Analysis of Regularized Linear Functions for Classification Problems

Tong Zhang

IBM Research Report RC-21572

Analysis of Regularized Linear Functions for Classification Problems

Tong Zhang
Mathematical Sciences Department
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
tzhang@watson.ibm.com

Abstract

Recently, sample complexity bounds have been derived for problems involving linear functions such as neural networks and support vector machines. In this paper, we extend some theoretical results in this area by providing convergence analysis for regularized linear functions with an emphasis on classification problems. The class of methods we study in this paper generalize support vector machines and are conceptually very simple. To analyze these methods within the traditional PAC learning framework, we derive dimensional independent covering number bounds for linear systems under certain regularization conditions, and obtain relevant generalization bounds. We also present an analysis for these methods from the asymptotic statistical point of view. We show that this technique provides better description for large sample behaviors of these algorithms. Furthermore, we shall investigate numerical aspects of the proposed methods, and establish their relationship with ill-posed problems studied in numerical mathematics.

1 Introduction

In this paper, we are interested in the generalization performance of linear classifiers obtained from certain algorithms. From computational learning theory point of view, such performance measurements, or sample complexity bounds, can be described by a quantity called covering number [26, 37, 35], which measures the size of a parametric function family. For two-class classification problem, the covering number can be bounded by a combinatorial quantity called VC-dimension [37, 28]. Following this work, researchers have found other combinatorial quantities (dimensions) useful for bounding the covering numbers. Consequently, the concept of VC-dimension has been generalized to deal with more general problems, for example in [26, 35]. Many of the more recent results have been summarized in [5] and [34].

Recently, Vapnik introduced the concept of support vector machine [36] which has been successful applied to many real problems. This method achieves good generalization by restricting the 2-norm of the weights of a separating hyperplane. A similar technique has been investigated by

Bartlett [3], where the author studied the performance of neural networks when the 1-norm of the weights is bounded. The same idea has also been applied in [29] to explain the effectiveness of the boosting algorithm.

In Section 4, we will see that support vector machines for training linear classifiers are special cases of the following general form of regularization method which has been widely studied for regression problems both in statistics and in numerical mathematics:

$$\inf_w E_{x,y} L(w, x, y) = \inf_w E_{x,y} f(w^T xy) + \lambda g(w), \tag{1}$$

where $E_{x,y}$ is the expectation over a distribution of (x, y), and $y \in \{-1, 1\}$ is the binary label of data vector x. To apply this formulation for the purpose of training linear classifiers, we can choose f as a decreasing function, such that $f(\cdot) \geq 0$, and choose $g(w) \geq 0$ as a function that penalizes large w ($\lim_{w\to\infty} g(w) \to \infty$). λ is an appropriately chosen positive parameter to balance the two terms.

In this paper, we study some theoretical aspects of using (1) to obtain a linear classifier w with the following decision rule to predict the label y of an unlabeled data x:

$$y = \begin{cases} 1 & \text{if} \quad w^T x > 0, \\ -1 & \text{if} \quad w^T x \leq 0. \end{cases}$$

To analyze this system, we shall first extend the covering number bounds of 1-norm regularized neural networks in [3], and emphasize the importance of dimensional independence. Generalization performance of (1) in the PAC learning framework is derived accordingly. This analysis shows that the new formulation (1) has similar theoretical advantages as support vector machines while conceptually is simpler (in the sense that the new formulation is in the form of traditional regularization methods in statistics and numerical mathematics people are familiar with). One disadvantage of PAC style analysis is that the derived small sample bounds are often not accurate for real problems. If f and g are smooth functions in (1), we show that exact asymptotic formulae exist. These asymptotic results provide better large sample descriptions for learning algorithms. Furthermore, since the form of (1) comes from regularization methods used in numerical mathematics for solving ill-posed problems, it is natural to study related numerical issues. We will demonstrate the relationship of the new method with support vector machines, and why a non-zero λ is important for stability of the system under observation noise. Note that although such issues are very important in practice, they are often ignored under the standard PAC learning framework.

In a separate report [38], we study efficient numerical algorithms for solving (1) and its generalizations. We demonstrate that the newly derived methods compare favorably with the quadratic programming formulation of support vector machines. The new methods based on (1) and its generalizations are more flexible and often more efficient to solve numerically. Conceptually, we may also regard (1) as a model of data distribution, and derive interesting new learning algorithms accordingly.

The paper is organized as follows. Section 2 studies (1) from the PAC learning point of view. In

Section 2.1, we briefly review the concept of covering numbers as well as the main results related to analyzing the performance of learning algorithms. In Section 2.2, we introduce the regularization idea. Our main goal is to construct regularization conditions so that dimension independent bounds on covering numbers can be obtained. Section 2.3 extends results from the previous section to nonlinear compositions of linear functions. We will present some generalization error bounds for (1). In Section 3, we derive an asymptotic formula for the expected generalization performance of a learning algorithm, which will then be used to analyze the proposed formulation. In Section 4, we study some numerical properties of the new formulation and compare it to the problem of solving ill-posed systems. Section 5 summarizes results obtained in this paper.

2 PAC style generalization bounds

2.1 Covering numbers

We formulate the learning problem as to find a parameter from random observations to minimize the expected loss (*risk*): given a loss function $L(\alpha, x)$ and n observations $X_1^n = \{x_1, \ldots, x_n\}$ independently drawn from a fixed but unknown distribution D, we want to find α that minimizes the expected loss over x:

$$R(\alpha) = E_x\, L(\alpha, x) = \int L(\alpha, x)\, dP(x). \tag{2}$$

Without any assumption of the underlying distribution x, the most natural method for solving (2) using a limited number of observations is by the *empirical risk minimization* (ERM) method (*cf.* [35, 36]). We simply choose a parameter α that minimizes the observed risk:

$$R(\alpha, X_1^n) = E_{X_1^n} L(\alpha, x) = \frac{1}{n} \sum_{i=1}^{n} L(\alpha, x_i), \tag{3}$$

where we use $E_{X_1^n}$ to denote the empirical expectation over the observed data.

We denote the parameter obtained in this way as $\alpha_{\mathrm{erm}}(X_1^n)$. The convergence behavior of this method can be analyzed by using the VC theory under the PAC framework, which relies on the uniform convergence of the empirical risk (the uniform law of large numbers): $\sup_\alpha |R(\alpha, X_1^n) - R(\alpha)|$. Such a bound can be obtained from quantities that measure the size of a Glivenko-Cantelli class. For a finite number of indices, the family size can be measured simply by its cardinality. For general function families, a well known quantity to measure the degree of uniform convergence is the *covering number* which can be dated back to Kolmogrov [21, 22]. The idea is to discretize (the discretization process can depend on the data X_1^n) the parameter space into N values $\alpha_1, \ldots, \alpha_N$ so that each $L(\alpha, \cdot)$ can be approximated by $L(\alpha_i, \cdot)$ for some i. We shall only describe a simplified version relevant for our purposes.

Definition 2.1 *Let B be a metric space with metric ρ. Given a norm p, observations $X_1^n = [x_1, \ldots, x_n]$, and vectors $f(\alpha, X_1^n) = [f(\alpha, x_1), \ldots, f(\alpha, x_n)] \in B^n$ parameterized by α, the covering*

number in p-norm, denoted as $\mathcal{N}_p(f, \epsilon, X_1^n)$*, is the minimum number of a collection of vectors* $v_1, \ldots, v_m \in B^n$ *such that* $\forall \alpha$, $\exists v_i$: $\|\rho(f(\alpha, X_1^n), v_i)\|_p \leq n^{1/p}\epsilon$. *We also denote* $\mathcal{N}_p(f, \epsilon, n) = \max_{X_1^n} \mathcal{N}_p(f, \epsilon, X_1^n)$.

Note that from the definition and the Jensen's inequality, we have $\mathcal{N}_p \leq \mathcal{N}_q$ for $p \leq q$. We will always assume the metric on R to be $|x_1 - x_2|$ if not explicitly specified otherwise. The following theorem is due to Pollard [26]:

Theorem 2.1 ([26]) $\forall n$, $\epsilon > 0$ *and distribution* D,

$$P[\sup_\alpha |R(\alpha, X_1^n) - R(\alpha)| > \epsilon] \leq 8E[\mathcal{N}_1(L, \epsilon/8, X_1^n)] \exp(\frac{-n\epsilon^2}{128M^2}),$$

where $M = \sup_{\alpha,x} L(\alpha, x) - \inf_{\alpha,x} L(\alpha, x)$, *and* $X_1^n = \{x_1, \ldots, x_n\}$ *are independently drawn from* D.

The constants in the above theorem can be improved for certain problems: see [6, 13, 35, 36] for related results. However, they yield very similar bounds. The result most relevant for this paper is a lemma in [3] where the 1-norm covering number is replaced by the ∞-norm covering number. The latter can be bounded by a scale-sensitive combinatorial dimension [1], which can be bounded from the 1-norm covering number if this covering number does not depend on n. These results can replace Theorem 2.1 to yield better estimates under certain circumstances.

Since Bartlett's lemma in [3] is only for classification problems, we shall extends it to general problems so that it is comparable to Theorem 2.1. In the following theorem, we replace the "margin" concept for classification problems by a notion of separation for general problems. We also avoid introducing the concept of "fat-shattering" dimension which leads to some complicated technical manipulations in [3]. The important differences between the following theorem and Theorem 2.1 are: firstly, with the existence of a γ-separating function, we are able to use different accuracies γ and ϵ respectively in the covering number estimate and the Chernoff bound; and secondly, the covering number itself is not that of the overall loss function.

Theorem 2.2 *Let* f_1 *and* f_2 *be two functions:* $R^n \to [0, 1]$ *such that* $|y_1 - y_2| \leq \gamma$ *implies* $f_1(y_1) \leq f_3(y_2) \leq f_2(y_1)$ *where* $f_3 : R^n \to [0, 1]$ *is a reference separating function, then*

$$P[\sup_\alpha [E_x f_1(L(\alpha, x)) - E_{X_1^n} f_2(L(\alpha, x))] > \epsilon] \leq 4E[\mathcal{N}_\infty(L, \gamma, X_1^n)] \exp(\frac{-n\epsilon^2}{32}),$$

Proof. We follow the standard techniques (*cf.* [26, 37]).

Step 1 (symmetrization by a replicate sample). For all $n\epsilon^2 \geq 2$, and consider i.i.d. random sample Y_1^n, independent of X_1^n,

$$\begin{aligned} &P[\sup_\alpha [E_x f_1(L(\alpha, x)) - E_{Y_1^n} f_2(L(\alpha, y))] > \epsilon] \\ \leq &2P[\sup_\alpha E_{X_1^n} f_1(L(\alpha, x)) - E_{Y_1^n} f_2(L(\alpha, y)) > \epsilon/2]. \end{aligned}$$

To see this, consider a function α^* such that $\alpha^*(Y_1^n)$ is a parameter that satisfies $E_x f_1(L(\alpha^*, x)) - E_{Y_1^n} f_2(L(\alpha^*, y)) > \epsilon$ if such a parameter exists; and let $\alpha^*(Y_1^n)$ be an arbitrary parameter if no such parameter exists. Note that for any Y_1^n, by the Chebyshev's inequality, the conditional probability

$$\begin{aligned}&P[E_x f_1(L(\alpha^*, x)) - E_{X_1^n} f_1(L(\alpha^*, y)) \leq \epsilon/2 | Y_1^n] \\ \geq & 1 - \frac{1}{n\epsilon^2/4} E_x f_1(L(\alpha^*, x))(1 - E_x f_1(L(\alpha^*, x))) \geq 1/2.\end{aligned}$$

We thus have

$$\begin{aligned}&\frac{1}{2} P[\sup_\alpha [E_x f_1(L(\alpha, x)) - E_{Y_1^n} f_2(L(\alpha, y))] > \epsilon] \\ =&\frac{1}{2} P[E_x f_1(L(\alpha^*, x)) - E_{Y_1^n} f_2(L(\alpha^*, y)) > \epsilon] \\ \leq & P[E_x f_1(L(\alpha^*, x)) - E_{Y_1^n} f_2(L(\alpha^*, y)) > \epsilon, E_x f_1(L(\alpha^*, x)) - E_{X_1^n} f_1(L(\alpha^*, y)) \leq \epsilon/2] \\ \leq & P[E_{X_1^n} f_1(L(\alpha^*, x)) - E_{Y_1^n} f_2(L(\alpha^*, y)) > \epsilon/2] \\ \leq & P[\sup_\alpha E_{X_1^n} f_1(L(\alpha, x)) - E_{Y_1^n} f_2(L(\alpha, y)) > \epsilon/2].\end{aligned}$$

Step 2 (symmetrization by random signs). Consider i.i.d. sign variables $\sigma_1, \ldots, \sigma_n$, independent of X_1^n and Y_1^n, with $P(\sigma_i = -1) = P(\sigma_i = 1) = 1/2$. Define

$$g_\sigma(\alpha, x) = (f_1(L(\alpha, x)) - f_2(L(\alpha, x)))/2 + \sigma(f_1(L(\alpha, x)) + f_2(L(\alpha, x)))/2,$$

and

$$h_\sigma(\alpha, y) = -(f_1(L(\alpha, y)) - f_2(L(\alpha, y)))/2 + \sigma(f_1(L(\alpha, y)) + f_2(L(\alpha, y)))/2.$$

It follows that the distribution of

$$\sup_\alpha \sum_{i=1}^n f_1(L(\alpha, x_i)) - f_2(L(\alpha, y_i))$$

is the same as that of

$$\sup_\alpha \sum_{i=1}^n g_{\sigma_i}(\alpha, x_i) - h_{\sigma_i}(\alpha, y_i).$$

Therefore

$$\begin{aligned}
&P[\sup_\alpha E_{X_1^n} f_1(L(\alpha,x)) - E_{Y_1^n} f_2(L(\alpha,y)) > \epsilon/2] \\
=&P[\sup_\alpha \frac{1}{n}\sum_{i=1}^n g_{\sigma_i}(\alpha,x_i) - h_{\sigma_i}(\alpha,y_i) > \epsilon/2] \\
\le&2P[\sup_\alpha \frac{1}{n}\sum_{i=1}^n g_{\sigma_i}(\alpha,x_i) > \epsilon/4].
\end{aligned}$$

