Data Dependent Concentration Bounds for Sequential Prediction Algorithms

Tong Zhang

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

Abstract. We investigate the generalization behavior of sequential prediction (online) algorithms, when data are generated from a probability distribution. Using some newly developed probability inequalities, we are able to bound the total generalization performance of a learning algorithm in terms of its observed total loss. Consequences of this analysis will be illustrated with examples.

1 Introduction

In statistical learning, we are interested in predicting output $Y \in \mathcal{Y}$ based on observation $X \in \mathcal{X}$. Given a set of n training examples $Z_1^n = \{Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n)\}$, a learning algorithm $\mathcal{A}$ produces a function $\mathcal{A}(Z_1^n; \cdot)$ on $\mathcal{X}$. With a future example $Z_{n+1} = (X_{n+1}, Y_{n+1})$, it produces an output $\mathcal{A}(Z_1^n; X_{n+1})$, and suffers a loss $L(\mathcal{A}(Z_1^n; X_{n+1}), Y_{n+1})$. Assume that the data are generated from an unknown underlying probability distribution D, then the instantaneous risk of the function produced by the algorithm is defined as the expected loss:

$$\mathbf{E}_{Z_{n+1} \sim D}\, L(\mathcal{A}(Z_1^n; X_{n+1}), Y_{n+1}).$$

In statistical learning, we assume that the training data $Z_1, \ldots, Z_n$ are drawn from the same underlying distribution D as the test data. In this paper, we are interested in the concentration of the total instantaneous generalization risk

$$\sum_{i=1}^{n} \mathbf{E}_{Z_i} L(\mathcal{A}(Z_1^{i-1}; X_i), Y_i) \tag{1}$$

to the total empirical loss of the algorithm on the training data

$$\sum_{i=1}^{n} L(\mathcal{A}(Z_1^{i-1}; X_i), Y_i). \tag{2}$$

The former is the generalization behavior of the algorithm on the test data, and the latter is the online performance of the algorithm on the training data. The problem of estimating (1) in terms of (2) has been investigated in [3, 4, 10]. There are two motivations for studying this problem. One is that this gives a probability inequality for the performance of online algorithm on future test data based on the observable "mistake" it makes on the training data. Such a concentration bound can also be used to convert many known online learning mistake bound results into PAC style probability bounds. The second motivation is somewhat different. As pointed out in [3], if we use the total empirical risk of an algorithm as a criterion to select the best learning algorithm (that is, we want to choose the algorithm with the smallest total risk), then the concentration behavior is similar to using n independent random samples. It can thus be argued that the total empirical risk of an algorithm makes better use of the data (than say, cross-validation), and thus is theoretically an attractively quantity for the purpose of model selection. We shall discuss both aspects in the paper.

The purpose of this paper is to develop data-dependent estimates of the total generalization performance (1) based on the observed total loss (2). In order to do so, we need to prove some new probability inequalities for dependent random variables that are suitable for this purpose.

2 Conditional probability inequalities for sums of dependent random variables

We consider a sequence of possibly dependent random variables $Z_1, Z_2, \ldots, Z_n$. For each k, Z_k may depend on the preceding random variables $Z_1, \ldots, Z_{k-1}$. Consider also a sequence of functionals $\xi_k(Z_1, \ldots, Z_k)$ ($k = 1, \ldots, n$). For example, in online mistake bound analysis, we may let $\xi_k = 1$ if a mistake is made on the k-th example, and $\xi_k = 0$ otherwise. Denote by $\mathbf{E}_{Z_k} \xi_k(Z_1, \ldots, Z_k)$ the conditional expectation of ξ_k with respect to Z_k, conditioned on $Z_1^{k-1} = \{Z_1, \ldots, Z_{k-1}\}$. Given an observed sequence $Z_1, \ldots, Z_n$, we are interested in the following two quantities:

$$s_n = \frac{1}{n} \sum_{i=1}^{n} \xi_i(Z_1, \ldots, Z_i), \quad \mu_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}_{Z_i} \xi_i(Z_1, \ldots, Z_i). \tag{3}$$

The first quantity is the empirical average of ξ_k, and the second quantity is the average (over k) of conditional expectation of ξ_k with respect to Z_k. We are interested in showing that s_n and μ_n are close with large probability. Note that if we let $Z_k = (X_k, Y_k)$ and $\xi_k = L(\mathcal{A}(Z_1^{k-1}; X_k), Y_k)$, then these two quantities can be interpreted as the total empirical and generalization risks of the learning algorithm $\mathcal{A}$ in equations (2) and (1).

