A Sequential Approximation Bound for Some Sample-Dependent Convex Optimization Problems with Applications in Learning

Tong Zhang

IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
tzhang@watson.ibm.com

Abstract. In this paper, we study a class of sample dependent convex optimization problems, and derive a general sequential approximation bound for their solutions. This analysis is closely related to the regret bound framework in online learning. However we apply it to batch learning algorithms instead of online stochastic gradient decent methods. Applications of this analysis in some classification and regression problems will be illustrated.

1 Introduction

An important aspect of a machine learning algorithm is its generalization ability. In the batch learning framework, an algorithm obtains a hypothesis from a finite number of training data. The generalization ability is measured by the accuracy of the learned hypothesis when it is tested on some previously unobserved data.

A popular method to derive generalization bounds is the so-called Vapnik-Chervonenkis (VC) style analysis [11]. This method depends on the uniform convergence of observed errors of the hypothesis family to their true errors. The rate of uniform convergence depends on an estimate of certain sample-dependent covering numbers (growth numbers) for the underlying hypothesis family. Although this framework is quite general and powerful, it also has many disadvantages. For example, the derived generalization bounds are often very loose.

Because of various disadvantages of VC analysis, other methods to estimate generalization performance have been introduced. In this paper, we propose a new style of analysis that is suitable for certain sample dependent convex optimization problems. This type of bounds are closely related to the leave-one-out analysis, which has received much attention recently. For example, see [3, 7, 8, 13] and some references therein. However, instead of estimating the leave-one-out cross-validation error, we estimate the convergence of the estimated parameter averaged over a sequence of data. This is closely related to the online regret bound framework (for example, see [1, 9]). However, we study the learning problems in batch setting. Another important technical difference is that since an

explicit regularization condition is used in a batch-form sample-dependent optimization formulation, we can avoid the limitation of "matching loss" and the "learning rate" parameter which requires to be adjusted in online learning analysis.

Our analysis also indicates that even though some gradient descent type online learning algorithms achieve good worst-case regret bounds, in practice they could still be inferior to the corresponding batch algorithms. This also justifies why practitioners apply an online algorithm repeatedly over the training data so that it effectively converges to the solution of a sample dependent optimization problem, although the online mistake bound analysis implies that this is not helpful. In addition, the sequential approximation analysis complements the leave-one-out analysis in batch learning. In many cases it can give better bounds than those from the leave-one-out analysis. The latter analysis may not yield bounds that are asymptotically tight.

We organize the paper as follows. In Section 2, we prove a sequential approximation bound for a class of sample dependent optimization problems. This bound is the foundation of our analysis. Section 3 applies this bound to a general formulation of linear learning machines. Section 4 and Section 5 contain specific results of this analysis on some classification and regression problems. Concluding remarks are given in Section 6.

2 A generic sequential approximation bound

In many machine learning problems, we are given a training set of input variable x and output variable y. Our goal is to find a function that can predict y based on x. Typically, one needs to restrict the hypothesis function family size so that a stable estimate within the function family can be obtained from a finite number of samples. We assume that the function family can be specified by a vector parameter $w \in H$, where H is a Hilbert space. The inner product of two vectors $w_1, w_2 \in H$ is denoted by $w_1^T w_2$. We also let $w^2 = w^T w$ and $\|w\| = (w^T w)^{1/2}$.

We consider a "learning" algorithm that determines a parameter estimate w_n from training samples $(x_1, y_1), \ldots, (x_n, y_n)$ by solving the following sample-dependent optimization problem:

$$w_n = \arg\min_w w^2 \tag{1}$$

$$\text{s.t.} \quad w \in C_n(x_1, y_1, \ldots, x_n, y_n). \tag{2}$$

We assume that C_n is a sample-dependent weakly closed convex set in H. That is,

- $\forall\, w \in H$: if $\exists$ sequence $\{w_i\}_{i=1,2,\ldots} \in C_n$ such that $\lim_{i\to\infty} w_i^T x = w^T x$ for all $x \in H$,[1] then $w \in C_n$.
- $\forall w_1, w_2 \in C_n$ and $\theta \in [0,1]$, we have $\theta w_1 + (1-\theta) w_2 \in C_n$.

[1] We say that the sequence $\{w_i\}$ converges weakly to w.

Under the above assumptions, the optimization problem (1) becomes a convex programming problem. The following proposition shows that it has a unique solution.

Proposition 1. *If $C_n(x_1, y_1, \ldots, x_n, y_n)$ is non-empty, then optimization problem (1) has a unique solution that belongs to $C_n(x_1, y_1, \ldots, x_n, y_n)$.*

Proof. Since C_n is non-empty, there exists a sequence $\{w_i\}_{i=1,2,\ldots}$ such that $\lim_{i\to\infty} w_i^2 = \inf_{w\in C_n} w^2$. Note that since the sequence $\{w_i^2\}$ converges, it is bounded. Therefore it contains a weakly convergent subsequence (cf. Proposition 66.4 in [6]). Without loss of generality, we assume the weakly convergent subsequence is the sequence $\{w_i\}$ itself. Denote its weak limit by w_*, then by the weakly closedness of C_n, we have $w_* \in C_n$. Also

$$w_*^2 = \lim_{i\to\infty} w *^T w_i \leq (w_*^2 \lim_{i\to\infty} w_i^2)^{1/2} \leq (w_*^2 \inf_{w\in C_n} w^2)^{1/2} \leq w_*^2.$$

This implies that w_* is a solution of (1).

To see that the solution is unique, we simply assume that there are two solutions denoted by $w_1 \in C_n$ and $w_2 \in C_n$. Note that $0.5w_1 + 0.5w_2 \in C_n$ by the convexity of C_n. We thus have $(0.5w_1 + 0.5w_2)^2 \geq 0.5w_1^2 + 0.5w_2^2$ by the definition of w_1 and w_2 as solutions of (1). This inequality is satisfied only when $w_1 = w_2$.

The following lemma, although simple to prove, is the foundation of our analysis.

