which leverage the unlabeled data during training to improve learning performance.

This paper deals with transductive regression, which arises in problems such as predicting the real-valued labels of the nodes of a known graph in computational biology, or the scores associated with known documents in information extraction or search engine tasks.

Several algorithms have been devised for the specific setting of transductive regression (Belkin et al., 2004b; Chapelle et al., 1999; Schuurmans & Southey, 2002; Cortes & Mohri, 2007). Several other algorithms introduced for transductive classification can be viewed in fact as transductive regression ones as their objective function is based on the squared loss, e.g., (Belkin et al.2004a; 2004b). Cortes and Mohri (2007) also gave explicit VC-dimension generalization bounds for transductive regression that hold for all bounded loss functions and coincide with the tight classification bounds of Vapnik (1998) when applied to classification.

This paper presents novel algorithm-dependent generalization bounds for transductive regression. Since they are algorithm-specific, these bounds can often be tighter than bounds based on general complexity measures such as the VC-dimension. Our analysis is based on the notion of algorithmic stability.

In Sec. 2 we give a formal definition of the transductive regression setting and the notion of stability for transduction. Our bounds generalize the stability bounds given by Bousquet and Elisseeff (2002) for the inductive setting and extend to regression the stability-based transductive classification bounds of (El-Yaniv & Pechyony, 2006). Standard concentration bounds

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

such as McDiarmid's bound (McDiarmid, 1989) cannot be readily applied to the transductive regression setting since the points are not drawn independently but uniformly without replacement from a finite set. Instead, a generalization of McDiarmid's bound that holds for random variables sampled without replacement is used, as in (El-Yaniv & Pechyony, 2006). Sec. 3.1 gives a simpler proof of this bound.

This concentration bound is used to derive a general transductive regression stability bound in Sec. 3.2. In Sec. 4, we present the stability coefficients for a family of local transductive regression algorithms. The analysis in this section is based on convexity. In Sec. 5, we study the stability of other transductive regression algorithms (Belkin et al., 2004a; Wu & Schölkopf, 2007; Zhou et al., 2004; Zhu et al., 2003) based on their closed form solution and propose a modification to the seemingly unstable algorithm that makes them stable and guarantees a non-trivial generalization bound. Finally, Sec. 6 shows the results of experiments with local transductive regression demonstrating the benefit of our stability bounds for model selection, in particular for determining the radius of the local neighborhood used by the algorithm. This provides a partial validation of our bounds and analysis.

2. Definitions

Let us first describe the transductive learning setting. Assume that a full sample X of $m+u$ examples is given. The learning algorithm further receives the labels of a random subset S of X of size m which serves as a training sample. The remaining u unlabeled examples, $x_{m+1}, \ldots, x_{m+u} \in X$, serve as test data. We denote by $X \vdash (S,T)$ a partitioning of X into the training set S and the test set T. The *transductive learning* problem consists of predicting accurately the labels $y_{m+1}, \ldots, y_{m+u}$ of the test examples, no other test examples will ever be considered (Vapnik, 1998).[1] The specific problems where the labels are real-valued numbers, as in the case studied in this paper, is that of *transduction regression*. It differs from the standard (*induction*) regression since the learning algorithm is given the unlabeled test examples beforehand and can thus exploit this information to improve performance.

We denote by $c(h,x)$ the cost of an error of a hypothesis h on a point x labeled with $y(x)$. The cost function commonly used in regression is the squared loss $c(h,x) = (h(x) - y(x))^2$. In the remaining of this paper, we will assume a squared loss but many of our results generalize to other convex cost functions. The training and test errors of h are respectively $\widehat{R}(h) = \frac{1}{m}\sum_{k=1}^{m} c(h,x_k)$ and $R(h) = \frac{1}{u}\sum_{k=1}^{u} c(h,x_{m+k})$. The generalization bounds we derive are based on the notion of transductive algorithmic stability.

Definition 1 (Transduction β-stability)**.** *Let L be a transductive learning algorithm and let h denote the hypothesis returned by L for $X \vdash (S,T)$ and h' the hypothesis returned for $X \vdash (S',T')$. L is said to be* uniformly β-stable *with respect to the cost function c if there exists $\beta \geq 0$ such that for any two partitionings $X \vdash (S,T)$ and $X \vdash (S',T')$ that differ in exactly one training (and thus test) point and for all $x \in X$,*

$$\left|c(h,x) - c(h',x)\right| \leq \beta. \tag{1}$$

3. Transduction Stability Bounds

3.1. Concentration Bound for Sampling without Replacement

Stability-based generalization bounds in the inductive setting are based on McDiarmid's inequality (1989). In the transductive setting, the points are drawn uniformly without replacement and thus are not independent. Therefore, McDiarmid's concentration bound cannot be readily used. Instead, a generalization of McDiarmid's bound for sampling without replacement is needed as in El-Yaniv and Pechyony (2006).

We will denote by $\mathbf{S}_1^m$ a sequence of random variables $S_1, \ldots, S_m$ and write $\mathbf{S}_1^m = \mathbf{x}_1^m$ as a shorthand for the m equalities $S_i = x_i$, $i = 1, \ldots, m$ and $\Pr[\mathbf{x}_{i+1}^m | \mathbf{x}_1^{i-1}, x_i] = \Pr[\mathbf{S}_{i+1}^m = \mathbf{x}_{i+1}^m | \mathbf{S}_1^{i-1} = \mathbf{x}_1^{i-1}, S_i = x_i]$.

Theorem 1 ((McDiarmid, 1989), 6.10)**.** *Let $\mathbf{S}_1^m$ be a sequence of random variables, each S_i taking values in the set X, and assume that a measurable function $\phi : X^m \mapsto \mathbb{R}$ satisfies: $\forall i \in [1,m], \forall x_i, x_i' \in X$,*

$$\left|\mathbb{E}_{\mathbf{S}_{i+1}^m}\left[\phi | \mathbf{S}_1^{i-1}, S_i = x_i\right] - \mathbb{E}_{\mathbf{S}_{i+1}^m}\left[\phi | \mathbf{S}_1^{i-1}, S_i = x_i'\right]\right| \leq c_i.$$

Then, $\forall \epsilon > 0, \quad \Pr\left[|\phi - \mathbb{E}\left[\phi\right]| \geq \epsilon\right] \leq 2\exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}\right)$.