The starting point of our analysis is the following simple equality.

Lemma 1. *Consider a sequence of random functionals $\xi_1(Z_1), \ldots, \xi_n(Z_1, \ldots, Z_n)$. We have*

$$\mathbf{E}_{Z_1, \ldots, Z_n} \exp\left(\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \ln \mathbf{E}_{Z_i} e^{\xi_i} \right) = 1.$$

Proof. We prove the lemma by induction on n. When $n = 1$, the equality is easy to verify. Assume that the claim holds for all $n \leq k$. Now for $n = k + 1$, we have

$$\begin{aligned}
&\mathbf{E}_{Z_1,\ldots,Z_n} \exp\left(\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \ln \mathbf{E}_{Z_i} e^{\xi_i}\right)\\
=&\mathbf{E}_{Z_1,\ldots,Z_{n-1}} \left[\exp\left(\sum_{i=1}^{n-1} \xi_i - \sum_{i=1}^{n-1} \ln \mathbf{E}_{Z_i} e^{\xi_i}\right) \mathbf{E}_{Z_n} \exp(\xi_n - \ln \mathbf{E}_{Z_n} e^{\xi_n})\right]\\
=&\mathbf{E}_{Z_1,\ldots,Z_{n-1}} \exp\left(\sum_{i=1}^{n-1} \xi_i - \sum_{i=1}^{n-1} \ln \mathbf{E}_{Z_i} e^{\xi_i}\right) = 1.
\end{aligned}$$

Note that the last equation follows from the induction hypothesis.

The following result is a direct consequence of Lemma 1, which we will use to develop more concrete concentration bounds later in the paper. In the literature, related inequalities have been used to derive conditional probability inequalities for Martingales. The technique used here simplifies and improves such results. Some tight probability bounds suitable for our purpose can be obtained as consequences of the following lemma.

Lemma 2. *Consider a sequence of random functionals* $\xi_1(Z_1), \ldots, \xi_n(Z_1, \ldots, Z_n)$. *Then* $\forall t \geq 0$ *and* ρ,

$$\Pr\left[-\sum_{i=1}^{n} \ln \mathbf{E}_{Z_i} e^{-\rho\xi_i} \geq \rho \sum_{i=1}^{n} \xi_i + t\right] \leq e^{-t}.$$

Proof. Let $\xi(\rho) = -\sum_{i=1}^{n} \ln \mathbf{E}_{Z_i} e^{-\rho\xi_i} - \rho \sum_{i=1}^{n} \xi_i$, then we have from Lemma 1: $\mathbf{E}\, e^{\xi(\rho)} = 1$. Now $\forall t$, we have

$$\Pr(\xi(\rho) \geq t)e^t \leq \mathbf{E}\, e^{\xi(\rho)} = 1.$$

Therefore $\Pr(\xi(\rho) \geq t) \leq \exp(-t)$.

Remark 1. Both in Lemma 1 and Lemma 2, the fixed size n can be replaced by a random stopping time that depends on $Z_1, \ldots, Z_n$. We can simply define $\xi_m = 0$ for $m > n$ when the sequence stops at n after seeing $Z_1, \ldots, Z_n$.

Remark 2. Given a random variable Z, the function $\ln \mathbf{E}_Z e^{-\rho Z}$ of ρ is often referred to as its logarithmic moment generating function. It is used in the large deviation literature to obtain tight asymptotic tail probability estimates. The left side of Lemma 2 is the sum of (conditional) logarithmic moment generating functions of ξ_k with respect to Z_k. The bound obtain is essentially identical to the large deviation bounds for independent variables. Therefore, we are able to translate well-known inequalities in the independent setting to the dependent setting with appropriate estimations of logarithmic moment generating functions. This is the approach we will take later on.