Lemma 1. *Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a sequence of observations. Assume that*

$$C_n(x_1, y_1, \ldots, x_n, y_n) \subseteq C_{n-1}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}).$$

Let w_k be the solution of (1) with respect to samples $(x_1, y_1), \ldots, (x_k, y_k)$ where $(k = n-1, n)$,[2] then if C_n is non-empty, we have the following one-step approximation bound:

$$(w_n - w_{n-1})^2 \leq w_n^2 - w_{n-1}^2.$$

Proof. Since $w_n^2 = w_{n-1}^2 + (w_n - w_{n-1})^2 + 2(w_n - w_{n-1})^T w_{n-1}$, to prove the lemma we only need to show $(w_n - w_{n-1})^T w_{n-1} \geq 0$. If this is not true, then assume $z = (w_n - w_{n-1})^T w_{n-1} < 0$. Let $\theta = \min(1, -z/(w_n - w_{n-1})^2)$, then $\theta \in (0, 1]$ and by the convexity of C_{n-1}, we know $w_{n-1} + \theta(w_n - w_{n-1}) \in C_{n-1}$. However,

$$\begin{aligned} & (w_{n-1} + \theta(w_n - w_{n-1}))^2 \\ &= w_{n-1}^2 + \theta^2(w_n - w_{n-1})^2 + 2\theta(w_n - w_{n-1})^T w_{n-1} \\ &\leq w_{n-1}^2 + \theta z < w_{n-1}^2, \end{aligned}$$

which contradicts the definition of w_{n-1}. Therefore the lemma holds.

[2] There is a little abuse of notation. We need to change the subscripts of n to k in (1) to define w_k. This convention, also used in later parts of the paper, should not cause any confusion.

Theorem 1. *Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a sequence of observations. Let $C_0 = H$, and assume that for all $k = m, \ldots, n$,*

$$C_k(x_1, y_1, \ldots, x_k, y_k) \subseteq C_{k-1}(x_1, y_1, \ldots, x_{k-1}, y_{k-1}).$$

Let w_k be the solution of (1) with respect to samples $(x_1, y_1), \ldots, (x_k, y_k)$ where $(k = m-1, \ldots, n)$, then we have the following sequential approximation bound:

$$\sum_{i=m}^{n} (w_i - w_{i-1})^2 \leq w_n^2 - w_{m-1}^2.$$

Proof. By Lemma 1, we have $(w_i - w_{i-1})^2 \leq w_i^2 - w_{i-1}^2$ for all $i = 1, \ldots, n$. Summing over $i = m, \ldots, n$, we obtain the theorem.

Note that the style of the above bound is similar to techniques widely used in online learning [1, 9, 4, 5]. However, the formulation we consider here is significantly different than what has been considered in the existing online learning literature. Furthermore, from a technical point of view, instead of bounding the regret loss as in online learning analysis, we directly bound the sum of squared distances of consecutive parameter estimates in a batch learning setting. Therefore our bound indicates the convergence of estimated parameter itself, which can then be used to bound the regret with respect to any loss function. The reason we can prove the convergence of parameter itself is due to our explicit use of regularization that minimizes w_n^2 in (1).

The concept of convergence of the estimated parameter has been widely used in traditional numerical mathematics and statistics. However, it has only recently been applied to analyzing learning problems. For example, techniques related to what we use here have also been applied in [12, 13]. The former leads to PAC style probability bounds, while the latter gives leave-one-out estimates. The convergence of the estimated parameter is also related to the algorithmic stability concept in [8]. However, the former condition is stronger. Consequently, better bounds can usually be obtained if we can show the convergence of the estimated parameter.

3 Linear Learning Methods

3.1 Linear learning formulations

To apply the general sequential approximation bound, we consider the linear prediction model where y is predicted as $y \approx w^T x$. We assume that $x \in H$ for all sample x. Given a training set of $(x_1, y_1), \ldots, (x_n, y_n)$, the parameter estimate w_n is obtained from (1) with the set C_n defined by the following type of constraints:

$$\begin{aligned} &C_n(x_1, y_1, \ldots, x_n, y_n) \\ &= \{w \in H : c_{n,k}(w^T x_1, x_1, y_1 \ldots, w^T x_n, x_n, y_n) \leq 0, \, (k = 1, \ldots, s_n)\}, \end{aligned} \quad (3)$$

where each $c_{n,k}$ is a continuous convex function of w.

Proposition 2. *The set C_n defined in (3) is convex and weakly closed.*

Proof. It is easy to check that the set C_n defined above is convex. C_n is also weakly closed since if a sequence $\{w_i\} \in C_n$ converges weakly to $w \in H$, then $\forall k$, by the continuity of $c_{n,k}$:

$$\begin{aligned} c_{n,k}(w^T x_1, x_1, y_1, \ldots, w^T x_n, x_n, y_n) &= c_{n,k}(\lim_i w_i^T x_1, x_1, y_1, \ldots, \lim_i w_i^T x_n, x_n, y_n) \\ &= \lim_i c_{n,k}(w_i^T x_1, x_1, y_1, \ldots, w_i^T x_n, x_n, y_n) \le 0. \end{aligned}$$

This means that $w \in C_n$.

For all concrete examples in this paper, we only consider the following functional form of $c_{n,k}$ in (3):

$$c_{n,k}(w^T x_1, x_1, y_1, \ldots, w^T x_n, x_n, y_n) = \sum_{i=1}^{n} f_{k,i}(w^T x_i, x_i, y_i),$$

where $f_{k,i}(a, b_1, b_2)$ is a continuous convex function of a. Specifically, we consider the following two choices of C_n. The first choice is

$$C_n = \{w \in H : a(x_i, y_i) \le w^T x_i \le b(x_i, y_i), \quad (i = 1, \ldots, n)\}. \tag{4}$$

Both $a(\cdot)$ and $b(\cdot)$ are functions that can take $\pm\infty$ as their values. The second choice is

$$C_n = \{w \in H : \sum_{i=1}^{k} L(w^T x_i, x_i, y_i) \le s\}, \tag{5}$$

where $L(a, b_1, b_2) \ge 0$ is a continuous convex function of a. $s \ge 0$ is a fixed parameter.