The following is a concentration bound for sampling without replacement needed to analyze the generalization of transductive algorithms.

Theorem 2. *Let $\mathbf{x}_1^m$ be a sequence of random variables, sampled from an underlying set X of $m+u$ elements without replacement, and let that $\phi : X^m \mapsto \mathbb{R}$*

[1] Another natural setting for transduction is one where the training and test samples are both drawn according to the same distribution and where the test points, but not their labels, are made available to the learning algorithm. However, as pointed out by Vapnik (1998), any generalization bound in the setting we analyze directly yields a bound for this other setting, essentially by taking the expectation.

be a symmetric function such that for all $i \in [1,m]$ and for all $x_1,\ldots,x_m \in X$ and $x'_1,\ldots,x'_m \in X$,

$$\left|\phi(x_1,\ldots,x_m) - \phi(x_1,\ldots,x_{i-1},x'_i,x_{i+1},\ldots,x_m)\right| \leq c.$$

Then, $\forall \epsilon > 0, \quad \Pr\left[\left|\phi - \mathbb{E}\left[\phi\right]\right| \geq \epsilon\right] \leq 2\exp\left(\frac{-2\epsilon^2}{\alpha(m,u)c^2}\right),$

where $\alpha(m,u) = \frac{mu}{m+u-1/2} \cdot \frac{1}{1-1/(2\max\{m,u\})}$.

Proof. For a fixed $i \in [1,m]$, let $g(\mathbf{S}_1^{i-1}) = \mathbb{E}_{\mathbf{S}_{i+1}^m}\left[\phi|\mathbf{S}_1^{i-1}, S_i{=}x_i\right] - \mathbb{E}_{\mathbf{S}_{i+1}^m}\left[\phi|\mathbf{S}_1^{i-1}, S_i{=}x'_i\right]$. Then, $g(\mathbf{x}_1^{i-1}) = \sum_{\mathbf{x}_{i+1}^m} \phi(\mathbf{x}_1^{i-1}, x_i, \mathbf{x}_{i+1}^m) \Pr[\mathbf{x}_{i+1}^m|\mathbf{x}_1^{i-1}, x_i] - \sum_{\mathbf{x'}_{i+1}^m} \phi(\mathbf{x}_1^{i-1}, x'_i, \mathbf{x'}_{i+1}^m) \Pr[\mathbf{x'}_{i+1}^m|\mathbf{x}_1^{i-1}, x'_i]$.

For uniform sampling without replacement, the probability terms can be written as: $\Pr\left[\mathbf{x}_{i+1}^m|\mathbf{x}_1^{i-1}, x_i\right] = \prod_{k=i}^{m-1} \frac{1}{m+u-k} = \frac{u!}{(m+u-i)!}$. Thus, $g(\mathbf{x}_1^{i-1}) = \frac{u!}{(m+u-i)!}[\sum_{\mathbf{x}_{i+1}^m} \phi(\mathbf{x}_1^{i-1}, x_i, \mathbf{x}_{i+1}^m) - \sum_{\mathbf{x'}_{i+1}^m} \phi(\mathbf{x}_1^{i-1}, x'_i, \mathbf{x'}_{i+1}^m)]$. To compute the expression between brackets, we divide the set of permutations $\{\mathbf{x'}_{i+1}^m\}$ into two sets, those that contain x_i and those that do not. If a permutation $\mathbf{x'}_{i+1}^m$ contains x_i we can write it as $\mathbf{x'}_{i+1}^{k-1} x_i \mathbf{x'}_{k+1}^m$, where k is such that $x'_k = x_i$. We then match it up with the permutation $x_i \mathbf{x'}_{i+1}^{k-1} \mathbf{x'}_{k+1}^m$ from the set $\{x_i \mathbf{x}_{i+1}^m\}$. These two permutations contain exactly the same elements, and since the function ϕ is symmetric in its arguments, the difference in the value of the function on the permutations is zero.

In the other case, if a permutation $\mathbf{x'}_{i+1}^m$ does not contain the element x_i, then we simply match it up with the same permutation in $\{\mathbf{x}_{i+1}^m\}$. The matching permutations appearing in the summation are then $x_i \mathbf{x'}_{i+1}^m$ and $x'_i \mathbf{x'}_{i+1}^m$ which clearly only differ with respect to x_i. The difference in the value of the function ϕ in this case can be bounded by c. The number of such permutations is $(m-i)!\binom{m+u-(i+1)}{m-i} = \frac{(m+u-i-1)!}{(u-1)!}$, which leads to the following upper bound: $\sum_{\mathbf{x}_{i+1}^m} \phi(\mathbf{x}_1^{i-1}, x_i, \mathbf{x}_{i+1}^m) - \sum_{\mathbf{x'}_{i+1}^m} \phi(\mathbf{x}_1^{i-1}, x'_i, \mathbf{x'}_{i+1}^m) \leq \frac{(m+u-i-1)!}{(u-1)!} c$, which implies that $|g(\mathbf{x}_1^{i-1})| \leq \frac{u!}{(m+u-i)!} \cdot \frac{(m+u-i-1)!}{(u-1)!} c \leq \frac{u}{m+u-i} c$. Then, combining Theorem 1 with the identity $\sum_{i=1}^m \frac{1}{(m+u-i)^2} \leq \frac{m}{m+u-1/2} \frac{1}{u-1/2}$, yields that $\Pr\left[\left|\phi - \mathbb{E}\left[\phi\right]\right| \geq \epsilon\right] \leq 2\exp\left(\frac{-2\epsilon^2}{\alpha_u(m,u)c^2}\right)$, where $\alpha_u(m,u) = \frac{mu}{m+u-1/2} \cdot \frac{1}{1-1/(2u)}$. The function ϕ is symmetric in m and u in the sense that selecting one of the sets uniquely determines the other set. The statement of the theorem then follows from a similar bound with $\alpha_m(m,u) = \frac{mu}{m+u-1/2} \cdot \frac{1}{1-1/(2m)}$, taking the tighter of the two. □

3.2. Transductive Stability Bound

To obtain a general transductive regression stability bound, we apply the concentration bound of Theorem 2 to the random variable $\phi(S) = R(h) - \widehat{R}(h)$. To do so, we need to bound $\mathbb{E}_S\left[\phi(S)\right]$, where S is a random subset of X of size m, and $|\phi(S) - \phi(S')|$ where S and S' are samples differing by exactly one point.