Step 3 (derandomizing data). To estimate $P[\sup_\alpha \frac{1}{n}\sum_{i=1}^n g_{\sigma_i}(\alpha,x_i) > \epsilon/4]$, we fix X_1^n and estimate the conditional probability

$$P[\sup_\alpha \frac{1}{n}\sum_{i=1}^n g_{\sigma_i}(\alpha,x_i) > \epsilon/4 | X_1^n].$$

Let $\{(z_{1,j},\ldots z_{n,j}) : j = 1,\ldots,m\}$ be an infinity-norm γ-covering of $L(\alpha, X_1^n)$, where $m = \mathcal{N}_\infty(L,\gamma,X_1^n)$, then by definition, $\forall\alpha$, $\exists j$ such that $|z_{i,j} - L(\alpha,x_i)| < \gamma$ for all i. Note that $g_1(\alpha,x_i) = f_1(L(\alpha,x_i)) \le f_3(z_{i,j})$ and $g_{-1}(\alpha,x_i) = -f_2(L(\alpha,x_i)) \le -f_3(z_{i,j})$, therefore $g_{\sigma_i}(\alpha,x_i) \le \sigma_i f_3(z_{i,j})$. We thus obtain

$$\begin{aligned}
&P[\sup_\alpha \frac{1}{n}\sum_{i=1}^n g_{\sigma_i}(\alpha,x_i) > \epsilon/4 | X_1^n] \\
\le&P[\sup_j \frac{1}{n}\sum_{i=1}^n \sigma_i f_3(z_{i,j}) > \epsilon/4 | X_1^n] \\
\le&\mathcal{N}_\infty(L,\gamma,X_1^n)\sup_j P[\frac{1}{n}\sum_{i=1}^n \sigma_i f_3(z_{i,j}) > \epsilon/4 | X_1^n] \\
\le&\mathcal{N}_\infty(L,\gamma,X_1^n)e^{-n\epsilon^2/32}.
\end{aligned}$$

The last inequality follows from the Hoeffding's inequality [16]. □

We say that f_1 and f_2 has a γ separator if there exists f_3 such that $|y_1 - y_2| \le \gamma$ implies $f_1(y_1) \le f_3(y_2) \le f_2(y_1)$. Note that if we define $f^\gamma(y) = \sup_{|z-y|<\gamma} f_1(z)$, then f_1 and $f^{2\gamma}$ has a γ separator f^γ.

The above theorem gives the following PAC style generalization error bound: $\forall\gamma,\eta > 0$, with probability of at least $1-\eta$,

$$E_x f_1(L(\alpha,x)) \le E_{X_1^n} f_2(L(\alpha,x)) + \sqrt{\frac{32}{n}(\ln 4\mathcal{N}_\infty(L,\gamma,n) + \ln\frac{1}{\eta})}.$$

If we consider a sequence of functions f_2^γ parameterized by γ, each having a γ separator, then we immediately notice that in the above bound, γ has to be data independent. However, by using an idea described in [30], it is not difficult to give a uniform bound so that γ can be chosen according

to data:

Corollary 2.1 *Let $0 \leq f_1 \leq f_2^\gamma \leq 1$ be two families of functions parameterized by $\gamma \in [0,1]$, such that f_1 and f_2^γ has a γ separator. Assume that $f_2^{\gamma_1}(y) \geq f_2^{\gamma_2}(y)$ if $\gamma_1 \geq \gamma_2$. Let $\gamma_1 > \gamma_2 > \cdots$ be a decreasing sequence of parameters, and p_i be a sequence of positive numbers such that $\sum_i p_i = 1$, then for all $\eta > 0$, with probability of at least $1-\eta$ over data:*

$$E_x f_1(L(\alpha, x)) \leq E_{X_1^n} f_2^\gamma(L(\alpha, x)) + \sqrt{\frac{32}{n}(\ln 4\mathcal{N}_\infty(L, \gamma_i, X_1^n) + \ln \frac{1}{p_i \eta})}$$

for all $\gamma \in [0,1]$, where i is the smallest index such that $\gamma_i < \gamma$.

Proof. The result follows from Theorem 2.2 and basic probability arguments presented in [30]. $\forall i > 0$ (let $\gamma_0 = 1$), with probability at most $p_i\eta$, we have

$$E_x f_1(L(\alpha, x)) > E_{X_1^n} f_2^{\gamma_{i-1}}(L(\alpha, x)) + \sqrt{\frac{32}{n}(\ln 4\mathcal{N}_\infty(L, \gamma_i, X_1^n) + \ln \frac{1}{p_i \eta})}.$$

Summing up over i, with probability at most η,

$$E_x f_1(L(\alpha, x)) > E_{X_1^n} f_2^{\gamma_{i-1}}(L(\alpha, x)) + \sqrt{\frac{32}{n}(\ln 4\mathcal{N}_\infty(L, \gamma_i, X_1^n) + \ln \frac{1}{p_i \eta})}.$$

for at least one i, which implies the corollary. □

If close to perfect generalization can be achieved, *i.e.* $E_{X_1^n} f_2^\gamma(L(\alpha, x)) \approx 0$, we can obtain better bounds by using a refined version of Chernoff bound where $-2n\epsilon^2$ is replaced by $-n\epsilon^2/2(Ef + \epsilon)$ for over estimation and $-n\epsilon^2/Ef$ for under estimation. In the extreme case that some choice of α achieves perfect generalization: $E_x f_2^\gamma(L(\alpha, x)) = 0$, and assume that our choices of $\alpha(X_1^n)$ always satisfy the condition $E_{X_1^n} f_2^\gamma(L(\alpha, x)) = 0$, then it is not hard to see that a generalization performance of $O(\frac{1}{n}\log \mathcal{N}_\infty)$ instead of $O(\sqrt{\frac{1}{n}\log \mathcal{N}_\infty})$ can be achieved.

2.2 Covering number bounds for linear systems

Theorems in Section 2.1 demonstrate the important roles of covering numbers for analyzing the generalization performance of a learning algorithm. In this section, we shall derive a few new bounds on covering numbers for the following form of real valued loss functions:

$$L(w, x) = x^T w = \sum_{i=1}^{d} x_i w_i. \tag{4}$$

As we shall see later, these bounds are relevant to the convergence properties of (1). For simplicity, we shall skip covering number results for vector valued functions since they are less relevant to the regularization method (1). A brief discussion on related issues will be given in Section 2.3 when we study nonlinear compositions of linear systems. Note that in order to apply Theorem 2.1, since $\mathcal{N}_1 \leq \mathcal{N}_2$, therefore it is sufficient to estimate $\mathcal{N}_2(L, \epsilon, n)$ for $\epsilon > 0$. It is clear that $\mathcal{N}_2(L, \epsilon, n)$ is

not finite if no restrictions on x and w are imposed. Therefore in the following, we will assume that each $\|x_i\|_p$ is bounded, and study conditions of $\|w\|_q$ so that $\log \mathcal{N}(f, \epsilon, n)$ is independent or weakly dependent of d.

We start our analysis with a lemma that is attributed to Maurey, also see [2, 18].

Lemma 2.1 (Maurey) *In a Hilbert space, let $f = \sum_{i=1}^d w_i g_i$, where each $\|g_i\| \leq b$, $w_i \geq 0$ and $\alpha = \sum_i w_i \leq 1$, then for every $n \geq 1$, there exist non-negative integers $k_1, \ldots, k_d \geq 0$, such that $\sum_{i=1}^d k_i \leq n$ and*

$$\|f - \frac{1}{n} \sum_{i=1}^d k_i g_i\|^2 \leq \frac{\alpha b^2 - \|f\|^2}{n}.$$

Our first result generalizes a theorem of Bartlett [3]. The original results is with $p = \infty$ and $q = 1$, and the related technique has also appeared in [23, 29].

Theorem 2.3 *If $\|x_i\|_p \leq b$ and $\|w\|_q \leq a$, where $1/p + 1/q = 1$ and $2 \leq p \leq \infty$, then*

$$\log_2 \mathcal{N}_2(L, \epsilon, n) \leq \lceil \frac{a^2 b^2}{\epsilon^2} \rceil \log_2(2d + 1).$$

Proof. Consider matrix $X = [x_1, \ldots, x_n]^T$. Denote the columns of X as $y_1, \ldots, y_d$. Let

$$g_i = \frac{n^{1/p} ab}{\|y_i\|_p} y_i, \qquad w_i' = \frac{\|y_i\|_p}{n^{1/p} ab} w_i.$$

By Hölder's inequality, it is easy to check that

$$\begin{aligned} \sum_i |w_i'| &= \left| \sum_i \frac{\|y_i\|_p}{n^{1/p} ab} w_i \right| \\ &\leq \frac{1}{n^{1/p} ab} (\sum_i \|y_i\|_p^p)^{1/p} (\sum_i |w_i|^q)^{1/q} \\ &\leq \frac{1}{n^{1/p} ab} (nb^p)^{1/p} a = 1. \end{aligned}$$

Since function $x^{p/2}$ is convex, thus by the Jensen's inequality, we obtain $n^{-1/2} \|y_i\|_2 \leq n^{-1/p} \|y_i\|_p$. This implies that $\|g_i\|_2 \leq n^{1/2} ab$. Therefore by Lemma 2.1, if we let $k \geq (ab/\epsilon)^2$, then $\forall z = \sum_i w_i y_i$, we can find integers $k_1, \ldots, k_d$ such that $\sum_i |k_i| \leq k$ and

$$\|z - \frac{1}{k} \sum_i k_i g_i\|_2^2 \leq \frac{n a^2 b^2}{k} \leq n\epsilon^2.$$

This means that the covering number $\mathcal{N}_2(L, \epsilon, n)$ is no larger than the number of integer solutions of $\sum_i |k_i| \leq k$, which is less than or equal to $(2d + 1)^k$. □

The above bound on the covering number depends logarithmically on d, which is already quite weak (as compared to linear dependence on d in the standard situation). However, the bound in Theorem 2.3 is not tight for $p < \infty$. For example, the following theorem improves the above bounds for $p = 2$. Our technique used in the proof relies on the SVD decomposition [10] for matrices, and improves a similar result in [30] (which relies on the "fat-shattering" dimension) by a logarithmic factor of n.

Theorem 2.4 *If $\|x_i\|_2 \leq b$ and $\|w\|_2 \leq a$, then*

$$\log_2 \mathcal{N}_2(L, \epsilon, n) \leq \lceil \frac{2a^2b^2}{\epsilon^2} \rceil \log_2(4a^2b^2/\epsilon^2 + 1).$$

Proof. Consider the matrix X in the proof of Theorem 2.3. Assume that the SVD decomposition of X is $X = USV$. Let $q = \min(n, d)$ and the non-zero diagonal elements of S are $\lambda_1, \ldots, \lambda_q$. Therefore after an orthogonal rotation which does not change the L_2 norm, the range R of Xw can be written as $\sum_i (z_i/\lambda_i)^2 \leq a^2$. We need to find an $\sqrt{n}\epsilon$-L_2 cover for R. Note that the Frobenius norm $\|X\|_F^2 \leq nb^2$, which implies that $\sum_i \lambda_i^2 \leq nb^2$. Therefore there are at most $p \leq 2a^2b^2/\epsilon^2$ of those λ_i which can be larger than or equal to $\sqrt{\frac{n}{2}}\epsilon/a$. Assume that they are $\lambda_1, \ldots, \lambda_p$, then by the Schwartz inequality, $\sum_{i=1}^p |z_i| \leq (\sum_i (z_i/\lambda_i)^2)^{1/2} (\sum_i (\lambda_i)^2)^{1/2} \leq \sqrt{n}ab$. Thus by Theorem 2.3, there exists an $\epsilon/\sqrt{2}$-covering of $R' = \{[z_1, \ldots, z_p] : \sum_{i=1}^p (z_i/\lambda_i)^2 \leq a^2\}$ with $\exp[\lceil \frac{2a^2b^2}{\epsilon^2} \rceil \log_2(2p+1)]$ vectors. Since $(\sum_{i=p+1}^n z_i^2)^{1/2} \leq \sqrt{\frac{n}{2}}\epsilon$, therefore, these vectors in the $\epsilon/\sqrt{2}$-covering of R' padded with zeros give an ϵ covering of R. $\square$

The next theorem shows that if $1/p + 1/q > 1$, then the 2-norm covering number is also independent of dimension.

Theorem 2.5 *If $\|x_i\|_p \leq b$ and $\|w\|_q \leq a$, where $1 \leq q \leq 2$ and $\delta = 1/p + 1/q - 1 > 0$, then*

$$\log_2 \mathcal{N}_2(L, \epsilon, n) \leq \lceil \frac{4a^2b^2}{\epsilon^2} \rceil \log_2(2(2ab/\epsilon)^{1/\delta} + 1)$$

Proof. Let $p' = q/(q-1)$, then as in the proof of Theorem 2.3, we obtain

$$\sum_i |\frac{w_i'}{\alpha_i}|^q \leq 1, \qquad \sum_i |\alpha_i|^p \leq 1,$$

where $w_i' = \frac{1}{ab}\|y_i\|_p n^{-1/p} w_i$ and $\alpha_i = \|y_i\|_p n^{-1/p}/b$.

Now we (re-order and) partition $\{\alpha_i\}$ into two parts such that $|\alpha_{u+1}|, \ldots, |\alpha_d| < \epsilon_1^{1/p}$ (the value of ϵ_1 is to be determined later) where $u \leq 1/\epsilon_1$. Since

$$\sum_{i=u+1}^d \alpha_i^{p'} \leq \epsilon_1^{\delta p'} \sum_{i=u+1}^d \alpha_i^p \leq \epsilon_1^{\delta p'},$$

therefore

$$\left\|\sum_{i=u+1}^{d} y_i w_i\right\|_2 \leq \sum_{i=u+1}^{d} \|g_i\|_2 |\alpha_i \cdot \frac{w_i'}{\alpha_i}| \leq n^{1/2} ab \left(\sum_{i=u+1}^{d} |\alpha_i|^{p'}\right)^{1/p'} \left(\sum_{i=u+1}^{d} |\frac{w_i'}{\alpha_i}|^q\right)^{1/q} \leq n^{1/2} ab\, \epsilon_1^{\delta}.$$

By Theorem 2.3, $\forall \epsilon_2$, there exists an ϵ_2-covering of $\sum_{i=1}^{u} w_i y_i$ with $\exp[\lceil \frac{a^2 b^2}{\epsilon_2^2} \rceil \log_2(2u+1)]$ vectors. These vectors also give an $(\epsilon_1^{\delta} ab + \epsilon_2)$-cover for $\sum_{i=1}^{d} w_i y_i$. Now, by setting $\epsilon_1 = (\epsilon/2ab)^{1/\delta}$ and $\epsilon_2 = \epsilon/2$, we obtain the theorem. □

One consequence of this theorem is a potentially refined explanation for the boosting algorithm. In [29], the boosting algorithm has been analyzed by using a technique related to results in [3] which essentially rely on Theorem 2.3 with $p = \infty$. Unfortunately, the bound contains a logarithmic dependency on d (in the general case) which does not seems to fully explain the fact that the performance of the boosting algorithm keeps improving as d increases. However, this seemingly mysterious behavior may be better understood from Theorem 2.5 under the assumption that the data is more restricted than simply being ∞-norm bounded. For example, when the contribution of the wrong predictions is bounded by a constant for all data, then we can regard its p-th norm bounded for some $p < \infty$. In this case, Theorem 2.5 implies dimensional independent generalization which is consistent with experiments. In general, this interpretation can be useful for ∞-norm bounded data with certain sparsity properties.