Clearly, either of the above choices of C_n satisfies the condition $C_k \subseteq C_{k-1}$. Hence Theorem 1 can be applied.

3.2 An equivalent formulation

From the numerical point of view, the parameter estimate w_n in (1) with C_n defined in (5) is closely related to the solution $\tilde{w}_n$ of the following penalized optimization formulation more commonly used in statistics:

$$\tilde{w}_n = \arg\min_{w \in H} [w^2 + C \sum_{i=1}^{k} L(w^T x_i, x_i, y_i)], \tag{6}$$

where $C > 0$ is a parameter. In fact, $\forall C$ in (6), if we let $s = \sum_{i=1}^{n} L(\tilde{w}^T x_i, x_i, y_i)$, then the solution w_n with C_n given in (5) is the same as $\tilde{w}_n$ in (6). To see this, just note that by the definition of w_n, we have $w_n^2 \le \tilde{w}_n^2$. Now compare (6) at $w = w_n$ and $w = \tilde{w}_n$, we obtain $w_n^2 \ge \tilde{w}_n^2$. This means that $\tilde{w}_n$ is the solution of (1) with C_n given in (5). Due to the uniqueness of solution, $w_n = \tilde{w}_n$.

This equivalence suggests that our analysis of the constrained formulation (1) with C_n defined in (5) can provide useful insights into the penalty type formulation (6). However in reality, there are some complications since typically the parameter C in (6) or s in (5) is determined by data-dependent cross-validation. A typical analysis either fixes C in (6) or fixes s in (5). These choices are not equivalent any more. An advantage of using (6) is that we do not need to worry about the feasibility condition (C_n is non-empty), although for many practical problems (even for problems with noise), the feasibility condition itself can be generally satisfied. The readers should be aware that although bounds given later in the paper assume that C_n is always non-empty, it is not difficult to generalize the bounds to handle the case where C_n may become empty with small probability.

There is no difficulty analyzing (6) directly using the same technique developed in this paper. We only need a slight generalization of Lemma 1 that allows a general penalized convex formulation in the objective function. Note that the proof of Lemma 1 essentially relies on the KKT condition of (1) at the optimal solution. In the more general situation, a similar KKT condition can be used to yield a desired inequality.

It is also easy to generalize the scheme to analyze non-square regularization conditions. Furthermore, by introducing slack variables (for example, this is done in the standard SVM formulation), it is not hard to rewrite general penalty type regularization formulations such as (6) as constrained regularization formulations such as (1), where we replace the minimization of w^2 by the minimization of an arbitrary convex function $g(w, \xi)$ of the weight vector w and the slack variable vector ξ. This provides a systematic approach to a very general learning formulation.

It is also possible to use a different technique to bound the squared distance of consecutive parameter estimates as in [13]. Although the method will yield a similar sequential approximation bound for penalty type formulation (6), it is not suitable for analyzing constrained formulation (1) which we study in this paper. In a related work, mistake bounds for some ridge-regression like online algorithms are derived in [4]. The resulting bounds are similar to what can be obtained by using our technique.

In this paper, we do not consider the general formulation which includes (6). We shall only mention that while our current approach is more suitable for the small noise situation (that is, s small, or equivalently C large), a direct analysis for (6) is more suitable for the large noise situation (that is, C small, or equivalently s large).

3.3 Kernel learning machines

Proposition 3. *The solution w_n of (1) with C_n defined in (3) belongs to X_n where X_n is the subspace of H that is spanned by x_i ($i = 1, \ldots, n$).*

Proof. Let $\bar{w}_n$ be the orthonormal projection of w_n onto X_n, then $\bar{w}_n^T x_i = w_n^T x_i$ for $i = 1, \ldots, n$. This implies that $\bar{w}_n \in C_n$. Since $\bar{w}_n^2 \leq w_n^2$, by Proposition 1, we have $w_n = \bar{w}_n$.

Under the assumption of Proposition 3, we can assume a representation of w as $w = \sum_{i=1}^n \alpha_i x_i$ in the optimization of (1). Using this representation, $w^T x = \sum_{i=1}^n \alpha_i x_i^T x$, and $w^2 = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j x_i^T x_j$. Therefore the only property we need to know about the Hilbert space H is a representation of its inner product $x^T y$. We may replace $x^T y$ by a symmetric positive-definite kernel function $K(x, y)$, which leads to a corresponding kernel method as follows:

$$\hat{\alpha} = \arg\min_{\alpha} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(x_i, x_j) \tag{7}$$

$$\text{s.t.} \quad c_{n,k}(\sum_{i=1}^n \alpha_i K(x_i, x_1), x_1, y_1 \ldots, \sum_{i=1}^n \alpha_i K(x_i, x_n), x_n, y_n) \le 0 \quad (k = 1, \ldots, s_n).$$

A properly behaved kernel function induces a Hilbert space (reproducing kernel Hilbert space) that consists of the closure of functions $f(x)$ of the form $\sum_i \alpha_i K(x_i, x)$. We can represent the functions linearly in a feature space as $f(x) = \sum_{i=1}^\infty w_i \phi_i(x)$. The inner product is $K(x, y) = \sum_{i=1}^\infty \lambda_i \phi_i(x) \phi_i(y)$, where $\lambda_i > 0$ are eigenvalues. See [2], chapter 3 for more details on kernel induced feature spaces. Our analysis in the reproducing kernel Hilbert space can thus be applied to study the general kernel method in (7).

4 Regression

4.1 Sequential-validation bounds

We consider regression problems. For simplicity, we only consider the q-norm loss with $1 \le q \le 2$. Our goal is to estimate w from the training data so that it has a small expected loss:

$$Q(w) = E_{(x,y)\sim D} |w^T x - y|^q,$$

where the expectation E is taken over an unknown distribution D. $1 \le q \le 2$ is a fixed parameter. The training samples (x_i, y_i) for $i = 1, \ldots, n$ are independently drawn from D.