Lemma 1. *Let H be a bounded hypothesis set ($\forall x \in X, |h(x) - y(x)| \leq B$) and L a β-stable algorithm returning the hypotheses h and h' for two training sets S and S of size m each, respectively, differing in exactly one point. Then,*

$$|\phi(S) - \phi(S')| \leq 2\beta + B^2(m+u)/(mu). \tag{2}$$

Proof. By definition, S and S' differ exactly in one point. Let $x_i \in S$, $x_{m+j} \in S'$ be the points in which the two sets differ. The lemma follows from the observation that for each one of the $m-1$ common labeled points in S and S', and for each one of the $u-1$ common test points in T and T' (recall $T = X \setminus S$, $T' = X \setminus S'$), the difference in cost is bounded by β, while for x_i and x_{m+j}, the difference in cost is bounded by B^2. Then, it follows that $|\phi(S) - \phi(S')| \leq \frac{(u-1)\beta}{u} + \frac{(m-1)\beta}{m} + \frac{B^2}{u} + \frac{B^2}{m} \leq 2\beta + B^2\left(\frac{1}{u} + \frac{1}{m}\right)$. □

Lemma 2. *Let h be the hypothesis returned by a β-stable algorithm L. Then, $|\mathbb{E}_S\left[\phi(S)\right]| \leq \beta$.*

Proof. By definition of $\phi(S)$, its expectation is $\frac{1}{u}\sum_{k=1}^u \mathbb{E}_S\left[c(h, x_{m+k})\right] - \frac{1}{m}\sum_{k=1}^m \mathbb{E}_S\left[c(h, x_k)\right]$. Since $\mathbb{E}_S\left[c(h, x_{m+j})\right]$ is the same for all $j \in [1,u]$, and $\mathbb{E}_S\left[c(h, x_i)\right]$ the same for all $i \in [1,m]$, for any i and j, $\mathbb{E}_S\left[\phi(S)\right] = \mathbb{E}_S\left[c(h, x_{m+j})\right] - \mathbb{E}_S\left[c(h, x_i)\right] = \mathbb{E}_{S'}\left[c(h', x_i)\right] - \mathbb{E}_S\left[c(h, x_i)\right]$. Thus, $\mathbb{E}_S\left[\phi(S)\right] = \mathbb{E}_{S,S' \sim X}\left[c(h', x_i) - c(h, x_i)\right] \leq \beta$. □

Theorem 3. *Let H be a bounded hypothesis set ($\forall x \in X, |h(x) - y(x)| \leq B$) and L a β-stable algorithm. Let h be the hypothesis returned by L when trained on $X \vdash (S,T)$. Then, for any $\delta > 0$, with prob. at least $1-\delta$,*

$$R(h) \leq \widehat{R}(h) + \beta + \left(2\beta + \frac{B^2(m+u)}{mu}\right)\sqrt{\frac{\alpha(m,u)\ln\frac{1}{\delta}}{2}}.$$

Proof. The result follows directly from Theorem 2 and Lemmas 1 and 2. □

This is a general bound that applies to *any* transductive algorithm. To apply it, the stability coefficient β, which depends on m and u, needs to be determined. In the subsequent sections, we derive bounds on β for a number of transductive regression algorithms (Cortes

& Mohri, 2007; Belkin et al., 2004a; Wu & Schölkopf, 2007; Zhou et al., 2004; Zhu et al., 2003).

4. Stability of Local Transductive Regression Algorithms

This section describes and analyzes a general family of local transductive regression algorithms (LTR) generalizing the algorithm of Cortes and Mohri (2007).

LTR algorithms can be viewed as a generalization of the so-called kernel regularization-based learning algorithms to the transductive setting. The objective function that they minimize is of the form:

$$F(h,S) = \|h\|_K^2 + \frac{C}{m}\sum_{k=1}^{m} c(h,x_k) + \frac{C'}{u}\sum_{k=1}^{u} \widetilde{c}(h,x_{m+k}), \quad (3)$$

where $\|\cdot\|_K$ is the norm in the reproducing kernel Hilbert space (RKHS) with associated kernel K, $C \geq 0$ and $C' \geq 0$ are trade-off parameters, and $\widetilde{c}(h,x) = (h(x) - \widetilde{y}(x))^2$ is the error of the hypothesis h on the unlabeled point x with respect to a pseudo-target $\widetilde{y}$.

Pseudo-targets are obtained from neighborhood labels $y(x)$ by a local weighted average. Neighborhoods can be defined as a ball of radius r around each point in the feature space. We will denote by β_{loc} the score-stability coefficient of the local algorithm used, that is the maximal amount by which the two hypotheses differ on an given point, when trained on samples disagreeing on one point. This notion is stronger than that of cost-based stability.