Another way to remove the dimensional dependency of covering numbers is to introduce a *damping* (*i.e.* to treat dimensions unequally), as demonstrated in the following theorem. The basic idea behind this theorem is to introduce some "compactness" which stabilizes numerical estimation (also see related discussions in Section 4). This technique *shrinks* the space so that the *effective dimension* is reduced. One can easily generalize this theorem by using linear operators instead of sequences of numbers a_j and b_j:

Theorem 2.6 *If* $\sum_j |x_{i,j}/b_j|^p \leq 1$ *and* $\sum_j |w_j/a_j|^q \leq 1$, *where* $1/p + 1/q = 1$, $p \in [2, \infty]$, *and* $|a_j b_j| \leq \frac{c}{2} f(j)^{-s}$ *for some* $c, s > 0$, *where* $f \geq 1$ *is a monotone increasing function, then*

$$\log_2 \mathcal{N}_2(L, \epsilon, n) \leq \lceil \frac{c^2}{\epsilon^2} \rceil \log_2[2 f^{-1}((c/\epsilon)^{1/s}) + 1].$$

Proof. Similar to the proof of the previous theorem. Just note that when $j \geq f^{-1}((c/\epsilon)^{1/s}) = u$, $|a_j b_j| \leq \epsilon/2$. Therefore

$$|\sum_{j \geq u} w_j x_j| \leq \frac{\epsilon}{2} \sum_{j \geq u} |w_j/a_j| \cdot |x_j/b_j| \leq \epsilon/2.$$

□

If we want to apply Theorem 2.2, then it is necessary to obtain bounds for infinity-norm covering numbers. The following theorem give such bounds by using a result from online learning.

Theorem 2.7 *If $\|x_i\|_p \leq b$ and $\|w\|_q \leq a$, where $2 \leq p < \infty$ and $1/p + 1/q = 1$, then $\forall \epsilon > 0$,*

$$\log_2 \mathcal{N}_\infty(L, \epsilon, n) \leq 36(p-1)\frac{a^2 b^2}{\epsilon^2} \log_2[2\lceil 4ab/\epsilon + 2\rceil n + 1].$$

Proof. If $\epsilon > ab$, then since $|x_i^T w| \leq ab$ for all i, we can choose 0 as a cover and the theorem follows trivially. In the following we assume that $\epsilon \leq ab$.

We divide the interval $[-ab - \epsilon/2, ab + \epsilon/2]$ with $m = \lceil 4ab/\epsilon + 2\rceil$ intervals, each of size no larger than $\epsilon/2$. Let $-ab - \epsilon/2 = \theta_0 < \theta_1 < \cdots < \theta_m = ab + \epsilon/2$ be the boundaries of the intervals so that $\theta_i - \theta_{i-1} \leq \epsilon/2$ for all i. For a sample $X_1^n = \{x_1, \ldots, x_n\}$, consider the sets $S_1 = \{(x_i, -\theta_j/a) : i = 1, \ldots, n; j = 0, \ldots, m-1\}$ and $S_2 = \{(-x_i, \theta_j/a) : i = 1, \ldots, n; j = 1, \ldots, m\}$.

$\forall \|w\|_q \leq a$, consider the set of values of w: $x_i^T w - \theta_{j_1(i,w)}$ and $-x_i^T w + \theta_{j_2(i,w)}$, where $j_1(i, w)$ is the maximum index of θ_j such that $x_i^T w - \theta_{j_1(i,w)} \geq \epsilon/2$; and $j_2(i, w)$ is the minimum index of θ_j such that $x_i^T w - \theta_{j_2(i,w)} \leq -\epsilon/2$. This implies that $\forall (y, z)$ such that if $\forall i$: $x_i^T y - z\theta_{j_1(i,w)} > 0$ and $-x_i^T y + z\theta_{j_2(i,w)} > 0$, then $z > 0$ and $\forall i : x_i^T y/z \in (\theta_{j_1(i,w)}, \theta_{j_2(i,w)})$. This implies that $|x_i^T y/z - x_i^T w| < \epsilon$ for all i.

We apply a mistake bound result for online algorithms from [12] which implies that $\forall \|w\|_q \leq a$, let

$$M = 36(p-1)\frac{a^2 b^2}{\epsilon^2} \geq \frac{(p-1)}{(\epsilon/2)^2}(\|w\|_q^q + a^q)^{2/q} \sup_i(\|x_i\|_p^p + (b + \epsilon/2a)^p)^{2/p},$$

then there exists non-negative integer sequences α_i and β_i, such that $\sum_{i=1}^n \alpha_i + \beta_i \leq M$ and if we let

$$(y, az) = f_p(\sum_i \alpha_i(x_i, -\theta_{j_1(i,w)}/a) + \sum_i \beta_i(-x_i, \theta_{j_2(i,w)}/a)),$$

where $f_p(z) = p \cdot \text{sign}(z)|z|^{p-1}$, then $x_i^T y - z\theta_{j_1(i,w)} > 0$ and $-x_i^T y + z\theta_{j_2(i,w)} > 0$ for all i.

It follows from the above discussion that the infinity-norm covering number $\mathcal{N}_\infty(L, \epsilon, n)$, is no more than the number of non-negative integer solutions of

$$\sum_{i,j} n_{i,j} + m_{i,j} \leq M,$$

where (i, j) go through the index of S_1 (and S_2). Since the number of solutions is no more than $(|S_1| + |S_2| + 1)^M$, thus

$$\log_2 \mathcal{N}_\infty(L, \epsilon, n) \leq 36(p-1)\frac{a^2 b^2}{\epsilon^2} \log_2[2\lceil 4ab/\epsilon + 2\rceil n + 1].$$

□

Note that in Theorem 2.7, we have made no attempt to optimize the constants. Since $\theta_0 = -ab - \epsilon/2$ and $\theta_m = ab + \epsilon/2$ are quite artificially introduced, and are only for the purpose of

consistent indexing, thus improvements can be obtained trivially by simply ignoring them. Also note that $\mathcal{N}_2 \leq \mathcal{N}_\infty$, therefore Theorem 2.7 implies dimensional independent 2-norm covering number bounds for $2 \leq p < \infty$, which gives better results than Theorem 2.3 in the sense of dimensional dependency. In the case of $p = \infty$, we show that an entropy condition can be used to obtain dimensional independent covering number bounds. This entropy condition is related to the multiplicative update algorithms widely studied for online learning algorithms. We shall first introduce the following definition:

Definition 2.2 *Let $\mu = [\mu_i]$ be a vector with positive entries such that $\|\mu\|_1 = 1$ (in this case, we call μ a distribution vector). Let $x = [x_i] \neq 0$ be a vector of the same length, then we define the weighted relative entropy of x with respect to μ as:*

$$\mathrm{entro}_\mu(x) = \sum_i |x_i| \ln \frac{|x_i|}{\mu_i \|x\|_1}.$$

It is a well-known fact that relative entropy as defined above is always non-negative, and $\mathrm{entro}_\mu(x) = 0$ only when $|x| = \|x\|_1 \cdot \mu$. Before the main theorem, we need a Lemma that refines and generalizes the discussion in Section 5 of [12] (their result is not directly applicable). Also see [8, 9, 19] and references therein for related techniques. In the following lemma, $x_{j,i}$ indicates the i-th component of vector x_j.

Lemma 2.2 *Let μ be a distribution vector and w be a vector with non-negative entries such that $\|w\|_1 \leq W$. $\forall \delta \in (0, \min_j w^T x_j]$, let*

$$m(\delta) = \frac{2 \sup_i \|x_i\|_\infty^2 W \cdot \mathrm{entro}_\mu(w)}{\delta^2}.$$

Then there exists an integer sequence $j_1, \ldots, j_k$ where $k \leq m(\delta)$, and a vector $\hat{w}$ defined as $\hat{w}_i = \mu_i \exp(\eta \sum_{\ell=1}^k x_{j_\ell,i})$, where $\eta = \delta / W \sup_j \|x_j\|_\infty^2$, so that $\hat{w}^T x_j > 0$ for all j.

Proof. Without loss of generality, we assume that $\|w\|_1 = 1$. Let z be a vector, consider

$$M(z) = \ln \sum_{i=1}^n \mu_i e^{z_i} - w^T z + \sum_{i=1}^n w_i \ln \frac{w_i}{\mu_i},$$

then it is easy to show that $M(z) \geq 0$ for all z.

Assume now that the Theorem is not true, then there exists a sequence of integers $j_1, \ldots, j_k$ where $k > m(\delta)$ such that if we define a sequence of vectors z_ℓ as $z_\ell = z_{\ell-1} + \eta x_{j_\ell}$ with $z_0 = 0$, then $\sum_i \mu_i \exp(z_{\ell-1,i}) x_{j_\ell,i} \leq 0$.

Note that for all pairs of vectors $v, \Delta v$:

$$\frac{d}{dt} \ln \sum_i \mu_i e^{v_i + \Delta v_i t} = \frac{\sum_i \mu_i e^{v_i + \Delta v_i t} \Delta v_i}{\sum_i \mu_i e^{v_i + \Delta v_i t}}$$

and

$$\frac{d^2}{dt^2}\ln\sum_i \mu_i e^{v_i+\Delta v_i t} \le \frac{\sum_i \mu_i e^{v_i+\Delta v_i t}\Delta v_i^2}{\sum_i \mu_i e^{v_i+\Delta v_i t}}.$$

Therefore from the Taylor expansion, we know that there exists $t \in [0,1]$ such that

$$\begin{aligned}
\ln\sum_i \mu_i e^{z_{\ell,i}} \le& \ln\sum_i \mu_i e^{z_\ell - 1,i} + \frac{\sum_i \mu_i e^{z_{\ell-1,i}}\eta x_{j_\ell,i}}{\sum_i \mu_i e^{z_{\ell-1,i}}} + \frac{\eta^2}{2}\frac{\sum_i \mu_i e^{z_{\ell-1,i}+\eta x_{j_\ell,i}t} x_{j_\ell,i}^2}{\sum_i \mu_i e^{z_{\ell-1,i}+\eta x_{j_\ell,i}t}} \\
\le& \ln\sum_i \mu_i e^{z_\ell - 1,i} + \frac{\eta^2}{2}\frac{\sum_i \mu_i e^{z_{\ell-1,i}+\eta x_{j_\ell,i}t} x_{j_\ell,i}^2}{\sum_i \mu_i e^{z_{\ell-1,i}+\eta x_{j_\ell,i}t}} \\
\le& \ln\sum_i \mu_i e^{z_\ell - 1,i} + \frac{\eta^2}{2}\|x_{j_\ell}\|_\infty^2.
\end{aligned}$$

Note that we have used the fact that $\sum_i \mu_i \exp(z_{\ell-1,i})x_{j_\ell,i} \le 0$. We obtain

$$\begin{aligned}
M(z_\ell) - M(z_{\ell-1}) =& \ln\frac{\sum_i \mu_i e^{z_{\ell,i}}}{\sum_i \mu_i e^{z_{\ell-1,i}}} - w^T\cdot \eta x_{j_\ell} \\
\le& \frac{\eta^2}{2}\|x_{j_\ell}\|_\infty^2 - \eta\delta.
\end{aligned}$$

Summing up over ℓ:

$$\begin{aligned}
M(z_k) <& M(z_0) + m(\delta)(\frac{\eta^2}{2}\sup_j \|x_j\|_\infty^2 - \eta\delta) \\
=& \mathrm{entro}_\mu(w) + m(\delta)(\frac{\eta^2}{2}\sup_j \|x_j\|_\infty^2 - \eta\delta) \le 0,
\end{aligned}$$

which is a contradiction. □

Theorem 2.8 *Given a distribution vector μ, If $\|x_i\|_\infty \le b$ and $\|w\|_1 \le a$ and $\mathrm{entro}_\mu(w) \le c$, where we assume that w has non-negative entries, then $\forall \epsilon > 0$,*

$$\log_2 \mathcal{N}_\infty(L,\epsilon,n) \le \frac{36b^2(a^2+ac)}{\epsilon^2}\log_2[2\lceil 4ab/\epsilon + 2\rceil n + 1].$$

Proof. The proof follows the same steps of Theorem 2.7. We let $\mu' = [\mu, 1]/2$ and $w' = [w, a]$. Thus $\|w'\|_1 \le 2a$, and $\mathrm{entro}_{\mu'}(w') \le \mathrm{entro}_\mu(w) + a\ln 2 < a + c$. Similarly, the expansion x_i' of x_i (by adding an entry θ/a) has norm $\|x_i'\|_\infty \le 1.5b$ (we assume that $\epsilon/a \le b$). We shall apply the mistake bound from Lemma 2.2, where we set $\delta = \epsilon/2$ and $W = 2a$, then define M as

$$M = \frac{36(a+c)ab^2}{\epsilon^2} \ge \frac{2}{\delta^2}\sup_i \|x_i'\|_\infty^2 W\cdot \mathrm{entro}_{\mu'}(w').$$

The rest of the proof follows from essentially the same arguments of Theorem 2.7's proof. □

Corollary 2.2 *Given a distribution vector μ, If $\|x_i\|_\infty \leq b$ and $\|w\|_1 \leq a$ and $\mathrm{entro}_\mu(w) \leq c$, then $\forall \epsilon > 0$,*

$$\log_2 \mathcal{N}_\infty(L, \epsilon, n) \leq 1 + \frac{144b^2(2a^2 + ac)}{\epsilon^2} \log_2[2\lceil 4ab/\epsilon + 2\rceil n + 1].$$

Proof. Let $u = \min(w, 0)$ and $v = \min(-w, 0)$, then $w = u - v$ and $\|u\|_1, \|v\|_1 \leq \|w\|_1$. Since for any $L = L_1 - L_2$, we have $\mathcal{N}_\infty(L, \epsilon, n) \leq \mathcal{N}_\infty(L_1, \epsilon/2, n) + \mathcal{N}_\infty(L_2, \epsilon/2, n)$, therefore we only need to show that $\mathrm{entro}_\mu(u) \leq \mathrm{entro}_\mu(w) + \|w\|_1$. To prove this, we shall assume that $\|w\|_1 = 1$ without loss of generality, and $u, v \neq 0$. Since $\|u\|_1 + \|v\|_1 = 1$, thus

$$\|u\|_1 \ln \frac{1}{\|u\|_1} + \|v\|_1 \ln \frac{1}{\|v\|_1} \leq \ln 2 \leq \ln 2 + \sum_i v_i \ln \frac{v_i}{\|v\|_1 \mu_i}.$$

The above inequality can be rewritten as

$$\sum_i u_i \ln \frac{u_i}{\mu_i \|u\|_1} \leq \ln 2 + \sum_i u_i \ln \frac{u_i}{\mu_i} + \sum_i v_i \ln \frac{v_i}{\mu_i}.$$

That is $\mathrm{entro}_\mu(u) \leq \mathrm{entro}_\mu(w) + \ln 2$. □

Note that we don't require the dimension to be finite. However, if the dimension d is finite, and we let $\mu_i = 1/d$, then it is easy to check that $\forall w, \mathrm{entro}_\mu(w) \leq \|w\|_1 \ln d$. Therefore by Corollary 2.2, we obtain the following result which gives a better bound than a similar result in [3] by a logarithmic factor of n.