Based on Lemma 2, we are now ready to derive results that are direct generalizations of the corresponding cases for independent variables, using appropriate estimates of the logarithmic moment generating functions. These generalizations are the main results of this section.

2.1 Conditional Hoeffding inequalities

For a bounded random variable $\xi \in [0,1]$, it is well-known that its logarithmic moment generating function can be estimated as (see [9]):

$$\ln \mathbf{E} e^{-\rho\xi} \leq \ln\left[1 + (e^{-\rho} - 1)\mathbf{E}\xi\right]. \tag{4}$$

In fact, this is a simple consequence of Jensen's inequality. Using this estimate, we can obtain

Lemma 3. *Assume that $\xi_k \in [0,1]$ for all $k = 1, \ldots, n$. Then $\forall t \geq 0$ and ρ,*

$$\Pr\left[-\ln\left[1 + \mu_n(e^{-\rho} - 1)\right] \geq \rho s_n + t\right] \leq e^{-nt},$$

where μ_n and s_n are defined in (3).

Proof. Using the concavity of logarithm, we have

$$\sum_{i=1}^{n} \ln \mathbf{E}_{Z_i} e^{-\rho\xi_i} \leq n \ln\left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{E}_{Z_i} e^{-\rho\xi_i}\right].$$

Now using (4) on the right hand side of the above inequality, we obtain the result as a direct consequence of Lemma 2.

Theorem 1. *Under the conditions of Lemma 3. We have*

$$\Pr\left[\mu_n \geq \frac{\rho s_n + t}{1 - e^{-\rho}}\right] \leq e^{-nt}, \quad \Pr\left[\mu_n \geq s_n + \sqrt{t/2}\right] \leq e^{-nt}.$$

Proof. Using the fact $-\ln(1-x) \geq x$, we obtain from Lemma 3 that with probability at most e^{-nt},

$$\mu_n(1 - e^{-\rho}) \geq \rho s_n + t.$$

This implies the first inequality.

For the second inequality, we substitute the following bound (which can be verified using Taylor expansion around $\rho = 0$; for example, see [9])

$$-\ln\left[1 + x(e^{-\rho} - 1)\right] \geq \rho x - \frac{\rho^2}{8}$$

into Lemma 3: with probability at most e^{-nt},

$$\mu_n - s_n \geq \frac{t + \rho^2/8}{\rho}.$$

Now take $\rho = \sqrt{t/8}$, we obtain the second inequality.

The second inequality is well-known [1]. We simply reproduce it here. The first inequality is superior (with any fixed ρ) when s_n is small. However, it is not tight when s_n is large. The best possible inequality can be obtained by picking the optimal ρ in Lemma 3. This is what we shall explore next.

Before introducing the next theorem, we shall introduce the following definitions: $\forall \alpha, \beta \in [0,1]$ and $t \geq 0$:

$$\begin{aligned}
&\mathrm{KL}(\alpha||\beta) = \alpha \ln(\alpha/\beta) + (1-\alpha)\ln((1-\alpha)/(1-\beta)),\\
&\mathrm{KL}_2^{-1}(\alpha||t) = \sup\{\beta : \mathrm{KL}(\alpha||\beta) \leq t\}.
\end{aligned}$$

Theorem 2. *Under the conditions of Lemma 3. We have* $\forall \alpha \in [0,1]$ *and* $t \geq 0$*:*

$$\Pr\left[\mu_n \geq \mathrm{KL}_2^{-1}(\alpha||t), s_n \leq \alpha\right] \leq e^{-nt}.$$

Proof. We know from Lemma 3 that $\forall \rho \geq 0$ and $\beta \in [\alpha, 1]$, the following two inequalities hold with probability of at most e^{-nt}:

$$\mu_n \geq \beta, \quad s_n \leq \alpha, \quad -\rho\alpha - \ln\left[1 + \beta(e^{-\rho} - 1)\right] \geq t.$$

Since the claim holds for all $\rho \geq 0$, we may take the parameter ρ that maximizes the left hand side of the third inequality. That is, we take $\rho = \ln(\beta(1-\alpha)) - \ln(\alpha(1-\beta))$, and the third inequality becomes $\mathrm{KL}(\alpha||\beta) \geq t$. Now, let $\beta = \mathrm{KL}_2^{-1}(\alpha||t)$, the third inequality is trivially satisfied. We thus obtain the statement of the theorem.