In this section, we use the bounded-labels assumption, that is $\forall x \in S$, $|y(x)| \leq M$. We also assume that for any $x \in X$, $K(x,x) \leq \kappa^2$. We will use the following bound based on the reproducing property and the Cauchy-Schwarz inequality valid for any hypothesis $h \in H : \forall x \in X$,

$$|h(x)| = |\langle h, K(x,\cdot)\rangle| \leq \|h\|_K \sqrt{K(x,x)} \leq \kappa \|h\|_K. \quad (4)$$

Lemma 3. *Let h be the hypothesis minimizing (3). Assume that for any $x \in X$, $K(x,x) \leq \kappa^2$. Then, for any $x \in X$, $|h(x)| \leq \kappa M\sqrt{C+C'}$.*

Proof. The proof is a straightforward adaptation of the technique of (Bousquet & Elisseeff, 2002) to LTR algorithms. By Eqn. 4, $|h(x)| \leq \kappa \|h\|_K$. Let $\mathbf{0} \in \mathbb{R}^{m+u}$ be the hypothesis assigning label zero to all examples. By definition of h,

$$F(h,S) \leq F(\mathbf{0},S) \leq (C+C')M^2.$$

Using $\|h\|_K \leq \sqrt{F(h,S)}$ yields the statement. □

Since $|h(x)| \leq \kappa M\sqrt{C+C'}$, this immediately gives us a bound on $|h(x) - y(x)| \leq M(1 + \kappa\sqrt{C+C'})$. Thus, we are in a position to apply Theorem 3 with $B = AM$, $A = 1 + \kappa\sqrt{C+C'}$.

We now derive a bound on the stability coefficient β. To do so, the key property we will use is the convexity of $h \mapsto c(h,x)$. Note, however, that in the case of $\widetilde{c}$, the pseudo-targets may depend on the training set S. This dependency matters when we wish to apply convexity with two hypotheses h and h' obtained by training on different samples S and S'. For convenience, for any two such fixed hypotheses h and h', we extend the definition of $\widetilde{c}$ as follows. For all $t \in [0,1]$,

$$\widetilde{c}(th+(1-t)h',x) = \big((th+(1-t)h')(x) - (t\widetilde{y}+(1-t)\widetilde{y}')\big)^2.$$

This allows us to use the same convexity property for $\widetilde{c}$ as for c for any two fixed hypotheses h and h', as verified by the following lemma, and does not affect the proofs otherwise.

Lemma 4. *Let h be a hypothesis obtained by training on S and h' by training on S'. Then, for all $t \in [0,1]$,*

$$t\widetilde{c}(h,x) + (1-t)\widetilde{c}(h',x) \geq \widetilde{c}(th + (1-t)h', x). \quad (5)$$

Proof. Let $\widetilde{y} = \widetilde{y}(x)$ be the pseudo-target value at x when the training set is S and $\widetilde{y}' = \widetilde{y}'(x)$ when the training set is S'. For all $t \in [0,1]$,

$$\begin{aligned} &tc(h,x) + (1-t)c(h',x) - c(th+(1-t)h',x) \\ &= t(h(x)-\widetilde{y})^2 + (1-t)(h'(x)-\widetilde{y}')^2 \\ &- \left[t(h(x)-\widetilde{y}) + (1-t)(h'(x)-\widetilde{y}')\right]^2. \end{aligned}$$

The statement of the lemma follows directly by the convexity of $x \mapsto x^2$ over real numbers. □

Let h be a hypothesis obtained by training on S and h' by training on S'. Let $\Delta = h - h'$. Then, for all $x \in X$, $|c(h,x) - c(h',x)| = |\Delta(x)\left((h(x)-y(x)) + (h'(x)-y(x))\right)| \leq 2M(1 + \kappa\sqrt{C+C'})|\Delta(x)|$. As in 4, for all $x \in X$, $|\Delta(x)| \leq \kappa\|\Delta\|_K$, thus for all $x \in X$,

$$|c(h,x) - c(h',x)| \leq 2M(1+\kappa\sqrt{C+C'})\kappa\|\Delta\|_K. \quad (6)$$

Lemma 5. *Assume that for all $x \in X$, $|y(x)| \leq M$. Let S and S' be two samples differing by exactly one point. Let h be the hypothesis returned by the algorithm minimizing the objective function $F(h,S)$, h' be the hypothesis obtained by minimization of $F(h,S')$ and let $\widetilde{y}$ and $\widetilde{y}'$ be the corresponding pseudo-targets. Then,*

$$\begin{aligned} &C\left[c(h',x_i) - c(h,x_i)\right]/m - C'\left[\widetilde{c}(h',x_i) - \widetilde{c}(h,x_i)\right]/u \\ &\leq 2AM\left(\kappa\|\Delta\|_K\left(C/m + C'/u\right) + \beta_{loc}C'/u\right). \end{aligned}$$

where $\Delta = h' - h$ and $A = 1 + \kappa\sqrt{C+C'}$.

Proof. Let $\tilde{c}(h_i, \tilde{y}_i) = \tilde{c}(h, x_i)$ and $\tilde{c}(h'_i, \tilde{y}'_i) = \tilde{c}(h', x_i)$. By Lemma 3 and the bounded-labels assumption,

$$\begin{aligned}
&|\widetilde{c}(h'_i, \widetilde{y}'_i) - \widetilde{c}(h_i, \widetilde{y}_i)| \\
&= |\widetilde{c}(h'_i, \widetilde{y}'_i) - \widetilde{c}(h'_i, \widetilde{y}_i) + \widetilde{c}(h'_i, \widetilde{y}_i) - \widetilde{c}(h_i, \widetilde{y}_i)| \\
&\leq |(\widetilde{y}'_i - \widetilde{y}_i)(\widetilde{y}'_i + \widetilde{y}_i - 2h'_i)| + |(h'_i - h_i)(h'_i + h_i - 2\widetilde{y}_i)|\,.
\end{aligned}$$

By the score-stability of local estimates, $\widetilde{y}'(x_i) - \widetilde{y}(x_i) \leq \beta_{loc}$. Thus,

$$|\widetilde{c}(h'_i, \widetilde{y}'_i) - \widetilde{c}(h_i, \widetilde{y}_i)| \leq 2AM(\beta_{loc} + \kappa\|\Delta\|_K). \quad (7)$$

Using 6 leads after simplification to the statement of the lemma. □

The proof of the following theorem is based on Lemma 4 and Lemma 5 and is reserved to a longer version of this paper.