Corollary 2.3 *If $\|x_i\|_\infty \leq b$ and $\|w\|_1 \leq a$, then $\forall \epsilon > 0$,*

$$\log_2 \mathcal{N}_\infty(L, \epsilon, n) \leq 1 + \frac{144a^2b^2(2 + \ln d)}{\epsilon^2} \log_2[2\lceil 4ab/\epsilon + 2\rceil n + 1].$$

We shall now discuss the relationship among the covering number bounds obtained in this section. Theorem 2.3 uses a reduction technique to generalize a result in [3] (with $p = \infty$ and $q = 1$) which employs the Maurey's Lemma. However, it is very difficult to remove the inherent logarithmic dependence on dimension through this method. As a comparison, Theorem 2.7 (note that $\mathcal{N}_2 \leq \mathcal{N}_\infty$) employs online-learning mistake bound results to remove the $\log d$ dependency by introducing a $\log n$ dependency. This trade-off of $\log d$ and $\log n$ is very natural from the computational point of view since Maurey's Lemma achieves an approximation by selecting columns (relevant features) of the data while an online algorithm achieves an approximation by selecting rows (related to support vectors) of the data. It follows that if $d \ll n$, than Theorem 2.7 gives a better result, and if $n \ll d$, Theorem 2.3 gives better result. Note that in the PAC style bounds, the $\log n$ dependency on the sample size usually does not cause significant problem. However, it is still of interests to obtain covering number bounds that are independent of both n and d.

Theorem 2.4 gives such an example. Although, we speculate that the same claim could be true for all $1/p + 1/q = 1$ and $2 \leq p < \infty$, we are unable to prove (or disprove) this at the moment. However, in Theorem 2.5 and Theorem 2.6, we are able to obtain such results either under the assumption that $1/p + 1/q > 1$, or by a damping technique.

In the proofs of Theorem 2.4 and Theorem 2.5, the effective dimension of the problem are reduced by a compacification of part of the dimensions. This idea of compacification is very important in practical algorithms and is related to numerical stability of formulation (1) discussed in Section 4. Theorem 2.6 achieves such compacification directly by shrinking the non-important dimensions. Compactness also plays an important role in the numerical properties of solving (1) which we will analyze in Section 4. Theorem 2.8 is closely related to Theorem 2.6 in the sense of compacification. However, it employs another regularization condition. If we regard μ_i as a prior measure and w as a posterior measure, then the entropy condition in Theorem 2.8 clearly corresponds to the maximum entropy principal in density estimation. Therefore our covering number result justifies the maximum entropy method from PAC learning point of view. However, we shall mention that in order to obtain a consistency result for the maximum entropy method in the sense of weak convergence in distribution, the PAC style analysis is insufficient and the sequentially weak* compactness [27] of the regularized parameter set becomes important.

As a comparison of Theorem 2.7 and Theorem 2.8, note that as $p \to \infty$, the covering number bound diverges in Theorem 2.7. This is true when we try to regularize the parameter w around the origin, as pointed out in [12]. It is possible to construct a regularization condition around a non-zero vector so that the bound in Theorem 2.7 becomes its limit as $p \to \infty$. Because of Theorem 2.8 and its relation to the well-established maximum entropy principle, it is reasonable to use the entropy condition as in Theorem 2.8 (instead of 1-norm) as the regularization condition for infinity-norm bounded data. The Winnow online multiplicative update algorithm [24] and its continuous version of EG algorithm [19], as well as the classical MART (multiplicative algebraic reconstruction technique) algorithm [11] implicitly include such an entropy condition. In [38], we propose some other numerical algorithms for solving a formulation of (1) with entropy regularization condition. Recently, it was also shown in [20] that the boosting algorithm has a tendency to minimize entropy , although the analysis given in [29] only used the fact that the algorithm tries to maximize margin with fixed 1-norm. In addition, from the discussion after Theorem 2.3, sparse structures presented in the data were often overlooked in the old theoretical analysis when the data is infinite-norm bounded, and either the entropy or the 1-norm regularization condition is used.

We have shown in this section that covering number bounds can be derived from online mistake bounds. This fact suggests that after running a small mistake bound online algorithm repeatedly over the training data, one could expect a comparable generalization error because the effective parameter space that the online learning algorithm has explored is small. We can readily see from the construction of $M(z)$ in the proof of Lemma 2.2 that the weight w and the data z are effectively Lagrange dual variables, and the first term and the third term are corresponding Legendre transforms. This suggests that a proper regularization condition can be constructed or replaced by a properly chosen convex duality. This observation has important consequences in algorithmic design since we can intentionally create auxiliary variables by duality without any regularization.

Finally, the proof technique of Lemma 2.2 is closely related to the potential-reduction method for linear programming (*cf.* [33] and references therein), where a variant of $M(z)$ with a flipped sign for the second term can be used to show the polynomial convergence of certain interior point algorithms. Similar to the proof of Lemma 2.2, the technique of bounding the number of steps is also based on constant reduction of the potential function at each step by choosing an appropriate η based on estimates of its first order term and its second order term in the Taylor expansion. However, since Newton steps are often taken, the proofs for bounding such terms are more involved.

2.3 Consequences of the covering number bounds

In order to apply the covering number bounds to the regularization scheme (1), we would like to extend these results to handle nonlinear compositions of linear functions. Such extensions are also useful for other related methods such as projection pursuit regression, neural networks or radial basis networks, etc. We consider the following system:

$$L([\alpha, w], x) = f(g(\alpha, x) + w^T h(\alpha, x)), \tag{5}$$

where x is the observation, and $[\alpha, w]$ is the parameter.

Definition 2.3 *A function $f : R^{m_1} \to R^{m_2}$ is said to satisfy the Lipschitz condition with parameter γ if $\forall x, y$: $\|f(x) - f(y)\| \leq \gamma \|x - y\|$, where $\|\cdot\|$ denote corresponding norms on the respective spaces.*

The default norm on R is $\|x\| = |x|$ unless otherwise indicated. It is very easy to verify that the following is valid if f is Lipschitz: $\log_2 \mathcal{N}_r(f \circ g, \epsilon, X_1^n) \leq \log_2 \mathcal{N}_r(g, \epsilon/\gamma, X_1^n)$, where "$\circ$" denotes function composition, and the metric in each space corresponds to the norm used in the Lipschitz definition. In order to obtain general covering number bounds on the composition of functions, we need to allow h to be a vector valued function. Note that the covering number results in Section 2.2 are not for vector valued functions — we have skipped such results since they are not very relevant to the regularization problem (1) we are interested in. However, for completeness and background purposes, we present an informally discussion on how to obtain more general covering number results.

The simplest way to obtain a bound for a vector function is by summing the corresponding bounds for its components: $\log_2 \mathcal{N}_r([f_1, \ldots, f_d], \epsilon, n) \leq \sum_i \log_2 \mathcal{N}_r(f_i, \epsilon, n)$, where the metric in the $[f_1, \ldots, f_d]$ is taken to be the infinity norm. Note that if we require the bound to be independent of d, then $\mathcal{N}_r(f_i, \epsilon, n) = 1$ for $i \geq u(\epsilon, n)$ where $u(\epsilon, n)$ is a value independent of d. This can be achieved by the same damping technique used in Theorem 2.6. We can also derive covering number bounds for vector valued functions directly. If we consider w as a matrix in (4), then bounds that generalize the corresponding theorems in Section 2.2 can be obtained. For example, in Theorem 2.4, if we allow w to be matrices with Frobenius norm regularization, and assume that we use 2-norms for all vectors in different spaces, then the same result holds with the same proof. Besides direct generalizations of theorems in Section 2.2, some special methods can be applied for

special problems. An example is to exploit the symmetry property in neural networks to remove the effect of dimensions in inner layers [3].

Since the vector valued covering number bounds discussed above are not relevant to the later part of this paper, we shall not go into details. In the following, we give a result for (5), which shows that if a linear classifier contains a d-dimensional non-regularized part, then this non-regularized part contributes a $O(d)$ term to the logarithmic covering number $\log \mathcal{N}$.

Definition 2.4 *The total variation of a function $f : R \to R$ is defined as*

$$\mathrm{TV}(f, x) = \sup_{x_0 < x_1 \cdots < x_\ell \le x} \sum_{i=1}^{\ell} |f(x_i) - f(x_{i-1})|.$$

We also denote $\mathrm{TV}(f, \infty)$ *as* $\mathrm{TV}(f)$.

Lemma 2.3 *If $L([\alpha, w], x) = f(g(\alpha, x) + w^T h(\alpha, x))$, where $f \in [0, M]$ is monotone and w is a d-dimensional vector. Assume that $f : R \to [0, M]$ is Lipschitz with parameter γ, and assume that $\|w\|_q \le c$, then $\forall \epsilon_1, \epsilon_2 > 0$, and $n > 2(d+1)$:*

$$\log_2 \mathcal{N}_r(L, \epsilon_1 + \epsilon_2, n) \le (d+1) \log_2 [\frac{en}{d+1} \max(\lfloor \frac{M}{2\epsilon_1} \rfloor, 1)] + \log_2 \mathcal{N}_r([g, h], \epsilon_2/\gamma, n),$$

where metric of $[g, h]$ is defined as $|g_1 - g_2| + c\|h_1 - h_2\|_p$ ($1/p + 1/q = 1$).

Proof. Let $\alpha_1, \ldots, \alpha_u$ be an ϵ_2/γ covering of $[g, h]$ in r-norm (enlarge the parameter family if necessary), and

$$L_i(w, x) = f(g(\alpha_i, x) + w^T h(\alpha_i, x)).$$

Note that $\{g(\alpha_i, x) + w^T h(\alpha_i, x) : i = 1, \ldots, u\}$ forms an ϵ_2/γ covering of $g(\alpha, x) + w^T h(\alpha, x)$, therefore by the Lipschitz condition of f, $\{L_i(w, x)\}$ gives an ϵ_2 covering of L in r-norm. We thus only need to show that each L_i can be ϵ_1-covered by at most $[\frac{en}{d+1} \max(\lfloor \frac{M}{2\epsilon_1} \rfloor, 1)]^{d+1}$ vectors in r-norm.

Since the pseudo-dimension (*cf.* [15]) of $w^T y + z$ is at most $d+1$ and a monotone function does not increase it, thus we have

$$\log_2 \mathcal{N}_\infty(L_i, \epsilon_1, n) \le \log_2 \sum_{i=0}^{d+1} \binom{n}{i} \lfloor \frac{M}{2\epsilon_1} \rfloor^i.$$

By a well-known bound $\sum_{i=0}^{d} \binom{n}{i} \le (en/d)^d$ for all $n > 2d$ (*cf.* [5], pp. 218), we obtain

$$\log_2 \mathcal{N}_r(L_i, \epsilon_1, n) \le \log_2 \mathcal{N}_\infty(L_i, \epsilon_1, n) \le (d+1) \log_2 [\frac{en}{d+1} \max(\lfloor \frac{M}{2\epsilon_1} \rfloor, 1)].$$

□

Theorem 2.9 *If $L([\alpha, w], x) = f(g(\alpha, x) + w^T h(\alpha, x))$, where $\mathrm{TV}(f) < \infty$ and f is Lipschitz with parameter γ. Assume also that w is a d-dimensional vector and $\|w\|_q \leq c$, then $\forall \epsilon_1, \epsilon_2 > 0$, and $n > 2(d+1)$:*

$$\log_2 \mathcal{N}_r(L, \epsilon_1 + \epsilon_2, n) \leq (d+1) \log_2 [\frac{en}{d+1} \max(\lfloor \frac{\mathrm{TV}(f)}{2\epsilon_1} \rfloor, 1)] + \log_2 \mathcal{N}_r([g, h], \epsilon_2/\gamma, n),$$

where metric of $[g, h]$ is defined as $|g_1 - g_2| + c\|h_1 - h_2\|_p$ $(1/p + 1/q = 1)$.