Given the training data, we define the empirical expected loss as

$$Q_n(w, x_1, y_1, \ldots, x_n, y_n) = \frac{1}{n} \sum_{i=1}^n |w^T x_i - y_i|^q.$$

We use algorithm (1) to compute an estimate of w_n from the training data. We consider two formulations of C_n in (1). The first employs

$$C_n = \{w \in H : |w^T x_i - y_i| \le \epsilon(x_i, y_i), \quad (i = 1, \ldots, n)\}, \tag{8}$$

where $\epsilon(x, y) \ge 0$ is a pre-defined noise tolerance parameter. The second employs

$$C_n = \{w \in H : \sum_{i=1}^n |w^T x_i - y_i|^q \le s\}, \tag{9}$$

where s is a data independent parameter. If a solution does not exist in one of the above formulations (that is, C_n is empty), then we let $w_n = 0$. In the following, we only consider the case that C_n is always non-empty for clarity. However as we have mentioned earlier, it is possible to deal with the situation that C_n is empty with a small probability.

Theorem 2. *Assume for all training data C_{n+1} is non-empty, and $C_k \subseteq C_{k-1}$ for all $k = m, \ldots, n+1$. For each k, w_k is computed from samples $(x_1, y_1), \ldots, (x_k, y_k)$ using (1). We have the following sequential expected generalization bound ($1 \leq q \leq 2$):*

$$[\sum_{i=m}^{n} E\, Q(w_i)]^{1/q}$$
$$\leq E^{1/q} \sum_{i=m+1}^{n+1} Q_i(w_i, x_1, y_1, \ldots, x_i, y_i) + E^{1/q}\,[\|w_{n+1}\|^q (\sum_{i=m}^{n} \|x_{i+1}\|^{2q/(2-q)})^{(2-q)/2}].$$

The expectation E is with respect to $n+1$ independent random training samples $(x_1, y_1), \ldots, (x_{n+1}, y_{n+1})$ from D.

Proof. Consider training samples $(x_1, y_1), \ldots, (x_{n+1}, y_{n+1})$.

$$(\sum_{i=m}^{n} |w_i^T x_{i+1} - y_{i+1}|^q)^{1/q}$$
$$= (\sum_{i=m}^{n} |w_{i+1}^T x_{i+1} - y_{i+1} + (w_i - w_{i+1})^T x_{i+1}|^q)^{1/q}$$
$$\leq [\sum_{i=m}^{n} |w_{i+1}^T x_{i+1} - y_{i+1}|^q]^{1/q} + [\sum_{i=m}^{n} |(w_i - w_{i+1})^T x_{i+1}|^q]^{1/q}$$
$$\leq (\sum_{i=m}^{n} |w_{i+1}^T x_{i+1} - y_{i+1}|^q)^{1/q} + (\sum_{i=m}^{n} (w_i - w_{i+1})^2)^{1/2} (\sum_{i=m}^{n} \|x_{i+1}\|^{2q/(2-q)})^{(2-q)/2q}.$$

The first inequality follows from the Minkowski inequality. The second inequality follows from the Hölder's inequality.

By taking expectation E and again applying the Minkowski inequality, we obtain

$$[\sum_{i=m}^{n} E\, Q(w_i)]^{1/q}$$
$$= [E \sum_{i=m}^{n} |w_i^T x_{i+1} - y_{i+1}|^q]^{1/q}$$
$$\leq E^{1/q}\,[(\sum_{i=m}^{n} |w_{i+1}^T x_{i+1} - y_{i+1}|^q)^{1/q} + (\sum_{i=m}^{n} (w_i - w_{i+1})^2)^{1/2} (\sum_{i=m}^{n} \|x_{i+1}\|^{2q/(2-q)})^{(2-q)/2q}]^q$$
$$\leq E^{1/q} \sum_{i=m}^{n} |w_{i+1}^T x_{i+1} - y_{i+1}|^q + E^{1/q}\,[(\sum_{i=m}^{n} (w_i - w_{i+1})^2)^{q/2} (\sum_{i=m}^{n} \|x_{i+1}\|^{2q/(2-q)})^{(2-q)/2}].$$

By Theorem 1, we have $(\sum_{i=m}^{n}(w_i - w_{i+1})^2)^{q/2} \leq \|w_{n+1}\|^q$. Also observe that

$$E\,(w_{i+1}^T x_{i+1} - y_{i+1})^q = E\,Q_{i+1}(w_{i+1}, x_1, y_1, \ldots, x_{i+1}, y_{i+1}).$$

We thus obtain the theorem.

Corollary 1. *Using formulation (8), we have*

$$[\frac{1}{n+1}\sum_{i=0}^{n} E\,Q(w_i)]^{1/q}$$
$$\leq [E_{(x,y)\sim D}\,\epsilon^q(x,y)]^{1/q} + E^{1/q}[\frac{\|w_{n+1}\|^q}{(n+1)^{q/2}}(\sum_{i=1}^{n+1}\frac{\|x_i\|^{2q/(2-q)}}{n+1})^{(2-q)/2}].$$

Using formulation (9), we have

$$[\frac{1}{n-m}\sum_{i=m+1}^{n} E\,Q(w_i)]^{1/q}$$
$$\leq [\frac{1}{n-m}\sum_{i=m+2}^{n+1}\frac{s}{i}]^{1/q} + E^{1/q}[\frac{\|w_{n+1}\|^q}{(n-m)^{q/2}}(\sum_{i=m+2}^{n+1}\frac{\|x_i\|^{2q/(2-q)}}{n-m})^{(2-q)/2}].$$

Proof. From formulation (8), we obtain

$$EQ_k(w_k, x_1, y_1, \ldots, x_k, y_k) \leq E\frac{1}{k}\sum_{i=1}^{k}\epsilon(x_i, y_i)^q = E_{(x,y)\sim D}\epsilon(x,y)^q.$$

From formulation (9), we obtain

$$Q_k(w_k, x_1, y_1, \ldots, x_k, y_k) \leq \frac{s}{k}.$$

The bounds follow from Theorem 2.