Theorem 4. *Assume that for all $x \in X$, $|y(x)| \leq M$ and there exists κ such that $\forall x \in X$, $K(x,x) \leq \kappa^2$. Further, assume that the local estimator has uniform stability coefficient β_{loc}. Let $A = 1+\kappa\sqrt{C+C'}$. Then,* `LTR` *is uniformly β-stable with*

$$\beta \leq 2(AM)^2\kappa^2\left[\frac{C}{m}+\frac{C'}{u}+\sqrt{\left(\frac{C}{m}+\frac{C'}{u}\right)^2+\frac{2C'\beta_{loc}}{AM\kappa^2 u}}\right].$$

Our experiments with `LTR` will demonstrate the benefit of this bound for model selection (Sec. 6).

5. Stability Based on Closed-Form Solutions

5.1. Unconstrained Regularization Algorithms

In this section, we consider a family of transductive regression algorithms that can be formulated as the following optimization problem:

$$\min_{\mathbf{h}} \mathbf{h}^T\mathbf{Q}\mathbf{h} + (\mathbf{h}-\mathbf{y})^T\mathbf{C}(\mathbf{h}-\mathbf{y}). \quad (8)$$

$\mathbf{Q} \in \mathbb{R}^{(m+u)\times(m+u)}$ is a symmetric regularization matrix, $\mathbf{C} \in \mathbb{R}^{(m+u)\times(m+u)}$ is a symmetric matrix of empirical weights (in practice it is often a diagonal matrix), $\mathbf{y} \in \mathbb{R}^{(m+u)\times 1}$ are the target values of the m labeled points together with the pseudo-target values of the u unlabeled points (in some formulations, the pseudo-target value is 0), and $\mathbf{h} \in \mathbb{R}^{(m+u)\times 1}$ is a column vector whose ith row is the predicted target value for the x_i. The closed-form solution of (8) is given by

$$\mathbf{h} = (\mathbf{C}^{-1}\mathbf{Q} + \mathbf{I})^{-1}\mathbf{y}. \quad (9)$$

The formulation (8) is quite general and includes as special cases the algorithms of (Belkin et al., 2004a; Wu & Schölkopf, 2007; Zhou et al., 2004; Zhu et al., 2003). We present a general framework for bounding the stability coefficient of these algorithms and then examine the stability coefficient of each of these algorithms in turn.

For a symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ we will denote by $\lambda_M(\mathbf{A})$ its largest eigenvalue and $\lambda_m(\mathbf{A})$ its smallest. Then, for any $\mathbf{v} \in \mathbb{R}^{n\times 1}$, $\lambda_m(\mathbf{A})\|\mathbf{v}\|_2 \leq \|\mathbf{A}\mathbf{v}\|_2 \leq \lambda_M(\mathbf{A})\|\mathbf{v}\|_2$. We will also use in the proof of the following proposition the fact that for symmetric matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$, $\lambda_M(\mathbf{AB}) \leq \lambda_M(\mathbf{A})\lambda_M(\mathbf{B})$.

Proposition 1. *Let $\mathbf{h}$ and $\mathbf{h}'$ solve (8), under test and training sets that differ exactly in one point and let $\mathbf{C}, \mathbf{C}', \mathbf{y}, \mathbf{y}'$ be the analogous empirical weight and the target value matrices. Then,*

$$\|\mathbf{h}'-\mathbf{h}\|_2 \leq \frac{\|\mathbf{y}'-\mathbf{y}\|_2}{\frac{\lambda_m(\mathbf{Q})}{\lambda_M(\mathbf{C})}+1} + \frac{\lambda_M(\mathbf{Q})\|\mathbf{C}'^{-1}-\mathbf{C}^{-1}\|_2\,\|\mathbf{y}\|_2}{\left(\frac{\lambda_m(\mathbf{Q})}{\lambda_M(\mathbf{C}')}+1\right)\left(\frac{\lambda_m(\mathbf{Q})}{\lambda_M(\mathbf{C})}+1\right)}.$$

Proof. Let $\Delta = \mathbf{h}' - \mathbf{h}$ and $\Delta\mathbf{y} = \mathbf{y}' - \mathbf{y}$. Let $\mathfrak{C} = (\mathbf{C}^{-1}\mathbf{Q} + \mathbf{I})$ and $\mathfrak{C}' = (\mathbf{C}'^{-1}\mathbf{Q} + \mathbf{I})$. By definition,

$$\begin{aligned}
\Delta &= \mathfrak{C}'^{-1}\mathbf{y}' - \mathfrak{C}^{-1}\mathbf{y} \\
&= \mathfrak{C}'^{-1}\Delta\mathbf{y} + (\mathfrak{C}'^{-1} - \mathfrak{C}^{-1})\mathbf{y} \\
&= \mathfrak{C}'^{-1}\Delta\mathbf{y} + (\mathfrak{C}^{-1}\left[(\mathbf{C}^{-1} - \mathbf{C}'^{-1})\mathbf{Q}\right]\mathfrak{C}'^{-1})\mathbf{y}.
\end{aligned}$$

$$\text{Thus, } \|\Delta\|_2 \leq \frac{\|\Delta\mathbf{y}\|_2}{\lambda_m(\mathfrak{C})} + \frac{\lambda_M(\mathbf{Q})\|\mathbf{C}'^{-1}-\mathbf{C}^{-1}\|_2 \cdot \|\mathbf{y}\|_2}{\lambda_m(\mathfrak{C}')\lambda_m(\mathfrak{C})}. \quad (10)$$

Furthermore, $\lambda_m(\mathfrak{C}) \geq \frac{\lambda_m(\mathbf{Q})}{\lambda_M(\mathbf{C})}+1$. Plugging this bound back into Eqn. 10 yields:

$$\|\Delta\|_2 \leq \frac{\|\Delta\mathbf{y}\|_2}{\frac{\lambda_m(\mathbf{Q})}{\lambda_M(\mathbf{C})}+1} + \frac{\lambda_M(\mathbf{Q})\|\mathbf{C}'^{-1}-\mathbf{C}^{-1}\|_2\|\mathbf{y}\|_2}{\left(\frac{\lambda_m(\mathbf{Q})}{\lambda_M(\mathbf{C}')}+1\right)\left(\frac{\lambda_m(\mathbf{Q})}{\lambda_M(\mathbf{C})}+1\right)}. \quad \square$$