Proof. Since $\mathrm{TV}(f, x)$ is monotone and it is easy to verify that $\mathrm{TV}(f, x)$ is Lipschitz with parameter γ, thus the bound is valid with $f(x)$ replaced by $\mathrm{TV}(f, x)$. Note that $\forall x_1, x_2$, $|f(x_1) - f(x_2)| \leq |\mathrm{TV}(f, x_1) - \mathrm{TV}(f, x_2)|$, thus the bound also holds for $f(x)$. □

An interesting observation from Theorem 2.9 is that we can allow some dimensions in (1) to be non-regularized:

Example 2.1 In the above theorem, a special case is when $h(\alpha, x) = h(x)$ independent of α. In this case, the covering number of $[g, h]$ is equivalent to the covering number of g. We can set $c = \infty$, which means that w is not regularized. The bound on covering number for $L([\alpha, w], x) = f(g(\alpha, x) + w^T h(x))$ is

$$\log_2 \mathcal{N}_r(L, \epsilon, n) \leq (d+1) \log_2 [\frac{en}{d+1} \max(\lfloor \frac{\mathrm{TV}(f)}{\epsilon} \rfloor, 1)] + \log_2 \mathcal{N}_r(g, \epsilon/2\gamma, n)$$

for all $\epsilon > 0$ and $n > 2(d+1)$. □

Example 2.2 Consider classification by hyperplane: $L(w, x) = I(w^T x \leq 0)$ where I is the set indicator function. Let $L'(w, x) = f_0(w^T x)$ be another loss function where

$$f_0(z) = \begin{cases} 1 & z < 0 \\ 1 - z & z \in [0, 1] \\ 0 & z > 1 \end{cases}.$$

Instead of using ERM to estimate the parameter that minimizes the risk of L, consider the scheme of minimizing the empirical risk associated with L', under the assumption that $\|x\|_2 \leq b$ and constraint that $\|w\|_2 \leq a$. Denote the estimated parameter by w_n. It follows from Theorem 2.1 and the covering number bound of Theorem 2.4 that

$$P(E_x I(w_n^T x \leq 0) > \inf_w E_x f_0(w^T x) + 16\epsilon) \leq 8 \exp(\frac{-n\epsilon^2}{2} + \lceil \frac{2a^2 b^2}{\epsilon^2} \rceil \ln(4a^2 b^2/\epsilon^2 + 1)).$$

The PAC style bound is: with probability of at least $1 - \eta$:

$$E_x I(w_n^T x \leq 0) \leq \inf_{\|w\|_2 \leq a} E_x f_0(w^T x) + O(\sqrt{\frac{n^{1/2} ab \ln(nab + 2) + \ln \frac{1}{\eta}}{n}}).$$

The corresponding average generalization error is bounded by

$$\inf_{\|w\|_2 \le a} E_x f_0(w^T x) + O(\frac{1}{\sqrt{n}} + \frac{(ab)^{1/2}}{n^{1/4}} \ln^{1/2}(nab+2)).$$

Note that this bound shows that if n is large, w_n is not much worse than the optimal w that minimizes the modified risk $E_x f_0(w^T x)$. This suggests the scheme of minimizing the empirical risk of $E_{emp} f_0(w^T x)$ which has a better convergence behavior than minimizing $E_{emp} I(w^T x \le 0)$ directly.

We can also apply Theorem 2.2 and the covering number bound of Theorem 2.7 to give the following PAC bound: for any $\gamma > 0$, with probability of at least $1-\eta$,

$$E_x I(w_n^T x \le 0) \le E_{X_1^n} I(w_n^T x \le 2\gamma) + \sqrt{\frac{32}{n}(\ln 4 + 36\frac{a^2b^2}{\gamma^2}\ln(2\lceil 4ab/\gamma + 2\rceil n + 1) + \ln\frac{1}{\eta})}.$$

Now, by taking γ_i as $1/2^i$ and $p_i = 1/i(i+1)$ in Corollary 2.1, we obtain with probability of at least $1-\eta$:

$$E_x I(w_n^T x \le 0) \le E_{X_1^n} I(w_n^T x \le 2\gamma) + O(\sqrt{\frac{1}{n}(\frac{a^2b^2}{\gamma^2}\ln(ab/\gamma+2) + \ln n + \ln\frac{1}{\eta})})$$

for all $\gamma \in (0,1]$. At this point, a still needs to be data independent. However, note that the inequality always holds when $ab \ge O(\gamma\sqrt{n})$, therefore by choosing $a_i = \gamma\sqrt{n}/(2^i b)$ with $p_i = 1/i(i+1)$, we obtain with probability of at least $1-\eta$:

$$E_x I(w_n^T x \le 0) \le E_{X_1^n} I(w_n^T x \le 2\gamma) + O(\sqrt{\frac{1}{n}(\frac{a^2b^2}{\gamma^2}\ln(ab/\gamma+2) + \ln n + \ln\ln(\gamma\sqrt{n}/ab+e) + \ln\frac{1}{\eta})})$$

for all $a > 0$ and $\gamma \in (0,1]$. The average generalization error is bounded by

$$E_n \inf_{\gamma,a}[E_{X_1^n} I(w_n^T x \le 2\gamma) + O(\frac{ab/\gamma \cdot \ln^{1/2}(ab/\gamma+2) + \ln^{1/2} n + \ln^{1/2}\ln(\gamma\sqrt{n}/ab+e)}{n^{1/2}})],$$

where $E_{X_1^n}$ is the empirical average and E_n is the expectation with respect to the joint distribution of X_1^n (where each of its component independently taken from the same distribution). This bound is better than the bound obtained with Theorem 2.1 if on average, $E_{X_1^n} I(w_n^T x \le 2\gamma)$ is small with a relatively large γ. However, compared with the asymptotic results in Section 3, PAC bounds are often sub-optimal. □

In the rest of this section, we shall study PAC style bounds for (1). We assume that the data is normalized so that $\sup_x \|x\|_p \le b$ for some $b > 0$. Assume also that $R_g(w) \le h(g(w))$ holds with an increasing function h, where we take $R_g(w)$ to be an appropriate regularization condition on w with respect to the observed data. For example, we can let $R_g(w) = \|w\|_{q'}$ with $q' \le p/(p-1)$ if $2 \le p < \infty$; and let $R_g(w) = \|w\|_1 + \text{entro}_\mu(w)$ if $p = \infty$. We shall also assume that $f, g \ge 0$ and f

is non-increasing for simplicity.

Let $\hat{w}$ be an arbitrarily chosen parameter. Given n random data, let w_n be the solution of (1) under the empirical distribution. It follows that

$$E_{emp} f(w_n^T xy) + \lambda g(w_n) \leq E_{emp} f(\hat{w}^T xy) + \lambda g(\hat{w}).$$

Therefore $g(w_n) \leq f(-\|\hat{w}\|_q b)/\lambda + g(\hat{w})$. Let $a = h(f(-\|\hat{w}\|_q b)/\lambda + g(\hat{w}))$, then $R_g(w_n) \leq a$, where $q = p/(p-1)$. Note that for regularization conditions we described earlier, if we choose $\hat{w} = 0$, then $g(\hat{w}) = 0$. Therefore $a = h(f(0)/\lambda)$. On the other hand, if we can choose $\hat{w}$ so that $P(f(\hat{w}^T xy) = 0) = 1$, then $a = h(g(\hat{w}))$ is independent of λ. In particular, let $f(z) = 0$ when $z \geq 1$, and assume that the data is linearly separable by a large margin with a small $\hat{w}$, then the regularization condition is independent of the choice of λ. In this case, we can choose λ such that it is close to zero. However, if the data is not linearly separable (or separable by a very small margin), then λ appears in the regularization condition. In such case, a non-zero choice of λ is important.

Similar to Example 2.2, we can obtain generalization bounds of (1) using Theorem 2.1 and Theorem 2.7: with probability of at least $1 - \eta$,

$$E_{x,y} I(w_n^T xy \leq 0) \leq E_{emp} f_0(w_n^T xy) + O(\sqrt{\frac{(p-1)n^{1/2} ab \ln(nab+2) + \ln \frac{1}{\eta}}{n}}),$$

for $2 \leq p < \infty$ with $q = p/(p-1)$ regularization. The average generalization error is bounded by

$$E_n E_{emp} f_0(w^T xy) + O(\frac{1}{\sqrt{n}} + p(ab)^{1/2} \ln^{1/2}(nab+2)/n^{1/4}).$$

We can also obtain generalization bounds using Theorem 2.2: with probability of at least $1-\eta$,

$$E_{x,y} I(w_n^T xy \leq 0) \leq E_{emp} I(w_n^T xy \leq 2\gamma) + O(\sqrt{\frac{p-1}{n}(\frac{a^2 b^2}{\gamma^2} \ln(ab/\gamma + 2) + \ln n + \ln \frac{1}{\eta})})$$

for all $\gamma \in (0, 1]$. Also the following bound holds with probability at least $1 - \eta$, for all $\gamma \in (0, 1]$ and λ:

$$E_{x,y} I(w_n^T xy \leq 0) \leq E_{emp} I(w_n^T xy \leq 2\gamma) + O(\sqrt{\frac{p}{n}(\frac{a^2 b^2}{\gamma^2} \ln(\frac{ab}{\gamma} + 2) + \ln n + \ln \ln(\frac{\gamma\sqrt{n}}{ab} + e) + \ln \frac{1}{\eta})}).$$

The corresponding average generalization error is bounded by

$$E_n \inf_{\gamma,\lambda} [E_{emp} I(w_n^T xy \leq 2\gamma) + O(\frac{p}{n^{1/2}}(1 + ab \ln^{1/2}(\frac{ab}{\gamma} + 2)/\gamma + \ln^{1/2} n + \ln^{1/2} \ln(\frac{\gamma\sqrt{n}}{ab} + e)))].$$

In all of the above bounds, the constants in the $O(\cdot)$ notation are universe. Although in the final bound, a as a function of λ has to be derived from a data independent choice of $\hat{w}$, we can replace a by $R_g(w_n)$ and employ techniques from [30] to obtain more refined bounds (for simplicity,

we skip such analysis in this paper). Also note that for $p = \infty$ with entropy regularization, we can derive similar bounds by setting $p = 2$ in the above inequalities. Another important observation is that for our problems, bounds obtained from Theorem 2.1 are inferior to the corresponding bounds obtained from Theorem 2.2. Generally speaking, this will be true if $\mathcal{N}_\infty$ is comparable with $\mathcal{N}_1$.

We shall now discuss the convergence of $R(w_n)$ to $\inf_w R(w)$. We assume that $\inf_w R(w)$ is achieved at w_*. Since f is non-increasing, by results obtained in this section and a modified version of Corollary 2.1, we know that with probability of at least $1 - \eta$, for all $\gamma \in (0, 1]$:

$$E_{x,y} f(w_n^T xy + \gamma) + \lambda g(w_n) \leq E_{emp} f(w_*^T xy) + \lambda g(w_*) + f(-ab) \cdot O(\sqrt{\frac{p}{n}(\frac{a^2b^2}{\gamma^2} \ln(\frac{ab}{\gamma} + 2) + \ln n + \ln \frac{1}{\eta})}).$$

This implies that with probability of at least $1 - \eta$, for all $\gamma \in (0, 1]$:

$$E_{x,y} f(w_n^T xy + \gamma) + \lambda g(w_n) \leq E_{x,y} f(w_*^T xy) + \lambda g(w_*) + f(-ab) \cdot O(\sqrt{\frac{p}{n}(\frac{a^2b^2}{\gamma^2} \ln(\frac{ab}{\gamma} + 2) + \ln n + \ln \frac{1}{\eta})}).$$

Now assume the following uniform convergence condition on the expectation of f (which holds if f is continuous):

$$\lim_{\gamma \to 0} E_{x,y} \sup_w [f(w^T xy) - f(w^T xy + \gamma)] = 0,$$

then it follows that $R(w_n) \to R(w_*)$ in probability as $n \to \infty$. Combined with stability results in Section 4, we can infer that under appropriate choices of f and g, the parameter w_n converges in probability to w_*. The bounds in this section also imply that the rate of convergence is exponential in n. The exponential rate can still be obtained if we replace the requirement that $\|x\|_p \leq b$ with an exponential decay condition on $P(\|x\|_p > b)$.

This consistency result on estimated parameter is the basis for the central limit theorem which characterizes the asymptotic distribution of w_n discussed in Section 3. An implication of the convergence of parameter w_n under regularization is that even though from Theorem 2.9, we can relax the regularization condition for part of the dimensions without affecting good PAC generalization bounds, it is still useful to impose such a condition due to the increased stability for estimating w_n (also see Section 4.1).

In the above discussion, the generalization error bounds are derived from over-estimates of the training errors. This seems to imply that the only role of f in our PAC analysis is to serve as a well-behaved (smooth, convex) upper bound for the mis-classification error. Therefore to minimize f, we also approximately minimize the mis-classification error. Although this point of view is plausible, it does not fully explain the practical effectiveness of the regularization method (1) since the upper bound provided by f is usually a very poor estimate of the mis-classification error. In [38], we shall discuss a much more insightful point of view concerned with the modeling aspects of

f, which is related to the invariance principle mentioned in Section 3.

3 Asymptotic analysis

The convergence results in the previous sections are in the form of convergence in probability, which has a combinatorial flavor. For problems involving differentiable function families with vector parameters, it is often convenient to derive precise asymptotic results by using the differential structure. The following derivation is motived from techniques appeared in [17]. Due to the scope of this paper, we shall only keep the analysis at an intuitive level, and assume that all conditions appearing in the derivation are met. A rigorous treatment that includes the necessary regularity conditions, as well as issues related to the rate of convergence to the asymptotic generalization error, and consequences of the minimax formulation discussed later in this section, will be given in another report.

Assume that the parameter $\alpha \in R^m$ in (2) is a vector and L is a smooth function. Let α^* denote the optimal parameter, then α^* satisfies the following estimation equation:

$$\int \nabla_\alpha L(\alpha^*, x)\, dP(x) = 0,$$

where ∇_α is the derivative with respect to α. We denote $\nabla_\alpha L$ by a vector function Ψ.