The function $\mathrm{KL}_2^{-1}(\alpha||t)$ may not be intuitive at first sight. The following result gives a more intuitive form, which can be used to replace $\mathrm{KL}_2^{-1}(\alpha||t)$ in Theorem 2 as well as other theorems in the sequel.

Proposition 1. *The following bound holds for all* $\alpha \in [0,1]$ *and* $t \geq 0$*:*

$$\mathrm{KL}_2^{-1}(\alpha||t) \leq \alpha + \sqrt{2\alpha(1-\alpha)t} + 1.5(1-\alpha)t.$$

Proof. Let $\Delta\alpha = \sqrt{2\alpha(1-\alpha)t + 0.75^2(1-\alpha)^2t^2} + 0.75(1-\alpha)t$. In the following, we assume $\alpha + \Delta\alpha \leq 1$ since the bound holds trivially otherwise. Using Taylor expansion, we have

$$\begin{aligned}
&\mathrm{KL}(\alpha||\alpha+\Delta\alpha) = \alpha \ln\left(1 - \frac{\Delta\alpha}{\alpha+\Delta\alpha}\right) - (1-\alpha)\ln\left(1 - \frac{\Delta\alpha}{1-\alpha}\right)\\
\geq &\alpha\left[-\frac{\Delta\alpha}{\alpha+\Delta\alpha} - \frac{\Delta\alpha^2}{2(\alpha+\Delta\alpha)^2}\frac{1-\frac{\Delta\alpha}{3(\alpha+\Delta\alpha)}}{1-\frac{\Delta\alpha}{(\alpha+\Delta\alpha)}}\right] - (1-\alpha)\left[-\frac{\Delta\alpha}{1-\alpha} - \frac{\Delta\alpha^2}{2(1-\alpha)^2}\right]\\
= &\frac{\Delta\alpha^2}{2(\alpha+\Delta\alpha)} + \frac{\Delta\alpha^3}{6(\alpha+\Delta\alpha)^2} + \frac{\Delta\alpha^2}{2(1-\alpha)}\\
\geq &\frac{\Delta\alpha^2}{2(\alpha+0.75\Delta\alpha)} + \frac{\Delta\alpha^2}{2(1-\alpha)} \geq \frac{\Delta\alpha^2}{2(1-\alpha)(\alpha+0.75\Delta\alpha)} = t.
\end{aligned}$$

This implies that $\mathrm{KL}_2^{-1}(\alpha||t) \leq \alpha + \Delta\alpha$.

The bound in the Theorem 2 is asymptotically best possible for large deviation probability with fixed t since it matches the large deviation lower bound for independent random variables (this claim is also true for moderate deviation when t decreases sufficiently slower than $O(1/n)$). However, in the above theorem, we require that α is chosen in advance. If we remove this condition, a slightly weaker data dependent inequality still holds. The extra penalty of the resulting deviation is no more than $O(\ln n/n) = o(1)$; consequently in the large deviation situation (with fixed t), the bound is also asymptotically the best possible. However, it might be possible to improve the extra $O(\ln n)/n$ penalty we pay for achieving data-dependency because our proof technique may be suboptimal.

The technique we use is rather standard in proving data-dependent generalization bounds in the statistical learning theory literature. The application here is new. We shall state a general result as a lemma (which will also be used later), and then use it to derive a more concrete theorem.