Since $\|\mathbf{h}'-\mathbf{h}\|_\infty$ is bounded by $\|\mathbf{h}'-\mathbf{h}\|_2$, the proposition provides a bound on the score-stability of $\mathbf{h}$ for the transductive regression algorithms of Zhou et al. (2004); Wu and Schölkopf (2007); Zhu et al. (2003). For each of these algorithms, the pseudo-targets used are zero. If we make the bounded labels assumption ($\forall x \in X, |y(x)| \leq M$, for some $M > 0$), it is not difficult to show that $\|\mathbf{y}-\mathbf{y}'\|_2 \leq \sqrt{2}M$ and $\|\mathbf{y}\|_2 \leq \sqrt{m}M$. We now examine each algorithm in turn.

Consistency method (`CM`) In the `CM` algorithm (Zhou et al., 2004), the matrix $\mathbf{Q}$ is a normalized Laplacian of a weight matrix $\mathbf{W} \in \mathbb{R}^{(m+u)\times(m+u)}$ that captures affinity between pairs of points in the full sample X. Thus, $\mathbf{Q} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$, where $\mathbf{D} \in \mathbb{R}^{(m+u)\times(m+u)}$ is a diagonal matrix, with $[\mathbf{D}]_{i,i} =$

$\sum_j [\mathbf{W}]_{i,j}$. Note that $\lambda_m(\mathbf{Q}) = 0$. Furthermore, matrices $\mathbf{C}$ and $\mathbf{C}'$ are identical in CM, both diagonal matrices with (i,i)th entry equal to a positive constant $\mu > 0$. Thus $\mathbf{C}^{-1} = \mathbf{C}'^{-1}$ and using Prop. 1, we obtain the following bound on the score-stability of the CM algorithm: $\beta_{\mathtt{CM}} \leq \sqrt{2}M$.

Local learning regularization ($\mathtt{LL-Reg}$) In the $\mathtt{LL-Reg}$ algorithm (Wu & Schölkopf, 2007), the regularization matrix $\mathbf{Q}$ is $(\mathbf{I}-\mathbf{A})^T(\mathbf{I}-\mathbf{A})$, where $\mathbf{I} \in \mathbb{R}^{(m+u)\times(m+u)}$ is an identity matrix and $\mathbf{A} \in \mathbb{R}^{(m+u)\times(m+u)}$ is a non-negative weight matrix that captures the local similarity between all pairs of points in X. $\mathbf{A}$ is normalized, i.e. each of its rows sum to 1. Let $C_l, C_u > 0$ be two positive constants. The matrix $\mathbf{C}$ is a diagonal matrix with $[\mathbf{C}]_{i,i} = C_l$ if $x_i \in S$ and C_u otherwise. Let $C_{\max} = \max\{C_l, C_u\}$ and $C_{\min} = \min\{C_l, C_u\}$. Thus, $\|\mathbf{C}'^{-1} - \mathbf{C}^{-1}\|_2 = \sqrt{2}\left(\frac{1}{C_{\min}} - \frac{1}{C_{\max}}\right)$. By the Perron-Frobenius theorem, its eigenvalues lie in the interval $(-1, 1]$ and $\lambda_M(\mathbf{A}) \leq 1$. Thus, $\lambda_m(\mathbf{Q}) \geq 0$ and $\lambda_M(\mathbf{Q}) \leq 4$ and we have the following bound on the score-stability of the $\mathtt{LL-Reg}$ algorithm: $\beta_{\mathtt{LL-Reg}} \leq \sqrt{2}M + 4\sqrt{m}M\left(\frac{1}{C_{\min}} - \frac{1}{C_{\max}}\right) \leq \sqrt{2}M + \frac{4\sqrt{m}M}{C_{\min}}$.

Gaussian Mean Fields algorithm GMF (Zhu et al., 2003) is very similar to the $\mathtt{LL-Reg}$, and admits exactly the same stability coefficient.

Thus, the stability coefficients of the algorithms of $\mathtt{CM}, \mathtt{LL-Reg}$, and GMF can be large. Without additional constraints on the matrix $\mathbf{Q}$, these algorithms do not seem to be stable enough for the generalization bound of Theorem 3 to converge. A particular example of constraint is the condition $\sum_{i=1}^{m+u} h(x_i) = 0$ used by Belkin et al.'s algorithm (2004a). In the next section, we give a generalization bound for this algorithm and then describe a general method for making the algorithms just examined stable.

5.2. Stability of Constrained Regularization Algorithms

This subsection analyzes constrained regularization algorithms such as the Laplacian-based graph regularization algorithm of Belkin et al. (2004a). Given a weighted graph $G = (X, E)$ in which edge weights represent the extent of similarity between vertices, the task consists of predicting the vertex labels. The hypothesis h returned by the algorithm is solution of the following optimization problem:

$$\min_{h \in H} \mathbf{h}^T \mathbf{L} \mathbf{h} + \frac{C}{m} \sum_{i=1}^{m} (h(x_i) - y_i)^2 \quad \text{subject to: } \sum_{i=1}^{m+u} h(x_i) = 0, \tag{11}$$

where $\mathbf{L} \in \mathbb{R}^{(m+u)\times(m+u)}$ is a smoothness matrix, e.g., the graph Laplacian, $\{y_i \mid i \in [1,m]\}$ are the target values of the m labeled nodes.