Corollary 2. *Using formulation (8), we have*

$$[\frac{1}{n+1}\sum_{i=0}^{n} E\,Q(w_i)]^{1/q}$$
$$\leq [E_{(x,y)\sim D}\,\epsilon^q(x,y)]^{1/q} + \frac{\sup\|w_{n+1}\|}{(n+1)^{1/2}}E_{x\sim D}^{(2-q)/2q}\|x\|^{2q/(2-q)}.$$

Using formulation (9), we have

$$[\frac{1}{n-m}\sum_{i=m+1}^{n} E\,Q(w_i)]^{1/q}$$
$$\leq [\frac{s}{m+2}]^{1/q} + \frac{\sup\|w_{n+1}\|}{(n-m)^{1/2}}E_{x\sim D}^{(2-q)/2q}\|x\|^{2q/(2-q)}.$$

Proof. We have

$$\begin{aligned}
&E^{1/q}[\|w_{n+1}\|^q(\sum_{i=m+2}^{n+1}\frac{\|x_i\|^{2q/(2-q)}}{n-m})^{(2-q)/2}] \\
\leq &\sup\|w_{n+1}\|E^{1/q}(\sum_{i=m+2}^{n+1}\frac{\|x_i\|^{2q/(2-q)}}{n-m})^{(2-q)/2} \\
\leq &\sup\|w_{n+1}\|E^{(2-q)/2q}(\sum_{i=m+2}^{n+1}\frac{\|x_i\|^{2q/(2-q)}}{n-m}) \\
= &\sup\|w_{n+1}\|E_{x\sim D}^{(2-q)/2q}\|x\|^{2q/(2-q)}.
\end{aligned}$$

The second inequality above follows from the Jensen's inequality. Note also $\frac{1}{n-m}\sum_{i=m+2}^{n+1}\frac{s}{i}\leq\frac{s}{m+2}$. The bounds follow from Corollary 1.

Note that in Corollary 2, $\sup\|w_{n+1}\|$ is with respect to all instances of training data. It is useful when there exists a "target" vector such that the imposed constraints are satisfied. In this case, $\sup\|w_{n+1}\|$ is well-bounded. In the following, we briefly discuss some consequences of our bounds.

Consider the bound for formulation (8) in Corollary 2. If there exists a "target" vector $w_*\in H$ such that $|w_*^Tx-y|\leq\epsilon(x,y)$ for all data, then there exists $k\leq n$ such that

$$(E\,Q(w_k))^{1/q}\leq[E_{(x,y)\sim D}\,\epsilon^q(x,y)]^{1/q}+\frac{\|w_*\|}{(n+1)^{1/2}}E_{x\sim D}^{(2-q)/2q}\|x\|^{2q/(2-q)}. \quad (10)$$

We can further define an estimator $\bar{w}_n=\frac{1}{n+1}\sum_{i=0}^n w_i$, then we obtain from the Jensen's inequality that

$$Q(\bar{w}_n)\leq\frac{1}{n+1}\sum_{i=0}^n Q(w_i).$$

Therefore from Corollary 1, we have

$$(E\,Q(\bar{w}_n))^{1/q}\leq[E_{(x,y)\sim D}\,\epsilon^q(x,y)]^{1/q}+\frac{\|w_*\|}{(n+1)^{1/2}}E_{x\sim D}^{(2-q)/2q}\|x\|^{2q/(2-q)}.$$

Although the estimator $\bar{w}_n$ gives a worst case expected bound that is as good as we can obtain for w_n using a leave-one-out cross validation analysis,[3] it is likely to be inferior to the estimator w_n. This is because the performance of $\bar{w}_n$ is comparable to the average performance of w_k for $k=0,\ldots,n$. However, in practice, we can make the very reasonable assumption that with more training data, we obtain better estimates. Under this assumption, w_n should perform much better than this average bound.

[3] We do not give a thorough comparison here due to the limitation of space and a lack of previous leave-one-out results for similar regression problems that we can cite.

Another observation we can make from (10) is that if we let $q = 2$, then we have to assume that $\|x\|$ is bounded almost everywhere to make the second term of right hand side bounded. However, if we use a formulation with $q < 2$, then we only require the moment $E_{x\sim D}\|x\|^{2q/(2-q)}$ to be bounded. This implies that the formulation with $q < 2$ is more robust to large data than the squared loss formulation is.

We may also consider bounds for formulation (9) in Corollary 2 and obtain similar results. In this case, we shall seek m that approximately minimizes the bound. For example, consider $q = 2$ and assume $\|x\| \leq M$ for all x. We let $m \approx (n+2)s^{1/3}(s^{1/3} + (\sup \|w_{n+1}\| M)^{2/3})^{-1} - 2$, then the bound in Corollary 2 is approximately

$$(n+2)^{-1/2}(s^{1/3} + (\sup \|w_{n+1}\| M)^{2/3})^{3/2}. \tag{11}$$

To interpret this result, we consider the following scenario. Let w be any fixed vector such that $Q(w) \leq A$. By the law of large numbers, we can find $\epsilon_n \to 0$ as $n \to \infty$ such that $P(Q_n(w, x_1, y_1, \ldots, x_n, y_n) > A + \epsilon_n) = o(1/\sqrt{n})$. If we let $s = n(A+\epsilon_n)$, then with large probability, the inequality $Q_n(w, x_1, y_1, \ldots, x_n, y_n) \leq s$ can be satisfied. Technically, under appropriate regularity assumptions, data such that $Q_n(w, x_1, y_1, \ldots, x_n, y_n) > s$ contribute an additional $o(1/\sqrt{n})$ term (which we can ignore) to the expected generalization error. The expected generalization performance is dominated by those training data for which the condition $Q_n(w, x_1, y_1, \ldots, x_n, y_n) \leq s$ is satisfied. This can be obtained from (11), which now becomes