Lemma 4. *Under the conditions of Lemma 3. Consider a finite sequence* $0 \leq \alpha_1 \leq \cdots \leq \alpha_m = 1$*, and a sequence* $\{\Delta t_\ell\}$ *such that* $\sum_{\ell=1}^m e^{-\Delta t_\ell} \leq 1$*. Let* $\ell_*(x) : [0,1] \to \{1, \ldots, m\}$ *be any function such that* $\ell_*(x) \geq \inf\{\ell : \alpha_\ell \geq x\}$*. Then for all* $t \geq 0$*, we have*

$$\Pr\left[\mu_n \geq \mathrm{KL}_2^{-1}(\alpha_{\ell_*(s_n)}||n^{-1}\Delta t_{\ell_*(s_n)} + t)\right] \leq e^{-nt}.$$

Proof. Let $t_\ell = t + \Delta t_\ell/n$. We obtain from Theorem 2 that for each ℓ:

$$\Pr\left[\mu_n \geq \mathrm{KL}_2^{-1}(\alpha_\ell||t_\ell), s_n \leq \alpha_\ell\right] \leq e^{-\Delta t_\ell}e^{-nt}.$$

Take a union bound over $\ell = 1, \ldots$, we have:

$$\begin{aligned}
&\Pr\left[\ell = \ell_*(s_n) : \mu_n \geq \mathrm{KL}_2^{-1}(\alpha_\ell||t_\ell)\right] \\
\leq &\Pr\left[\exists \ell \in \{1, \ldots, m\} : \mu_n \geq \mathrm{KL}_2^{-1}(\alpha_\ell||t_\ell), s_n \leq \alpha_\ell\right] \\
\leq &\sum_{\ell=1}^m e^{-\Delta t_\ell}e^{-nt} \leq e^{-nt}.
\end{aligned}$$

This proves the lemma.

Theorem 3. *Under the conditions of Lemma 3. For all* $t \geq 0$*, we have*

$$\Pr\left[\mu_n \geq \mathrm{KL}_2^{-1}(n^{-1}\lceil ns_n\rceil||2n^{-1}\ln(\lceil ns_n\rceil + 2) + t)\right] < e^{-nt}.$$

Proof. In Lemma 4, we take $\alpha_\ell = (\ell-1)/n$ for $\ell = 1, \ldots, n+1$, $\Delta t_\ell = 2\ln(\ell+1)$, and $\ell_*(x) = \lceil nx\rceil + 1$.

2.2 Conditional Bennett inequalities

In Bernstein and Bennett inequalities, the resulting bounds depend on the variance of the random variables (for example, see [2]).

These inequalities are useful for some statistical estimation problems including least squares regression and density estimation with log-loss. This is because for these problems, the variance of a random variable can be bounded by its mean: $\exists b > 0 : \mathbf{E}_{Z_k}(\xi_k - \mathbf{E}_{Z_k}\xi_k)^2 \leq b\mathbf{E}_{Z_k}\xi_k$. Probability inequalities that use variance become crucial to obtain good bounds.

Bernstein inequalities for dependent random variables have been investigated in the literature (for example, see [6, 7] and references therein). However, they were not in the form most suitable for our purpose. We shall thus derive some new bounds here that are directly applicable to statistical estimation problems. Our bounds depend on the following additional quantity:

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}_{Z_k}(\xi_k - \mathbf{E}_{Z_k}\xi_k)^2.$$

A standard estimate of moment generating function leads to the following result.

Lemma 5. *Assume that $\xi_k - \mathbf{E}_{Z_k}\xi_k \geq -1$ for each k. We have $\forall \rho > 0$:*

$$\Pr\left[\mu_n \geq \frac{e^\rho - \rho - 1}{\rho}\sigma_n^2 + s_n + \frac{t}{\rho}\right] \leq e^{-nt}.$$