The hypothesis set H in this case can be thought of as a hyperplane in $\mathbb{R}^{m+u}$ that is orthogonal to the vector $\mathbf{1} \in \mathbb{R}^{m+u}$. Maintaining the notation used in (Belkin et al., 2004a), we let P_H denote the operator corresponding to the orthogonal projection on H. For a sample S drawn without replacement from X, define $\mathbf{I}_S \in \mathbb{R}^{(m+u)\times(m+u)}$ to be the diagonal matrix with $[\mathbf{I}_S]_{i,i} = 1$ if $x_i \in S$ and 0 otherwise. Similarly, let $\mathbf{y}_S \in \mathbb{R}^{(m+u)\times 1}$ be the column vector with $[\mathbf{y}_S]_{i,1} = y_i$ if $x_i \in S$ and 0 otherwise. The closed-form solution on a training sample S is given by (Belkin et al., 2004a):

$$\mathbf{h}_S = \left(P_H\left(\frac{m}{C}\mathbf{L} + \mathbf{I}_S\right)\right)^{-1} \mathbf{y}_S. \tag{12}$$

Theorem 5. *Assume that the vertex labels of the graph $G = (X, E)$ and the hypothesis h obtained by optimizing Eqn. 11 are both bounded ($\forall x, |h(x)| \leq M$ and $|y(x)| \leq M$ for some $M > 0$). Let $A = 1 + \kappa\sqrt{C}$. Then, for any $\delta > 0$, with probability at least $1-\delta$,*

$$R(h) \leq \widehat{R}(h) + \beta + \left(2\beta + \frac{(AM)^2(m+u)}{mu}\right)\sqrt{\frac{\alpha(m,u)\ln\frac{1}{\delta}}{2}},$$

with $\alpha(m,u) = \frac{mu}{m+u-1/2} \cdot \frac{1}{1-1/(2\max\{m,u\})}$ and $\beta \leq (4\sqrt{2}M^2)/(m\lambda_2/C - 1) + (4\sqrt{2m}M^2)/(m\lambda_2/C - 1)^2$, λ_2 is the second smallest eigenvalue of the Laplacian.

Proof. The proof is similar to that of (Belkin et al., 2004a) but uses our general transductive regression bound instead. □

The generalization bound we just presented differs in several respects from that of Belkin et al. (2004a). Our bound explicitly depends on both m and u while theirs shows only a dependency on m. Also, our bound does not depend on the number of times a point is sampled in the training set (parameter t), thanks to our analysis based on sampling without replacement.

Contrasting the stability coefficient of Belkin's algorithm with the stability coefficient of LTR (Theorem 4), we note that it does not depend on C' and β_{loc}. This is because unlabeled points do not enter the objective function, and thus $C' = 0$ and $\widetilde{y}(x) = 0$ for all

$x \in X$. However, the stability does depend on the second smallest eigenvalue λ_2 and the bound diverges as λ_2 approaches $\frac{C}{m}$. In all our regression experiments, we observed that this algorithm does not perform as well in comparison with LTR.

5.3. Making Seemingly Unstable Algorithms Stable

In Sec. 5.2, we saw that imposing additional constraints on the hypothesis, e.g., $\mathbf{h} \cdot \mathbf{1} = 0$, allowed one to derive non-trivial stability bounds. This idea can be generalized and similar non-trivial stability bounds can be derived for "stable" versions of the algorithms presented in Sec. 5.1 CM, LL − Reg, and GMF. Recall that the stability bound in Prop. 1 is inversely proportional to the smallest eigenvalue $\lambda_m(\mathbf{Q})$. The main difficulty with using the proposition for these algorithms is that $\lambda_m(\mathbf{Q}) = 0$ in each case. Let $\mathbf{v}_m$ denote the eigenvector corresponding to $\lambda_m(\mathbf{Q})$ and let λ_2 be the second smallest eigenvalue of $\mathbf{Q}$. One can modify (8) and constrain the solution to be orthogonal to $\mathbf{v}_m$ by imposing $\mathbf{h} \cdot \mathbf{v}_m = 0$. In the case of (Belkin et al., 2004a), $\mathbf{v}_m = \mathbf{1}$. This modification, motivated by the algorithm of (Belkin et al., 2004a), is equivalent to increasing the smallest eigenvalue to be λ_2.

As an example, by imposing the additional constraint, we can show that the stability coefficient of CM becomes bounded by $O(C/\lambda_2)$, instead of $\Theta(1)$. Thus, if $C = O(1/m)$ and $\lambda_2 = \Omega(1)$, it is bounded by $O(1/m)$ and the generalization bound converges as $O(1/m)$.

6. Experiments

6.1. Model Selection Based on Bound

This section reports the results of experiments using our stability-based generalization bound for model selection for the LTR algorithm. A crucial parameter of this algorithm is the stability coefficient $\beta_{loc}(r)$ of the local algorithm, which computes pseudo-targets $\widetilde{y}_x$ based on a ball of radius r around each point. We derive an expression for $\beta_{loc}(r)$ and show, using extensive experiments with multiple data sets, that the value r^* minimizing the bound is a remarkably good estimate of the best r for the test error. This demonstrates the benefit of our generalization bound for model selection, avoiding the need for a held-out validation set.

The experiments were carried out on several publicly available regression data sets: *Boston Housing*, *Elevators* and *Ailerons*[2]. For each of these data sets, we used $m = u$, inspired by the observation that, all other parameters being fixed, the bound of Theorem 3 is tightest when $m = u$. The value of the input variables were normalized to have mean zero and variance one. For the Boston Housing data set, the total number of examples was 506. For the Elevators and the Ailerons data set, a random subset of 2000 examples was used. For both of these data sets, other random subsets of 2000 samples led to similar results. The Boston Housing experiments were repeated for 50 random partitions, while for the Elevators and the Ailerons data set, the experiments were repeated for 20 random partitions each. Since the target values for the Elevators and the Ailerons data set were extremely small, they were scaled by a factor 1000 and 100 respectively in a pre-processing step.