$$(A + \epsilon_n)^{1/2}(1 + O((n(A + \epsilon_n))^{-1/3}(\|w\| M)^{2/3})$$

as $n \to \infty$. Assume that A is not small, then this bound, in a style of $A + \epsilon_n + O(n^{-1/3})$, is not optimal as $n \to \infty$. A better bound in the style of $A + O(n^{1/2})$ can be obtained by directly dealing with formulation (6), which we do not consider in this paper. However, this bound is good when A is small. It also clearly implies that if $\inf_w Q(w) = 0$, then we can choose $A \to 0$ (or equivalently $s = o(n)$) such that there is a sequence of increasing sample size k: $\lim_{k\to\infty} E\, Q(w_k) = 0$. Note that $\inf_w Q(w) = 0$ does not imply that there exists w such that $Q(w) = 0$. Many Hilbert functional spaces H (such as those generated by various kernels) are dense in the set of continuous functions. On the other hand, an arbitrary continuous target function does not correspond to a vector $w \in H$.

It is useful to mention that in many cases (for example, in many kernel methods such as the exponential kernel $K(x, y) = \exp(-(x - y)^2)$), C_n in (8) is not empty as long as either the training data x_i are non-duplicate or y is a function of x. The noise tolerance $\epsilon(x, y)$ in (10) (or s in (11)) can be used to trade-off the two terms on the right-hand-side of the bound in (10) (or the two terms in (11)): when we increase ϵ (or s), we increase the first term but decrease $\|w\|$ in the second term.

4.2 Noise-free formulation

Assume y is generated from the following exact formulation $y = w_*^T x$. Where $w_* \in H$ is the target weight vector. We consider the following noise free estimation formulation:

$$C_n = \{w \in H : w^T x_i = y_i \quad (i = 1, \ldots, n)\}.$$

For simplicity, we assume that $\|x\| \leq M$, and let $q = 2$. Our sequential approximation bound implies

$$\sum_{i=0}^{n} (w_i^T x_{i+1} - y_{i+1})^2 \leq \|w_*\|^2 M^2. \tag{12}$$

The above bound is similar to the regret bound using a stochastic gradient descent rule from $i = 0, \ldots, n$ (we let $w_0' = 0$):

$$w_{i+1}' = w_i' - \frac{1}{x_{i+1}^2}(w_i'^T x_{i+1} - y_{i+1}) x_{i+1}. \tag{13}$$

Using techniques in [1], we may consider the following equality:

$$(w_{i+1}' - w_*)^2 = (w_i' - w_*)^2 - \frac{1}{x_{i+1}^2}(w_i'^T x_{i+1} - y_{i+1})^2.$$

Summing over i, we obtain

$$\sum_{i=0}^{n} (w_i'^T x_{i+1} - y_{i+1})^2 \leq \|w_*\|^2 M^2. \tag{14}$$

Both (12) and (14) can be achieved with a set of orthonormal vectors x_i and $y_i = 1$. So both bounds are tight in the worst case. However, in practice w_n is likely to be a better estimator than w_n'. We illustrate this in the following.

Observe that $w_n = P_{x_1,\ldots,x_n}(w_*)$, where P is the orthogonal projection operator onto the subspace of H spanned by $x_1, \ldots, x_n$. We also denote $w - P_{x_1,\ldots,x_n}(w)$ by $P^{\perp}_{x_1,\ldots,x_n}(w)$. We have the following inequality:

$$(w_i^T x_{i+1} - y_{i+1})^2 = (P^{\perp}_{x_1,\ldots,x_i}(w_*)^T P^{\perp}_{x_1,\ldots,x_i}(x_{i+1}))^2 \leq P^{\perp}_{x_1,\ldots,x_i}(w_*)^2 P^{\perp}_{x_1,\ldots,x_i}(x_{i+1})^2.$$

The effective length of x_{i+1} is thus $P^{\perp}_{x_1,\ldots,x_i}(x_{i+1})$ which can be substantially smaller than M. As an extreme, we can bound $EQ(w_n)$ using the following inequality:

$$E\, Q(w_n) \leq w_*^2 E\, P^{\perp}_{x_1,\ldots,x_n}(x_{n+1})^2.$$

Now, we assume x_i is in an d-dimensional space, and the training data $x_1, \ldots, x_n$ ($n \geq d$) have rank d with probability 1, then $E\, Q(w_n) = 0$. Clearly in general we still have $E\, Q(w_n') > 0$. In practice, even though the data may not lie in finite dimension, the effective dimension measured by $E\, P^{\perp}_{x_1,\ldots,x_n}(x_{n+1})^2$ can decrease

rapidly as $n \to \infty$. In this case, the estimator w_n will be superior to w'_n although both have the same worst-case regret bounds.

This also justifies why in practice, one often run an online method repeatedly over the data. In the noise-free squared loss regression case, if $w'_{n,m}$ is obtained by applying (13) repeatedly m times over the training data, then $w'_{n,m} \to w_n$ as $m \to \infty$. This is because $\lim_{m\to 0} w'^T_{n,m} x_i = y_i = w_n^T x_i$, which easily follows from the regret bound of $w'_{n,m}$. This means $w'_{n,m}$ converges weakly to w_n, which also implies strong convergence since the vectors are in the finite dimensional subspace of H spanned by x_i $(i = 1, \ldots, n)$.

5 Classification

5.1 Sequential-validation bounds

We consider the classification problem: to find w that minimizes the classification error

$$Q(w) = E_{(x,y)\sim D} I(w^T xy \le 0),$$

where D is an unknown distribution. $I(z \le 0)$ is the indicator function: $I(z \le 0) = 1$ if $z \le 0$ and $I(z \le 0) = 0$ otherwise.