Proof. Let $\tilde{\xi}_k = \xi_k - \mathbf{E}_{Z_k}\xi_k$. We start with the following estimate

$$\begin{aligned}\ln \mathbf{E}_{Z_k} e^{-\rho\tilde{\xi}_k} \leq & \mathbf{E}_{Z_k} e^{-\rho\tilde{\xi}_k} - 1 \\ = & \mathbf{E}_{Z_k}\tilde{\xi}_k^2 \frac{e^{-\rho\tilde{\xi}_k} + \rho\tilde{\xi}_k - 1}{\tilde{\xi}_k^2} \\ \leq & \mathbf{E}_{Z_k}\tilde{\xi}_k^2 \frac{e^\rho - \rho - 1}{1^2}.\end{aligned}$$

The first inequality uses $\ln x \leq x - 1$. The last inequality uses the fact that $f(x) = x^{-2}(e^x - x - 1)$ is a non-decreasing function of x, and $-\rho\tilde{\xi}_k \leq \rho$. By using this estimate in Lemma 2, we obtain the desired bound.

The condition $\xi_k - \mathbf{E}_{Z_k}\xi_k \geq -1$ was considered by Bennett. It can be changed to appropriate moment conditions (in Bernstein inequalities). The proof of Lemma 5 follows a standard argument for proving Bennett bounds. This same method was also used in [7] to obtain a related bound (also see [6]). However, for statistical estimation problems, results in [7] are not directly applicable. This is because the most common application of this theorem is under the assumption that there exists a constant $b > 0$ such that $b\mathbf{E}_{Z_k}\xi_k \geq \mathbf{E}_{Z_k}(\xi_k - \mathbf{E}_{Z_k}\xi_k)^2$. Under this assumption, a more suitable bound can be derived from Lemma 5 as follows.

Theorem 4. *Assume that $\xi_k - \mathbf{E}_{Z_k}\xi_k \geq -1$ for each k. If there exists $b > 0$ such that $b\mathbf{E}_{Z_k}\xi_k \geq \mathbf{E}_{Z_k}(\xi_k - \mathbf{E}_{Z_k}\xi_k)^2$ for all k, then $\forall\alpha \geq 0$ and $t \geq 0$:*

$$\Pr\left[\mu_n \geq s_n + \sqrt{2\alpha b t} + c_b t, \quad s_n \leq \alpha\right] \leq e^{-nt},$$

where $c_b = \sqrt{(b+2/3)b} + b + 1/3$.

Proof. It is easy to verify that for $\rho \in (0,3)$:

$$e^\rho - \rho - 1 \leq \frac{\rho^2}{2}\sum_{u=0}^{\infty}(\rho/3)^u = \frac{\rho^2}{2(1-\rho/3)}.$$

Using Lemma 5, for any fixed $\rho \in (0, 6/(3b+2))$, we have with probability at least $1 - e^{-nt}$, $s_n \leq \alpha$ and

$$\mu_n < \left(1 - \frac{\rho b}{2(1-\rho/3)}\right)^{-1}(s_n + t/\rho) \leq s_n - \alpha + \left(1 - \frac{\rho b}{2(1-\rho/3)}\right)^{-1}(\alpha + t/\rho).$$

Since this inequality is true for all ρ, we may optimize over ρ. In particular, let $\rho^{-1} = (0.5b + 1/3) + \sqrt{0.5b(\alpha + 0.5bt + t/3)/t}$. and simplify, we obtain the theorem.

Similar to Lemma 4, we may obtain a data-dependent version of Theorem 4 with the same proof.

Lemma 6. *Under the conditions of Theorem 4. Consider a sequence $\alpha_1 \leq \alpha_2 \cdots$ and a sequence $\{\Delta t_\ell\}$ such that $\sum_\ell e^{-\Delta t_\ell} \leq 1$. Let $\ell_*(x)$ be a integer valued function such that $\ell_*(x) \geq \inf\{\ell : \alpha_\ell \geq x\}$. Then for all $t \geq 0$, we have*

$$\Pr\left[\mu_n \geq s_n + \sqrt{2b\alpha_{\ell_*(s_n)} t(s_n)} + c_b t(s_n)\right] \leq e^{-nt},$$

where $t(s_n) = t + \Delta t_{\ell_*(s_n)}/n$ and $c_b = \sqrt{(b+2/3)b} + b + 1/3$.