Now consider the empirical risk minimization estimator $\alpha_{\rm erm}(X_1^n)$ from n observations X_1^n. Let $P_n(x)$ be the empirical distribution of x with the n observations, then

$$\int \Psi(\alpha_{\rm erm}(X_1^n), x)\, dP_n(x) = 0. \tag{6}$$

If n is large, then $|\alpha^* - \alpha_{\rm erm}(X_1^n)|$ is small with high probability (see discussions at the end of Section 2.3), so that

$$\Psi(\alpha_{\rm erm}(X_1^n), x) - \Psi(\alpha^*, x) \approx \nabla_\alpha \Psi(\alpha^*, x)(\alpha_{\rm erm}(X_1^n) - \alpha^*).$$

Substituting into equation (6), we obtain

$$\int \Psi(\alpha^*, x)\, dP_n(x) \approx \int \nabla_\alpha \Psi(\alpha^*, x)(\alpha^* - \alpha_{\rm erm}(X_1^n, x))\, dP_n(x). \tag{7}$$

Assume that

$$V = \int \nabla_\alpha \Psi(\alpha^*, x)\, dP(x) \tag{8}$$

is non-singular, and let

$$\Psi_n = \int \Psi(\alpha^*, x)\, dP_n(x).$$

Then $\alpha^* - \alpha_{\rm erm} \approx V^{-1}\Psi_n$, and

$$\Delta R(\alpha_{\rm erm}) \approx \int \frac{1}{2}(\alpha_{\rm erm} - \alpha^*)^T \nabla_\alpha \Psi(\alpha^*, x)(\alpha_{\rm erm} - \alpha^*)\, dP(x) \approx \frac{1}{2}\Psi_n^T V^{-1}\Psi_n.$$

The most interesting case is when the central limit theorem holds: $\sqrt{n}\Psi_n$ asymptotically observes a Gaussian distribution with $O(\frac{1}{\sqrt{n}})$ mean and covariance matrix U, where

$$U = \int \Psi(\alpha^*, x)\Psi(\alpha^*, x)^T\, dP(x). \tag{9}$$

The asymptotic mean squared error of the empirical risk minimization is thus

$$E\ (\alpha_{\rm erm} - \alpha^*)^2 \approx \frac{1}{n}{\rm tr}(V^{-2}U), \tag{10}$$

where we use the symbol "tr" to denote the trace of a matrix. More generally, for any evaluation function $h(\alpha)$ such that $\nabla h(\alpha^*) = 0$:

$$E\ h(\alpha_{\rm erm}) \approx h(\alpha^*) + \frac{1}{2n}{\rm tr}(V^{-1}\nabla^2 h \cdot V^{-1}U), \tag{11}$$

where $\nabla^2 h$ is the Hessian matrix of h at α^*. A more complicated formula can be derived for $\nabla h(\alpha^*) \neq 0$, which we shall not describe in this paper. In particular, let $h(\alpha) = R(\alpha)$, then the asymptotic expected generalization error is

$$E\ R(\alpha_{\rm erm}) \approx R(\alpha^*) + \frac{1}{2n}{\rm tr}(V^{-1}U). \tag{12}$$

Note that this approach assumes that the optimal solution is unique. For (1), the uniqueness of solution can be guaranteed by the convexity condition in Lemma 4.1.

If (12) is valid, then it provides a better large sample description than PAC style bounds since the latter predict a generalization performance of at the best $R(\alpha^*) + O(1/\sqrt{n})$ (unless perfect generalization can be achieved: see discussions at the end of Section 2.1). At the first glance, this discrepancy may appear mysterious since it is well-known that many of the PAC bounds are worst case tight for each fixed sample size. However, there are a number of reasons that can explain why this discrepancy is not contradictory: firstly, the worst case tightness in fixed sample size does not imply the asymptotic tightness of PAC bounds with a fixed distribution; also (12) is distribution dependent, and could not handle certain problems of combinatoric nature; another important reason is that the technique for proving the PAC style bounds (see the proof of Theorem 2.2) is inherently sub-optimal asymptotically since it makes the worst case assumption that all points in a covering of the loss function is equally likely to become an empirical estimate of the parameter (for example, see [14] for discussions of this point).

For classification problem (1), the estimation equation becomes

$$E_{x,y} f'(w_*^T xy)xy + \lambda \nabla g(w_*) = 0. \tag{13}$$

Therefore let "cov" denote covariance, we have

$$\begin{aligned}\Psi(w_*, x, y) &= f'(w_*^T xy)xy + \lambda\nabla g(w_*) = f'(w_*^T xy)xy - E_{x,y}f'(w_*^T xy)xy,\\ U(w_*) &= \mathrm{cov}(f'(w_*^T xy)xy),\\ V(w_*) &= E_{x,y}f''(w_*^T xy)xx^T + \lambda\nabla^2 g(w_*).\end{aligned}$$

If $E_{x,y}f''(w^T xy)xx^T$ is a semi-positive definite matrix as in the case of convex f, then we have $V(w) \geq \lambda\nabla^2 g(w)$. It is not difficult to see that if $A \geq B > 0$ and $C \geq 0$, then $\mathrm{tr}(A^{-1}C) \leq \mathrm{tr}(B^{-1}C)$ assuming that A, B and C are all symmetric matrices. The reason is that the trace of a matrix is the sum of its eigenvalues. By the Courant-Fischer minimax theorem for symmetric generalized eigenproblems (*cf.* [10]), it is easy to verify that each eigenvalue of $A^{-1}C$ is no larger than the corresponding eigenvalue of $B^{-1}C$.

Now if we assume that f is convex and $\nabla^2 g(w_*)$ is positive definite, then we obtain the following asymptotic bound for expected generalization error:

$$R(w_*) + \frac{1}{2\lambda n}\mathrm{tr}[(\nabla^2 g(w_*))^{-1}\mathrm{cov}(f'(w_*^T xy)xy)].$$

In this case, the PAC style analysis in Section 2 and the stability analysis in Section 4.1 imply that the rate of convergence of w_{emp} to w_* is independent of dimension under an appropriate distance measure (but it may depend on λ), which suggests that the asymptotic results obtained in this section can be applicable even with relatively small sample size when λ is not close to zero.

Example 3.1 We would like to study a variant of the support vector machine: Consider $L(\alpha, x) = f(\alpha^T x) + \frac{1}{2}\lambda\alpha^2$,

$$f(z) = \begin{cases} 1 - z & z \leq 1 \\ 0 & z > 1 \end{cases}.$$

Because of the discontinuity in the derivative of f, the estimation equation may not hold. However, if we make assumption on the smoothness of the distribution x, then the expectation of the derivate over x can still be smooth, so that the estimation equation (6) and its linear approximation (7) are valid.

Consider the optimal parameter α^* and let $S = \{x : \alpha^{*T} x \in [0, 1]\}$. Note that $\lambda\alpha^* = E_{x\in S}x$, and $U = E_{x\in S}(x - E_{x\in S}x)(x - E_{x\in S}x)^T$. Assume that $\exists\gamma > 0$ s.t. $P(\alpha^{*T} x \leq \gamma) = 0$, then $V = \lambda I + B$ where B is a positive semi-definite matrix. It follows that

$$\mathrm{tr}(V^{-1}U) \leq \mathrm{tr}(U)/\lambda \leq \frac{E_{x\in S}x^2}{E_{x\in S}\alpha^{*T}x}\|\alpha^*\|_2^2 \leq \sup\|x\|_2^2\|\alpha^*\|_2^2/\gamma.$$

Now, consider α_n obtained from observations $X_1^n = [x_1, \ldots, x_n]$ by minimizing empirical risk

associated with loss function $L(\alpha, x)$, then

$$E_x L(\alpha_{emp}, x) \leq \inf_{\alpha} E_x L(\alpha, x) + \frac{1}{2\gamma n} \sup \|x\|_2^2 \|\alpha^*\|_2^2$$

asymptotically. Let $\lambda \to 0$, this scheme becomes the optimal separating hyperplane [36]. The asymptotic bound is better than the bound provided by the PAC bounds with fixed λ. In fact, if we consider an upper bound f of the classification error such that the derivative vanishes at α^*, then it follows that when λ is sufficiently small, the expected classification error goes to zero at a rate faster than $O(1/n)$ asymptotically; and if we choose a smooth function f, then the rate can be faster than any polynomial of $1/n$ asymptotically.

In [38] , the following loss function will also be considered which is numerically easier to minimize: $L(\alpha, x) = g(\alpha^T x) + \frac{1}{2}\lambda\alpha^2$, where

$$g(z) = \begin{cases} (z-1)^2 & z \in [0,1] \\ \rho(z-1)^2 & z > 1 \end{cases}.$$

For linearly separable problems, the method also becomes the optimal separating hyperplane when $\lambda \to 0$ and $\rho \to 0$. $\square$

Note that although the bound obtained in the above example is very similar to the mistake bound for the perceptron online update algorithm, we may in practice obtain much better estimates from (12) by plugging in the empirical data. For a non-square regularization condition, an appropriate transformation of the parameter space w is needed in order to obtain desirable convergence behaviors (*e.g.* dimensional independent convergence to the central limit distribution). This transformation can be achieved by using a *link function*, which has appeared in the proofs of Theorem 2.7 and Lemma 2.2. For our purpose, equation (13) has to be modified to take the transformation into account. Essentially, a link function transforms the metric of regularization in the parameter space so that it behaves like the square metric, and hence all the derivations in this section remain similar. More details about the link function approach in the context of linear classification can be found in [8] and references therein.

If no regularization is used, then the stability results in Section 4.1 do not hold. This implies that the solution does not have the good asymptotic behavior demonstrated in this section. Since the estimated parameter is less stable, therefore a larger portion of the parameter space needs to be explored in order to find an optimal solution of (1), and hence in the worst case, there can be more chance of choosing an inferior parameter.

There are a few different ways to use (11) for the purpose of analyzing learning problems. Assume that $h(\alpha) = E_{x \in D} L_h(\alpha, x)$ is the true risk we are interested in. If we know that the loss function obeys an invariance principle with respect to the distribution so that the optimal solution $\alpha_*(D)$ is also the minimum of $\bar{h}(\alpha) = E_{x \in D} L_{\bar{h}}(\alpha, x)$ for all $\bar{h}$ belonging to a certain function family H and for D belonging to a certain distribution family Γ. In this case, we can define the best

estimation rule corresponding to the empirical estimate with $\bar{h} \in H$ as the minimax solution of

$$\bar{h} = \arg\min_{\bar{h}} \max_{D \in \Gamma} \mathrm{tr}(V_{\bar{h}}^{-2} U_{\bar{h}})$$

if we are interested in the convergence of the parameter (as in regression problems), or

$$\bar{h} = \arg\min_{\bar{h}} \max_{D \in \Gamma} \mathrm{tr}(V_{\bar{h}}^{-1} \nabla^2 h \cdot V_{\bar{h}}^{-1} U_{\bar{h}})$$

if we are interested in the L_h loss. This approach has been used in statistics. For example, the maximum likelihood estimate is asymptotically optimal under quite general conditions in the sense of parameter estimation. A similar criterion was used in [17] to obtain asymptotically optimal robust estimators under certain invariance assumptions of the data distribution.

Another way to apply (11) is to consider a parametric family of functions h_γ that converges to h. We want to decide for a fixed large sample size n, which γ is an appropriate choice so that empirical estimation with h_γ should be employed. Besides some standard statistical techniques (such as cross validation methods), we can use PAC style bounds to select γ with data dependent uniform convergence of h_γ as in Corollary 2.1. However, PAC bounds are often too loose for most real problems, therefore a parameter selection method based on such bounds can often be quite sub-optimal. On the other hand, the asymptotic formulation (11) is precise for large sample sizes. It can also be much more useful for real problems even when the sample size is small. In this case, although a uniform rate of convergence in γ is helpful, it is not crucial for practical purposes. There are also techniques to impose uniform convergence, such as the prior penalty method on the γ space (*cf.* Corollary 2.1). Despite of its relevance to the problem of selecting the regularization parameter γ in (1), we shall not study this topic further in this report due to the limitation of space.

4 Numerical aspects

In the previous discussions, we have indicated that the numerical stability of the solution is very important to learning problems. One reason is that this is required for the convergence of parameter which is the basic assumption of asymptotic analysis. Another reason is that in reality, the PAC learning model (2) is often violated in the sense that $X_1^n = \{x_1, \ldots, x_n\}$ are not independently drawn from a fixed distribution due to a number of reasons: one possible cause is dependency among data; another possible cause is the questionable assumption that the training data is drawn from the same distribution as the test (future) data. In case that the formulation is unstable, a slightly error in model assumption can cause a large change in parameter, and such large change in parameter often has unpredictable behavior as far as the generalization performance is concerned.

From the computational point of view, a stable formulation is very desirable and is often required for designing efficient and robust numerical algorithms. Since a stable and efficient algorithm tends to search a small portion of the parameter space, it can be implied that a stable solution leads to good generalization performance.

To investigate the stability problem, we shall first note that the proposed formulation (1) resembles the standard form of regularization method used for regression problems in numerical analysis and statistics. Since the method was originally proposed in the context of solving ill-posed linear systems $Ax = b$ [31, 32], it is very useful to analyze (1) from the ill-posed system point of view. Although in general, rigorous mathematical treatments of ill-posed problems were often done for systems in infinite dimensional spaces, we shall assume that the spaces we consider are finite dimensional. We replace some traditional functional analysis aspects of ill-posed problems by sensitivity analysis which will lead to meaningful results in finite dimensional spaces.

The readers shall keep in mind that most of our analysis will also hold for infinite dimension: the compactness assumption of Lemma 4.1 of $\{z : g(z) \leq r\}$ can be replaced by weak* compactness and sequentially weak* compactness with appropriate regularity assumptions on f, g and the distribution of (x, y). Note that if we assume that w is in the dual space of a separable normed linear space containing data x, then any bounded closed subset is sequentially weak* compact [27], therefore we can simply assume that $\{z : g(z) \leq r\}$ is bounded which is a relatively mild assumption. For simplicity, we shall skip this type of analysis, although infinite dimensional spaces can occur in learning problems, such as the kernel formulation of support vector machines.

4.1 Stability analysis

For simplicity, we assume that $f \geq 0$ is convex and non-increasing, and $g \geq 0$ is a strictly convex. We also assume that $f(0) > 0$, and both f and g are differentiable.

Lemma 4.1 *If $f \geq 0$ is continuous convex and $g \geq 0$ is continuous strictly convex, then $\forall \lambda > 0$, and any distribution (x, y), the function $E_{x,y} f(w^T xy) + \lambda g(w)$ is strictly convex. Assume also that $E_{x,y} f(w^T xy)$ is a continuous function of w on a closed convex set Ω, and that $\forall r > 0$, the set $\Omega \cap \{z : g(z) \leq r\}$ is compact, then there exists a unique w that minimizes (1) on Ω.*

Proof. To verify that $E_{x,y} f(w^T xy) + \lambda g(w)$ is strictly convex, we simply need to check that $\forall \alpha \in (0, 1)$ and $w_1 \neq w_2$: $E_{x,y} f((\alpha w_1 + (1-\alpha) w_2)^T xy) \leq E_{x,y} \alpha f(w_1^T xy) + (1-\alpha) f(w_2^T xy)$ and $\lambda g(\alpha w_1 + (1-\alpha) w_2) < \lambda(\alpha g(w_1) + (1-\alpha) g(w_2))$.