Given the training data, we define the empirical expected loss with margin $\gamma > 0$ as:

$$Q_n^\gamma(w, x_1, y_1, \ldots, x_n, y_n) = \frac{1}{n} \sum_{i=1}^{n} I(w^T xy < \gamma).$$

We consider two formulations of C_n in (1). The first formulation is the maximum margin algorithm in [11] (also know as the separable SVM) that employs the following hard-margin constraints:

$$C_n = \{w \in H : w^T x_i y_i \ge 1, \quad (i = 1, \ldots, n)\}. \tag{15}$$

The second formulation employs a soft-margin constraint:

$$C_n = \{w \in H : \sum_{i=1}^{n} L(w^T x_i y_i) \le s\}, \tag{16}$$

where s is a data independent parameter and L is a non-negative and non-increasing convex function. If a solution does not exist in one of the above formulations (that is, C_n is empty), then we let $w_n = 0$. In the following discussion, we only consider the case that C_n is non-empty for all training data.

Theorem 3. *Assume for all training data C_n is non-empty, and $C_k \subseteq C_{k-1}$ for all $k = m, \ldots, n$. For each k, w_k is computed from samples $(x_1, y_1), \ldots, (x_k, y_k)$ using (1). We have the following average expected generalization bound ($1 \le q \le$*

2*):*

$$\sum_{i=m}^{n} E\, Q(w_i)$$
$$\leq \sum_{i=m+1}^{n+1} E\, Q_i^{\gamma}(w_i, x_1, y_1, \ldots, x_i, y_i) + E\,[\frac{\|w_{n+1}\|^q}{\gamma^q}(\sum_{i=m}^{n} |x_{i+1}|^{2q/(2-q)})^{(2-q)/2}].$$

The expectation E is with respect to $n+1$ independent random training samples $(x_1, y_1), \ldots, (x_{n+1}, y_{n+1})$ from D. Note that $\gamma > 0$ can be a function of $(x_1, y_1), \ldots, (x_{n+1}, y_{n+1})$.

Proof. Note that

$$I(w_i x_{i+1} y_{i+1} \leq 0) \leq I(w_{i+1}^T x_{i+1} y_{i+1} < \gamma) + \frac{1}{\gamma^q}|(w_{i+1} - w_i)^T x_{i+1}|^q.$$

Summing over $i = m, \ldots, n$, and using the Hölder's inequality, we have

$$\sum_{i=m}^{n} I(w_i x_{i+1} y_{i+1} \leq 0)$$
$$\leq \sum_{i=m}^{n} I(w_{i+1}^T x_{i+1} y_{i+1} < \gamma) + \frac{1}{\gamma^q}(\sum_{i=m}^{n}(w_i - w_{i+1})^2)^{q/2}(\sum_{i=m}^{n} \|x_{i+1}\|^{2q/(2-q)})^{(2-q)/2}.$$

Taking expectation E, we obtain

$$\sum_{i=m}^{n} E\, Q(w_i)$$
$$\leq \sum_{i=m}^{n} E\, I(w_{i+1}^T x_{i+1} y_{i+1} < \gamma) + E\,\frac{1}{\gamma^q}(\sum_{i=m}^{n}(w_i - w_{i+1})^2)^{q/2}(\sum_{i=m}^{n} \|x_{i+1}\|^{2q/(2-q)})^{(2-q)/2}.$$

From Theorem 1, we have $(\sum_{i=m}^{n}(w_i - w_{i+1})^2)^{q/2} \leq \|w_{n+1}\|^q$. Also observe that

$$E\, I(w_{i+1}^T x_{i+1} y_{i+1} < \gamma) = E\, Q_{i+1}^{\gamma}(w_{i+1}, x_1, y_1, \ldots, x_{i+1}, y_{i+1}).$$

We thus obtain the theorem.

Corollary 3. *Using formulation (15), we have*

$$\sum_{i=0}^{n} E\, Q(w_i) \leq E\,[\|w_{n+1}\|^q(\sum_{i=1}^{n+1} |x_i|^{2q/(2-q)})^{(2-q)/2}].$$

Using formulation (16), we have

$$\sum_{i=m+1}^{n} E\, Q(w_i) \leq \sum_{i=m+2}^{n+1} \frac{s}{i}\, E\,\frac{1}{L(\gamma)} + E\,[\frac{\|w_{n+1}\|^q}{\gamma^q}(\sum_{i=m+1}^{n} |x_{i+1}|^{2q/(2-q)})^{(2-q)/2}].$$

Proof. If $\gamma = 1$, we obtain from formulation (15)

$$Q_k^\gamma(w_k, x_1, y_1, \ldots, x_k, y_k) = 0.$$

From formulation (16), we obtain

$$Q_k^\gamma(w_k, x_1, y_1, \ldots, x_k, y_k) \leq \frac{1}{k}\sum_{i=1}^{k} \frac{L(w_k^T x_i y_i)}{L(\gamma)} \leq \frac{s}{kL(\gamma)}.$$

The bounds follow from Theorem 3.

Corollary 4. *Using formulation (15), we have*

$$\frac{1}{n+1}\sum_{i=0}^{n} E\,Q(w_i) \leq \frac{\sup \|w_{n+1}\|^q}{(n+1)^{q/2}} E_{x\sim D}^{(2-q)/2} \|x\|^{2q/(2-q)}.$$

Using formulation (16), with fixed γ, we have

$$\frac{1}{n-m}\sum_{i=m+1}^{n} E\,Q(w_i) \leq \frac{s}{(m+2)L(\gamma)} + \frac{\sup \|w_{n+1}\|^q}{(n-m)^{q/2}\gamma^q} E_{x\sim D}^{(2-q)/2} \|x\|^{2q/(2-q)}.$$

Proof. Using Corollary 3, the proof is the essentially the same as that of Corollary 2.

The separable case (15) has also been considered in [10], where they derived a result that is similar to the corresponding bound in Corollary 4 with $q = 2$.

We shall mention that in most of previous theoretical studies, the quantity $\|x\|$ were assumed to be bounded. However, our bounds can still be applied when $\|x\|$ is not bounded, as long as there exists $1 \leq q < 2$ such that the moment $E_{x\sim D}\|x\|^{2q/(2-q)}$ is finite. On the other hand, this causes a slow down of convergence in our bounds.