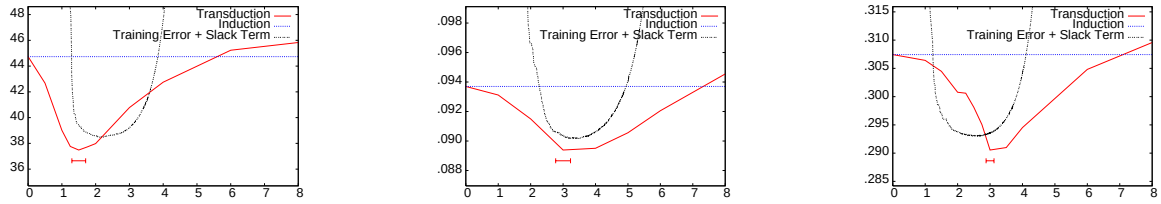
In our experiments, we estimated the pseudo-target of a point $x' \in T$ as a weighted average of the labeled points $x \in N(x')$ in a neighborhood of x'. Thus, $\widetilde{y}_{x'} = \sum_{x \in N(x')} \alpha_x y_x / \sum_{x \in N(x')} \alpha_x$. Weights are defined in terms of a similarity measure $K(x, x')$ captured by a kernel K: $\alpha_x = K(x, x')$. Let $m(r)$ be the number of labeled points in $N(x')$. Then, it is easy to show that $\beta_{loc} \leq 4\alpha_{\max} M/(\alpha_{\min} m(r))$, where $\alpha_{\max} = \max_{x \in N(x')} \alpha_x$ and $\alpha_{\min} = \min_{x \in N(x')} \alpha_x$. Thus, for a Gaussian kernel with parameter σ, $\beta_{loc} \leq 4M/(m(r)e^{-2r^2/\sigma^2})$. To estimate β_{loc}, one needs an estimate of $m(r)$, the number of samples in a ball of radius r from an unlabeled point x'. In our experiments, we estimated $m(r)$ as the number of samples in a ball of radius r from the origin. Since all features are normalized to mean zero and variance one, the origin is also the centroid of the set X.

We implemented a dual solution of LTR and used Gaussian kernels, for which, the parameter σ was selected using cross-validation on the training set. Experiments were repeated across 36 different pairs of values of (C, C'). For each pair, we varied the radius r of the neighborhood used to determine estimates from zero to the radius of the ball containing all points.

Figure 1(a) shows the mean values of the test MSE of our experiments on the Boston Housing data set for typical values of C and C'. Figures 1(b)-(c) show similar results for the Ailerons and Elevators data sets. For the sake of comparison, we also report results for induction. The relative standard deviations on the MSE are not indicated, but were typically of the order of 10%. LTR generally achieves a significant improvement over induction.

The generalization bound we derived in Eqn. 3 consists of the training error and a complexity term that depends on the parameters of the LTR algorithm $(C, C', M, m, u, \kappa, \beta_{loc}, \delta)$. Only two terms depend

[2] www.liaad.up.pt/~ltorgo/Regression/DataSets.html.



(a) (b) (c)

Figure 1. MSE against the radius r of LTR for three data sets: (a) Boston Housing. (b) Ailerons. (c) Elevators. The small horizontal bar indicates the location (mean ± one standard deviation) of the minimum of the empirically determined r.

upon the choice of the radius r: $\widehat{R}(h)$ and β_{loc}. Thus, keeping all other parameters fixed, the theoretically optimal radius r^* is the one that minimizes the training error plus the slack term. The figures also include plots of the training error combined with the complexity term, appropriately scaled. The empirical minimization of the radius r coincides with or is close to r^*. The optimal r based on test MSE is indicated with error bars.

6.2. Stable Versions of Unstable Algorithms

We refer to the stable version of the CM algorithm presented in Sec. 5.1 as CM − STABLE. We compared CM and CM − STABLE empirically on the same datasets, again using $m = u$. For the normalized Laplacian we used k-nearest neighbors graphs based on Euclidean distance. The parameters k and C were chosen by five-fold cross-validation over the training set. The experiment was repeated 20 times with random partitions. The averaged mean-squared errors with standard deviations, are reported in Table 6.2.

Dataset	CM	CM − STABLE
Elevators	0.3228 ± 0.0264	0.3293 ± 0.0286
Ailerons	0.1149 ± 0.0081	0.1184 ± 0.0087
Housing	57.93 ± 6.5	57.92 ± 6.5

We conclude from this experiment that CM and CM − STABLE have the same performance. However, as we showed previously, CM − STABLE has a non-trivial risk bound and thus comes with some guarantee.

7. Conclusion

We presented a comprehensive analysis of the stability of transductive regression algorithms with novel generalization bounds for a number of algorithms. Since they are algorithm-dependent, our bounds are often tighter than those based on complexity measures such as the VC-dimension. Our experiments also show the effectiveness of our bounds for model selection and the good performance of LTR algorithms.

References

Belkin, M., Matveeva, I., & Niyogi, P. (2004a). Regularization and semi-supervised learning on large graphs. *COLT* (pp. 624–638).

Belkin, M., Niyogi, P., & Sindhwani, V. (2004b). *Manifold regularization* (Technical Report TR-2004-06). University of Chicago.

Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. *JMLR*, *2*, 499–526.

Chapelle, O., Vapnik, V., & Weston, J. (1999). Transductive Inference for Estimating Values of Functions. *NIPS 12* (pp. 421–427).

Cortes, C., & Mohri, M. (2007). On Transductive Regression. *NIPS 19* (pp. 305–312).

El-Yaniv, R., & Pechyony, D. (2006). Stable transductive learning. *COLT* (pp. 35–49).

McDiarmid, C. (1989). On the method of bounded differences. *Surveys in Combinatorics* (pp. 148–188). Cambridge University Press, Cambridge.

Schuurmans, D., & Southey, F. (2002). Metric-Based Methods for Adaptive Model Selection and Regularization. *Machine Learning*, *48*, 51–84.

Vapnik, V. N. (1982). *Estimation of dependences based on empirical data*. Berlin: Springer.

Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley-Interscience.

Wu, M., & Schölkopf, B. (2007). Transductive classification via local learning regularization. *AISTATS* (pp. 628–635).

Zhou, D., Bousquet, O., Lal, T., Weston, J., & Schölkopf, B. (2004). Learning with local and global consistency. *NIPS 16* (pp. 595–602).

Zhu, X., Ghahramani, Z., & Lafferty, J. (2003). Semi-supervised learning using gaussian fields and harmonic functions. *ICML* (pp. 912–919).