If $\{z : g(z) \leq r\}$ is compact for all $r > 0$, then $\forall \hat{w} \in \Omega$, (1) is equivalent to minimizing $E_{x,y} f(w^T xy) + \lambda g(w)$ under the condition that $w \in \{z : g(z) \leq g(\hat{w}) + f(\hat{w})/\lambda\}$. Therefore a solution of (1) exists. Note that the set $\{z : g(z) \leq r\}$ is convex, and $E_{x,y} f(w^T xy) + \lambda g(w)$ is strictly convex, therefore the uniqueness of the solution follows from the standard results in convex programming (*cf.* [7], chap 6). □

Under the assumptions of Lemma 4.1, for a fixed distribution of (x, y), it follows that there is a unique solution $w(\lambda)$ that minimizes (1) for $\lambda > 0$. We shall now discuss the stability of $w(\lambda)$ when (1) is only approximately minimized: assume $\epsilon > 0$ and w satisfies

$$E_{x,y} f(w^T xy) + \lambda g(w) \leq E_{x,y} f(w(\lambda)^T xy) + \lambda g(w(\lambda)) + \epsilon. \tag{14}$$

Since g is strictly convex, we can define a distance function $d_g(w, w')$ (also called Bregman function [4]) as $d_g(w, w') = g(w') - g(w) - (w' - w)^T \nabla g(w)$. It has the property that $d_g(w, w') \geq 0$

and $d_g(w, w') = 0$ if and only if $w = w'$.

Since $w(\lambda)$ is the optimal solution of the convex programming problem (1), therefore the first order condition holds:

$$E_{x,y} f'(w(\lambda)^T xy)xy + \lambda \nabla g(w(\lambda)) = 0.$$

Combine this equality with (14), we obtain

$$[E_{x,y} f(w^T xy) - E_{x,y} f(w(\lambda)^T xy) - (w - w(\lambda))^T f'(w(\lambda)^T xy)xy] + \lambda d_g(w(\lambda), w) \leq \epsilon.$$

Since f is convex, the Bregman function of Ef is non-negative, thus we get the following bound:

Theorem 4.1 *Under the assumption of (14), we have*

$$d_g(w(\lambda), w) \leq \epsilon/\lambda.$$

In general, one can obtain a lower bound of d_g by Taylor expansion:

Example 4.1 Consider the regularization function $g(w) = \|w\|_2^2$,

$$\nabla^2 g(w) = 2\|w - w(\lambda)\|_2^2,$$

we obtain $\|w - w(\lambda)\|_2 \leq \sqrt{\epsilon/2\lambda}$.

Consider the regularization function $g(w) = \|w\|_q^q$ with $q \geq 2$. The following inequality holds for any real number $x > 0$ and Δ:

$$|\Delta/2|^q \leq |x + \Delta|^q - x^q - qx^{q-1}\Delta.$$

Therefore we obtain $\|w - w(\lambda)\|_q \leq 2(\epsilon/\lambda)^{1/q}$. $\square$

If a lower bound of d_g can be obtained locally (for example, by local approximation from Taylor expansion), then a perturbation bound of the solution can still be obtained:

Corollary 4.1 *Assume that there exists a continuous function l_g such that $l_g(w_1, w_2) \leq \eta$ implies $l_g(w_1, w_2) \leq d_g(w_1, w_2)$, then if $\epsilon < \lambda\eta$, we have*

$$l_g(w(\lambda), w) \leq d_g(w(\lambda), w) \leq \epsilon/\lambda.$$

Proof. Assume the claim is not true, then by the convexity of g, there exists a w such that $l_g(w(\lambda), w) = \eta$ and $d_g(w(\lambda), w) \leq \epsilon$. It follows from the assumptions that

$$l_g(w(\lambda), w) \leq d_g(w(\lambda), w) \leq \epsilon/\lambda < \eta,$$

which is a contradiction to $l_g(w(\lambda), w) = \eta$. $\square$

Our analysis points out that the approximation required to achieve the same parameter estimation accuracy needs to be decreased linearly with respect to λ. This is consistent with the limiting optimization problem of Theorem 4.2 as $\lambda \to 0$ since a perturbation of ϵ to the limiting objective function in the feasible set contributes $\lambda\epsilon$ to (1).

4.2 Properties of the solution curve

We are interested in the behavior of $w(\lambda)$ as $\lambda \to 0$.

Theorem 4.2 *Under the assumptions of Lemma 4.1, consider the solution curve $w(\lambda)$ of (1).*

1. *if $0 < \lambda_1 < \lambda_2$, then*

$$E_{x,y} f(w(\lambda_1)^T xy) + \lambda_1 g(w(\lambda_1)) \leq E_{x,y} f(w(\lambda_2)^T xy) + \lambda_2 g(w(\lambda_2));$$

2. *if $0 < \lambda_1 < \lambda_2$, then*

$$\begin{aligned} E_{x,y} f(w(\lambda_1)^T xy) &\leq E_{x,y} f(w(\lambda_2)^T xy), \\ g(w(\lambda_1)) &\geq g(w(\lambda_2)); \end{aligned}$$

3. *if $D = \{w : E_{x,y} f(w^T xy) = \inf_w E_{x,y} f(w^T xy)\}$ is nonempty, then as $\lambda \to 0$, $w(\lambda)$ converges to the unique solution of the following problem: $\inf_{w \in D} g(w)$;*

4. *if $D = \{w : E_{x,y} f(w^T xy) = \inf_w E_{x,y} f(w^T xy)\}$ is empty, then*

$$\lim_{\lambda \to 0} w(\lambda) = \infty.$$

Proof. Note that by definition, $E_{x,y} f(w(\lambda_1)^T xy) + \lambda_1 g(w(\lambda_1)) \leq E_{x,y} f(w(\lambda_2)^T xy) + \lambda_1 g(w(\lambda_2))$, therefore *1.* is established.

2.: Note that

$$E_{x,y} f(w(\lambda_1)^T xy) + \lambda_1 g(w(\lambda_1)) \leq E_{x,y} f(w(\lambda_2)^T xy) + \lambda_1 g(w(\lambda_2)), \tag{15}$$

$$E_{x,y} f(w(\lambda_2)^T xy) + \lambda_2 g(w(\lambda_2)) \leq E_{x,y} f(w(\lambda_1)^T xy) + \lambda_2 g(w(\lambda_1)). \tag{16}$$

We multiply the first of the above two inequalities by λ_2/λ_1, and add the resulting inequality to the second inequality:

$$\frac{\lambda_2}{\lambda_1} E_{x,y} f(w(\lambda_1)^T xy) + E_{x,y} f(w(\lambda_2)^T xy) \leq \frac{\lambda_2}{\lambda_1} E_{x,y} f(w(\lambda_2)^T xy) + E_{x,y} f(w(\lambda_1)^T xy).$$

That is,

$$(\frac{\lambda_2}{\lambda_1} - 1)(E_{x,y} f(w(\lambda_1)^T xy) - E_{x,y} f(w(\lambda_2)^T xy)) \leq 0,$$

which implies that $E_{x,y}f(w(\lambda_1)^T xy) \leq E_{x,y}f(w(\lambda_2)^T xy)$. To prove the second part of *2,* we add (15) and (16) to obtain:

$$\lambda_1 g(w(\lambda_1)) + \lambda_2 g(w(\lambda_2)) \leq \lambda_1 g(w(\lambda_2)) + \lambda_2 g(w(\lambda_1)),$$

which implies that $(\lambda_1 - \lambda_2)(g(w(\lambda_1)) - g(w(\lambda_2))) \leq 0$. That is $g(w(\lambda_1)) \geq g(w(\lambda_2))$.

3.: Since $g(w)$ is strictly convex and it is easy to verify that D is a convex set, thus by the same argument used in Lemma 4.1, there is a unique solution of the problem $\inf_{w\in D} g(w)$ as long as D is nonempty. Let $\tilde{w}$ be the solution, then

$$E_{x,y}f(w(\lambda)^T xy) + \lambda g(w(\lambda)) \leq E_{x,y}f(\tilde{w}^T xy) + \lambda g(\tilde{w}).$$

Since $E_{x,y}f(\tilde{w}^T xy) \leq E_{x,y}f(w(\lambda)^T xy)$, we obtain $g(w(\lambda)) \leq g(\tilde{w})$. Therefore Let $D_\lambda = \{w : E_{x,y}f(w^T xy) \leq E_{x,y}f(\tilde{w}^T xy) + \lambda g(\tilde{w})\}$ and $\tilde{D} = \{z : g(z) \leq g(\tilde{w})\}$, then $w(\lambda) \in D_\lambda \cap \tilde{D}$. Note that $D_\lambda \cap \tilde{D}$ is a compact convex set, and for $\lambda_1 \leq \lambda_2$, $D_{\lambda_1} \subset D_{\lambda_2}$, therefore if the claim is not true, then there exists a subsequence of positive $\lambda_1 > \lambda_2 > \cdots$ such that $\lim_i \lambda_i = 0$ and $\hat{w} = \lim_i w(\lambda_i) \neq \tilde{w}$. However, note that $\hat{w} \in \cap_{\lambda>0} D_\lambda = D$, and $g(\hat{w}) = \lim_i g(w(\lambda_i)) \leq \lim_i g(\tilde{w}) = g(\tilde{w})$, thus by definition, $\hat{w} = \tilde{w}$, which is a contradiction.

4.: Assume that there is a sequence of positive values $\lambda_1 > \lambda_2 > \cdots$ such that $\lim_i \lambda_i = 0$ and $w(\lambda_i)$ is bounded, then there is a convergent subsequence of $w(\lambda_i)$. We can assume the subsequence to be the whole sequence λ_i. Let $\hat{w} = \lim_i w(\lambda_i)$, then by the arguments used in *3.*: $\hat{w} \in \cap_{\lambda>0} D_\lambda = D$. This implies that D is not empty, which is a contradiction. □

If (x, y) is linearly separable and we pick f such that $f(z) > 0$ when $z < 1$ and $f(z) = 0$ when $z \geq 1$, then from the above theorem, the limit of $w(\lambda)$ as $\lambda \to 0$ is the solution of the following problem:

$$\inf g(w), \qquad \text{s.t.} P(w^T xy \geq 1) = 1,$$

which gives the optimal margin classifier when we pick $g(w) = w^2$. Clearly, the choice of $g(w) = w^2$ is good when the 2-norm of x is well bounded. If the ∞-norm of x is bounded, and if we only consider w so that each components is positive and $\|w\|_1 = \sum_i w_i = 1$, then it is standard to pick a choice $g(w) = -\sum_i w_i \log w_i$.

To demonstrate that $D = \{w : E_{x,y}f(w^T xy) = \inf_w E_{x,y}f(w^T xy)\}$ may be empty, we consider the case of exponential function $f(z) = \exp(-z)$, and assume that (x, y) is linearly separable by $\tilde{w}$: $P(\tilde{w}^T xy > 0) = 1$. In this case, let $\alpha \to +\infty$, then $E_{x,y}f(\alpha\tilde{w}^T xy) \to 0$, but there does not exist w such that $E_{x,y}f(w^T xy) = 0$. Similar phenomenon can still happen even when (x, y) is not linearly separable. If $f(z) = 0$ for some $z > 0$, then D is always non-empty when (x, y) is drawn from a finite sample. If (x, y) is drawn from an arbitrary infinite sample, D can be empty.

In the following, we shall assume that f and g are sufficiently smooth functions and it is valid to change the order of taking derivative and expectation $E_{x,y}$ (this will be true by simply assuming x is bounded and f sufficiently smooth).

Theorem 4.2 implies that the curve $w(\lambda)$ has interesting properties around $\lambda = 0$. In particular, if (x, y) is linearly separable, and f is appropriately chosen, then $w(\lambda)$ converges to the optimal separating hyperplane in the $g(\cdot)$ criterion. It is useful to study more carefully the behavior of $w(\lambda)$ around $\lambda = 0$. Such study resembles the trajectory analysis in nonlinear programming using interior point/penalty function approaches (*cf.* [7]).

The optimal solution $w(\lambda)$ (denoted as w for simplicity) is the unique solution of the first order condition:

$$E_{x,y} f'(w^T xy)xy + \lambda \nabla g(w) = 0. \tag{17}$$

We differentiate with respect to λ and w in (17), and obtain

$$E_{x,y} f''(w^T xy)xx^T dw + \lambda \nabla^2 g(w) dw + \nabla g(w) d\lambda = 0,$$

where $\nabla^2 g(w)$ indicates the Hessian matrix of $g(w)$ with respect to w. It follows that $w(\lambda)$ satisfies the following differential equation:

$$\frac{dw}{d\lambda} = -(E_{x,y} f''(w^T xy)xx^T + \lambda \nabla^2 g(w))^{-1} \nabla g(w). \tag{18}$$

Let $\tilde{w} = w(0)$. If $E_{x,y} f''(\tilde{w}^T xy)xx^T$ is full rank, then the system is well-behaved at 0. However, in case $E_{x,y} f''(\tilde{w}^T xy)xx^T$ is not full rank (which will happen for example, when (x, y) is linearly separable and $f(z) = 0$ for some $z > 0$ or the problem dimension is larger than the sample size), then more careful analysis has to be carried out. Note that if $\nabla g(\tilde{w}) = 0$, then $\tilde{w}$ achieves both the minimum of $g(\tilde{w})$ and the minimum of $E_{x,y} f(w^T xy)$, therefore $w(\lambda) = \tilde{w}$ for all $\lambda > 0$. We shall now assume that $\nabla g(\tilde{w}) \neq 0$.

We assume that for a norm $\|\cdot\|$ of w, there exists a continuous function $G(\cdot)$ defined on the unit ball $B = \{\Delta w : \|\Delta w\| = 1\}$, such that $\exists k > 0$,

$$E_{x,y}[f'((\tilde{w} + \alpha \Delta w)^T xy) - f'(\tilde{w}^T xy)]xy = \alpha^k G(\Delta w) + o(\alpha^k) \tag{19}$$

uniformly as $\alpha \to 0^+$ when $\Delta w \in B$ (that is, the rate of $o(\alpha^k)$ is independent of Δw).