Consider formulation (16) in Corollary 4. We may seek m to approximately minimize the bound. For example, consider $q = 2$ and assume $\|x\| \leq M$ for all x. We let $m \approx (n+2)s^{1/2}L(\gamma)^{1/2}(s^{1/2} + L(\gamma)^{1/2}\sup\|w_{n+1}\|M\gamma^{-1})^{-1} - 2$, then the bound in Corollary 2 is approximately

$$(n+2)^{-1}(s^{1/2}L(\gamma)^{-1/2} + \sup\|w_{n+1}\|M\gamma^{-1})^2. \tag{17}$$

Similar to the discussion after (11), the above bound can be improved by directly analyzing the penalty type formulation (6) if the problem is not linearly separable. On the other hand, the bound is good if the problem is nearly separable. If $\inf_w Q(w) = 0$, then we can find $s = o(n)$ such that there is a sequence of increasing sample size k: $\lim_{k\to\infty} E\,Q(w_k) = 0$.

5.2 Separable linear classification

We consider the separable case using formulation (15). Similar to the regression case, we may compare our results with the perceptron mistake bound. However, in this section we consider the comparison with Vapnik's leave-one-out bound in [11] and illustrate why our sequential validation analysis can provide useful information that cannot be obtained by using the leave-one-out bound alone.

To be compatible with Vapnik's result, we consider the special case of $q = 2$ in Corollary 3:

$$\sum_{i=0}^{n} E\, Q(w_i) \leq E\, w_{n+1}^2 \max_{i=1,\ldots,n+1} x_i^2.$$

This implies the following: there exists a sample size $k \leq n$ such that

$$E\, Q(w_k) \leq \frac{1}{n+1} E\, w_{n+1}^2 \max_{i=1,\ldots,n+1} x_i^2.$$

This can be compared with Vapnik's leave-one-out bound in [11], which can be expressed as:

$$E\, Q(w_n) \leq \frac{1}{n+1} E\, w_{n+1}^2 \max_{i=1,\ldots,n+1} x_i^2.$$

Using our bound, we may also consider an estimator w_n' that is randomly selected among w_i from $i = 0, \ldots, n$. Clearly,

$$E\, Q(w_n') \leq \frac{1}{n+1} E\, w_{n+1}^2 \max_{i=1,\ldots,n+1} x_i^2.$$

This gives a comparable bound as the leave-one-out bound of w_n. However, as we have argued before, despite of the same worst case bound, w_n' is likely to be inferior to w_n. If we make the reasonable assumption that with more training data, one can obtain better results, than our sequential validation result implies that the performance of w_n should be better than what is implied by the leave-one-out bound.

Specifically, we consider the situation that there exists an estimator M such that $\|x\| \leq M$ and there exists a weight parameter w_* so that the condition $w_*^T xy \geq 1$ is always satisfied. Vapnik's bounds implies that

$$E\, Q(w_n) \leq \frac{1}{n} w_*^2 M^2.$$

That is, the expected generalization error decreases at an order of $O(1/n)$. However, our bound implies that

$$\sum_{n=0}^{\infty} E\, Q(w_n) \leq w_*^2 M^2.$$

This implies that asymptotically the expected generalization error decreases faster than $O(1/n)$ since $\sum_{n=1}^{\infty} 1/n = \infty$.

This example shows that the sequential approximation analysis proposed in this paper provides useful insights into classification problems that cannot be obtained by using the leave-one-out analysis.

6 Summary

In this paper, we derived a general sequential approximation bound for a class of sample dependent convex optimization problems. Based on this bound, we are able to obtain sequential cross validation bounds for some learning formulations. A unique aspect that distinguishes this work from many previous works on mistake bound analysis is that we directly bound the convergence of consecutive parameter estimates in a batch learning setting.

The specific analysis given in this paper for constrained regularization formulation is more suitable for problems that contain small noise. However, it is easy to generalize the idea to analyze penalty type regularization formulations including the standard forms of Gaussian processes and soft-margin support vector machines. A direct analysis of penalty type regularization formulations is more suitable for large noise problems. Note that we have already demonstrated in the paper that the constrained regularization formulation considered here is numerically equivalent to the penalty type regularization formulation.

References

1. N. Cesa-Bianchi, P. Long, and M. K. Warmuth. Worst-case quadratic loss bounds for prediction using linear functions and gradient descent. *IEEE Transactions on Neural Networks*, 7:604–619, 1996.
2. Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and other Kernel-based Learning Methods.* Cambridge University Press, 2000.
3. Jüergen Forster and Manfred Warmuth. Relative expected instantaneous loss bounds. In *COLT 00*, pages 90–99, 2000.
4. Jüergen Forster and Manfred Warmuth. Relative loss bounds for temporal-difference learning. In *ICML 00*, pages 295–302, 2000.
5. Geoffrey J. Gordon. Regret bounds for prediction problems. In *COLT 99*, pages 29–40, 1999.
6. Harro G. Heuser. *Functional analysis.* John Wiley & Sons Ltd., Chichester, 1982. Translated from the German by John Horváth, A Wiley-Interscience Publication.
7. T. Jaakkola and D. Haussler. Probabilistic kernel regression models. In *Proceedings of the 1999 Conference on AI and Statistics*, 1999.
8. Michael Kearns and Dana Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6):1427–1453, 1999.
9. J. Kivinen and M.K. Warmuth. Additive versus exponentiated gradient updates for linear prediction. *Journal of Information and Computation*, 132:1–64, 1997.
10. Yi Li and Philip M. Long. The relaxed online maximum margin algorithm. In S.A. Solla, T.K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 498–504. MIT Press, 2000.
11. V.N. Vapnik. *Statistical learning theory.* John Wiley & Sons, New York, 1998.
12. Tong Zhang. Convergence of large margin separable linear classification. In *Advances in Neural Information Processing Systems 13*, pages 357–363, 2001.
13. Tong Zhang. A leave-one-out cross validation bound for kernel methods with applications in learning. In *COLT*, 2001.