To give an example for (19), consider $f(z) = 0$ when $z \geq 1$ and $f(z) = (1 - z)^{k+1}$ when $z < 1$. Also assume that the data distribution (x, y) comes from a finite sample (x_i, y_i) $(i = 1, \ldots, n)$ and is separable. Note that by Theorem 4.2, $\tilde{w}^T x_i y_i \geq 1$. Let $j = 1, \ldots, \beta$ denote the indices such that $\tilde{w}^T x_j y_j = 1$ and $\|\cdot\|$ is the 2-norm, then for $j \leq \beta$,

$$\begin{aligned}
& f'((\tilde{w} + \alpha \Delta w)^T x_j y_j) \\
=& f'(1 - \alpha \|x_j\| \cos(\Delta w, -x_j y_j)) \\
=& (1 + k)\alpha^k \|x_j\|^k \max(\cos(\Delta w, -x_j y_j), 0)^k.
\end{aligned}$$

It follows that we can set

$$G(\Delta w) = \frac{1+k}{n} \sum_{j=1}^{\beta} \|x_j\|^k \max(\cos(\Delta w, -x_j y_j), 0)^k.$$

We shall note that in the infinite sample case or as $n \to \infty$, the finite sample analysis is not very meaningful since α needs to be very small so that a sample such that $\tilde{w}^T x_i y_i = 1 + \epsilon$ for small ϵ does not hit the boundary $(\tilde{w} + \alpha \Delta w)^T x_i y_i = 1$. However, (19) can still be valid if the distribution is "smooth".

Under the assumption of (19), we let $w(\lambda) = \tilde{w} + \alpha(\lambda)\Delta w(\lambda)$ where $\|\Delta w(\lambda)\| = 1$. We approximate (17) using (19) and the Taylor expansion of $\lambda \nabla g(w)$ at $\tilde{w}$ to obtain:

$$\alpha(\lambda)^k G(\Delta w(\lambda)) + \lambda \nabla g(\tilde{w}) + o(\lambda + \alpha(\lambda)^k) = 0.$$

Therefore in general when $\nabla g(\tilde{w}) \neq 0$, we have $\lambda = O(\alpha(\lambda)^k)$. In the non-degenerate situation where $\underline{\lim}_{\lambda \to 0} \|G(\Delta w(\lambda))\| \neq 0$:

$$\nabla g(\tilde{w}) = \lim_{\lambda \to 0} -\frac{\alpha(\lambda)^k}{\lambda} G(\Delta w(\lambda)),$$

and $\alpha(\lambda) = O(\lambda^{1/k})$. It follows that the more smooth f is (as k is large), the slower the convergence of $w(\lambda) \to \tilde{w}$ as $\lambda \to 0$ when (x, y) is separable. As we have indicated before: if we take $f(z) = \exp(-z)$ which is in C_∞, then $\tilde{w}$ is infinity, which means that $w(\lambda)$ diverges.

Even though $w(\lambda)$ can converge slowly or be non-differentiable at $\lambda = 0$, we show that $R'(\lambda) = E_{x,y} f(w(\lambda)^T xy)$ is always differentiable at $\lambda = 0$. In fact, since

$$E_{x,y} f(w(\lambda)^T xy) + \lambda g(w(\lambda)) \leq E_{x,y} f(\tilde{w}^T xy) + \lambda g(\tilde{w}),$$

therefore we always have

$$E_{x,y} f(w(\lambda)^T xy) - E_{x,y} f(\tilde{w}^T xy) \leq \lambda [g(\tilde{w}) - g(w(\lambda))].$$

Since $\lim_{\lambda \to 0} g(w(\lambda)) = g(\tilde{w})$, therefore

$$R'(\lambda) - R'(0) = o(\lambda),$$

which means that the derivative of $E_{x,y} f(w(\lambda)^T xy)$ at $\lambda = 0$ is zero. We also obtain an inequality

$$R'(\lambda) - R'(0) \leq \lambda g(\tilde{w}),$$

indicating that the convergence rate of $R'(\lambda)$ depends on the magnitude of $\tilde{w}$.

4.3 Perturbation analysis

In the following analysis, we are interested in the behavior of the solution of (1) under a modification of the system. Such modification can arise from different sources, such as error caused by numerical procedures for solving (1), or violation of i.i.d. sample assumption, or violation of the assumption that training sample and test sample are drawn from the same distribution, or even mis-labels in the training set. We assume that the perturbation can either be a small change of coordinate $x \to x + \Delta x$, or a small percentage of wrong data including possible mis-labeling of y or a large change of Δx. We can formulate the problem as follows:

$$w_{\Delta x,\sigma}(\lambda) = \inf_w E_{x,y,\Delta x,\sigma} f(w^T(x + \Delta x)\sigma y) + \lambda g(w), \tag{20}$$

where $E_{\Delta x}\|\Delta x\|$ is small under certain norm of x and $\sigma = \pm 1$ is a random sign parameter having a small probability to be -1.

In the following, we study the solution behavior under the assumption that

$$|E_{x,y,\Delta x,\sigma} f(w^T(x + \Delta x)\sigma y) - E_{x,y} f(w^T xy)| \leq \epsilon \tag{21}$$

when $d_g(w(\lambda), w) \leq \eta$, where $\epsilon < \eta/2\lambda$. Under this model of perturbation, outliers that may cause extremely damaging effect to the estimation are not considered. Since

$$\begin{aligned} &E_{x,y,\Delta x,\sigma} f(w_{\Delta x,\sigma}(\lambda)^T(x + \Delta x)\sigma y) + \lambda g(w_{\Delta x,\sigma}(\lambda)) \\ \leq &E_{x,y,\Delta x,\sigma} f(w(\lambda)^T(x + \Delta x)\sigma y) + \lambda g(w(\lambda)), \end{aligned}$$

therefore if $d_g(w(\lambda), w_{\Delta x,\sigma}(\lambda)) \leq \eta$, then

$$E_{x,y} f(w_{\Delta x,\sigma}(\lambda)^T xy) + \lambda g(w_{\Delta x,\sigma}(\lambda)) \leq E_{x,y} f(w(\lambda)^T xy) + \lambda g(w(\lambda)) + 2\epsilon.$$

By Corollary 4.1, we obtain the following bound:

Theorem 4.3 *Under the assumption of (21), we have*

$$d_g(w(\lambda), w_{\Delta x,\sigma}(\lambda)) \leq 2\epsilon/\lambda.$$

This theorem implies that when $\lambda > 0$, the solution $w(\lambda)$ is stable under a small perturbation. However, when $\lambda \to 0$, $w(\lambda)$ is more and more sensitive to perturbation. At $\lambda = 0$, for many problems, a small perturbation of the data can cause a large modification of $\tilde{w}$.

For example, assume that (x, y) is linearly separable and we choose a function f such that $f(z) > 0$ when $z < 1$ and $f(z) = 0$ when $z \geq 1$, then $P(\tilde{w}^T xy \geq 1) = 1$. If we add a non-zero percentage of (x', y') such that $\|x'\| = \epsilon$ is small and $\tilde{w}^T x'y' > 0$, then by the characterization in Theorem 4.2, the solution $\tilde{w}'$ must satisfy the condition $\|\tilde{w}'\| \geq 1/\epsilon$ with norm of w dual of that of x'. Such modification has completely unpredictable effect since $\tilde{w}'$ can approach ∞ under small perturbation even if $\tilde{w}$ is small.

Note that bounds from the PAC analysis in Section 2 depends heavily on $\|\tilde{w}'\|$, therefore a

small perturbation of the data can cause a significant decrease of generalization performance. This analysis again demonstrate the importance of regularization which improves the numerical stability of the solution.

5 Discussions

In this paper, we have studied some theoretical aspects of using the regularization formulation (1) for classification problems. We show that with appropriate regularization condition, we can achieve the same dimensional independent generalization performance enjoyed by support vector machines.

In Section 2, the separation concept introduced in Theorem 2.2 suggests that the "margin" concept for linear classifiers can be extended naturally to general problems. The important feature of this theorem is its independent of the smoothness of the loss function itself. Note that in general, the covering numbers of the loss function depend on such smoothness characterized by the Lipschitz condition (see Theorem 2.9). Recently in [25, 39], McAllester and Zhang studied randomized algorithms that select posterior distributions inducing small average risks under certain regularization conditions. The dimensional independent covering number bounds provided in this paper explain naturally why these algorithms can give good generalization performance within the traditional PAC analysis framework. However, the techniques used in their papers are very different from techniques employed in this paper. In particular, their results are better than results we would obtain with a direct application of our covering number bounds to their problems.

Results provided in Section 3 illustrate that the PAC bounds are often asymptotically suboptimal. A desirable asymptotic behavior of a learning algorithm requires the numerical stability for solving the optimization problem (1). These numerical issues has been investigated in Section 4. We have also demonstrated that numerical stability issue becomes very important under a small perturbation of the system. Such stability requires a non-zero regularization parameter λ and is closely related to techniques for solving ill-posed problems in traditional numerical mathematics.

Finally, a good PAC generalization bound can be obtained only when the data and the parameter are small. In case of nearly separable problems and small regularization parameter, this requires that the magnitude of data $\|x\|$ is clustered since if there exist some very small data, then the optimal solution would be large as explained at the end of Section 4.3, which leads to poor PAC bound. This implies that our theory can predict good performance only when the projection of x onto $w^T x$ tends to be clustered around a positive point. Clearly, this is also the phenomenon we expect if we simply try to find w by optimizing the degree of such clustering of $w^T x$ using the least square regression (where we pick $f(z) = (z-1)^2$ and $g(w) = w^2$).

On one hand, our theory can be used to predict good generalization performance of the least square algorithm which has been very successfully used in the literature; on the other hand, we seem to conclude from our theory that support-vector machine like regularization techniques will have good performance only when the least square regression also works well (without considering outliers). In particular, we shall not expect miracle results from support vector machine like algorithms when the simple least square method fails. A very important feature of the least square algorithm is that we can conceptually think it as being derived from the assumption that the

distribution of xy is an isotropic Gaussian. Similarly, we can pose (1) as an assumption on the distribution of xy (note we allow such distribution to be improper). Some aspects of this view point will be explored in [38].

References

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):615–631, 1997.

[2] A.R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

[3] P.L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998.

[4] L. M. Bregman. The relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7:200–217, 1967.

[5] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition.* Springer-Verlag, New York, 1996.

[6] R.M. Dudley. *A course on empirical processes*, volume 1097 of *Lecture Notes in Mathematics.* 1984.

[7] Anthony V. Fiacco and Garth P. McCormick. *Nonlinear Programming: Sequential Unconstrained Minimization Techniques.* John Wiley and Sons, New York, 1968. Republished by SAIM, Philadelphia, 1990.

[8] C. Gentile and N. Littlestone. The robustness of the p-norm algorithms. In *COLT'99*, pages 1–11, 1999.

[9] C. Gentile and M. K. Warmuth. Linear hinge loss and average margin. In *Proc. NIPS'98*, 1998.

[10] G.H. Golub and C.F. Van Loan. *Matrix computations.* Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

[11] R. Gordon, R. Bender, and G. T. Herman. Algebraic reconstruction techniques. *Journal of Theoretical Biology*, pages 471–481, 1970.

[12] A.J. Grove, N. Littlestone, and D. Schuurmans. General convergence results for linear discriminant updates. In *Proc. 10th Annu. Conf. on Comput. Learning Theory*, pages 171–183, 1997.

[13] D. Haussler. Generalizing the PAC model: sample size bounds from metric dimension-based uniform convergence results. In *Proc. 30th IEEE Symposium on Foundations of Computer Science*, pages 40–45, 1989.

[14] D. Haussler, M. Kearns, H.S. Seung, and N. Tishby. Rigorous learning curve bounds from statistical mechanics. *Machine Learning*, 25(2-3):195–236, 1996.

[15] David Haussler and Philip M. Long. A generalization of Sauer's lemma. *J. Combin. Theory Ser. A*, 71(2):219–240, 1995.

[16] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.

[17] P.J. Huber. *Robust statistics*. John Wiley & Sons Inc., New York, 1981. Wiley Series in Probability and Mathematical Statistics.

[18] L.K. Jones. A simple lemma on greedy approximation in hilbert space and convergence rates for projection pursuit regression and neural network training. *Ann. Statist.*, 20:608–613, 1992.

[19] J. Kivinen and M. K. Warmuth. Additive versus exponentiated gradient updates for linear prediction. *Journal of Information and Computation*, 132:1–64, 1997.

[20] J. Kivinen and M. K. Warmuth. Boosting as entropy projection. In *COLT'99*, pages 134–144, 1999.

[21] A.N. Kolmogorov. Asymptotic characteristics of some completely bounded metric spaces. *Dokl. Akad. Nauk. SSSR*, 108:585–589, 1956.

[22] A.N. Kolmogorov and V.M. Tihomirov. ϵ-entropy and ϵ-capacity of sets in functional spaces. *Amer. Math. Soc. Transl.*, 17(2):277–364, 1961.

[23] Wee Sun Lee, P.L. Bartlett, and R.C. Williamson. Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory*, 42(6):2118–2132, 1996.

[24] N. Littlestone. Learning quickly when irrelevant attributes abound: a new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.

[25] David McAllester. PAC-Bayesian model averaging. In *COLT'99*, pages 164–170, 1999.

[26] D. Pollard. *Convergence of stochastic processes*. Springer-Verlag, New York, 1984.

[27] Walter Rudin. *Functional Analysis*. McGraw-Hill, New York, 2nd edition, 1991.

[28] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory (Series A)*, 13:145–147, 1972.

[29] R.E. Schapire, Y. Freund, P. Bartlett, and Wee Sun Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. *The Annals of Statistics*, to appear.

[30] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Trans. Inf. Theory*, 44(5):1926–1940, 1998.

[31] A.N. Tikhonov. On solving ill-posed problem and method of regularization. *Doklady Akademii Nauk USSR*, 153:501–504, 1963.

[32] A.N. Tikhonov and V.Y. Arsenin. *Solution of ill-posed problems*. W.H. Winston, Washington, DC., 1977.

[33] Michael J. Todd. Potential-reduction methods in mathematical programming. *Math. Programming*, 76(1, Ser. B):3–45, 1997. Interior point methods in theory and practice (Iowa City, IA, 1994).

[34] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[35] V.N. Vapnik. *Estimation of dependences based on empirical data*. Springer-Verlag, New York, 1982. Translated from the Russian by Samuel Kotz.

[36] V.N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag, New York, 1995.

[37] V.N. Vapnik and A.J. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Applications*, 16:264–280, 1971.

[38] T. Zhang and F.J. Oles. Algorithms for training linear classifiers by regularization methods. manuscript.

[39] Tong Zhang. Theoretical analysis of a class of randomized regularization methods. In *COLT'99*, pages 156–163, 1999.