Theorem 5. *Under the conditions of Theorem 4. For all $t \geq 0$, we have*

$$\Pr\left[\mu_n \geq s_n + \sqrt{2bn^{-1}\max(0, \lceil ns_n \rceil) t(s_n)} + c_b t(s_n)\right] \leq e^{-nt},$$

where $t(s_n) = t + 2n^{-1}\ln(\lceil ns_n \rceil + 2)$ and $c_b = \sqrt{(b+2/3)b} + b + 1/3$.

Proof. In Lemma 6, we take $\alpha_\ell = (\ell - 1)/n$ for $\ell = 1, 2, \cdots$, $\Delta t_\ell = 2\ln(\ell + 1)$, and $\ell_*(x) = \lceil nx \rceil + 1$.

Note that if $\xi \in [0,1]$, then the condition of Theorem 5 is satisfied with $b = 1$. However, Theorem 3 is tighter in this case (using Proposition 1).

3 Generalization Bounds for Some Online Algorithms

We consider two scenarios. One is classification, which requires the Hoeffding inequality developed in Section 2.1. The other is regression, which utilizes the Bennett inequality in Section 2.2.

3.1 Classification

We consider multi-category classification problem, with zero-one classification loss. We are interested in a classification function $h : \mathcal{X} \to \mathcal{Y} = \{1, \ldots, K\}$, with the classification loss

$$\mathrm{err}(h(x), y) = \begin{cases} 0 & \text{if } h(x) = y \\ 1 & \text{otherwise.} \end{cases}$$

The risk (expected classification error) of h is

$$\mathrm{err}(h) = \mathbf{E}_{(X,Y)\sim D}\mathrm{err}(h(X), Y).$$

Consider training data $Z_1^n = (Z_1, \ldots, Z_n)$ that are independently drawn from D. Consider a learning algorithm $\mathcal{A}$ that learns from the first k samples Z_1^k a classifier $\hat{h}_k(x) = \mathcal{A}(Z_1^k; x) : \mathcal{X} \to \mathcal{Y}$. We restate from Theorem 3 the following generalization bound.

Theorem 6. *Let $\hat{h}_k$ be a classifier learned from an algorithm after seeing training data $Z_1, \ldots, Z_k$. Let $\hat{M}_n = \sum_{i=1}^n \mathrm{err}(\hat{h}_{i-1}(X_i), Y_i)$ be the number of mistakes the algorithm makes online after n examples. Then with probability at most e^{-nt},*

$$\frac{1}{n}\sum_{i=1}^n \mathrm{err}(\hat{h}_i) \geq \mathrm{KL}_2^{-1}(n^{-1}\hat{M}_n || 2n^{-1}\ln(\hat{M}_n + 2) + t).$$

We may also apply the analysis to specific algorithms with known mistake bounds. For example, we may consider the multi-category perceptron algorithm [5], and obtain a margin bound accordingly. The multi-category perceptron algorithm is a natural generalization of the binary perceptron algorithm, popular in natural language processing due to its simplicity and effectiveness.

In the setting of multi-category perceptron [5], a data-point $x \in \mathcal{X}$ is represented by K vectors $\{x[1], \ldots, x[K]\}$, each corresponding to a class-label. A classifier h is represented by a linear weight vector w, with the corresponding classification rule:

$$h(x) = \arg\max_{\ell=1,\ldots,K} w^T x[\ell].$$

Let y be the true label of x. One may define the corresponding margin for the data point $z = (x, y)$ as

$$\gamma(w, z) = \frac{1}{\|w\|_2^2}\left(w^T x[y] - \max_{\ell \neq y} w^T x[\ell]\right).$$

The multi-category perceptron method maintains a weight vector starting from $w_0 = 0$. After seeing a data-point (X_i, Y_i), the algorithm uses the current weight w_{i-1} to make a prediction, which produces a label Y. We then update the weight vector as $w_i = w_{i-1} + (X_i[Y_i] - X_i[Y])$.

Assume there is a linear separator w_* for the training data $Z_1, \ldots, Z_n$ such that the margin $\inf_i \gamma(w_*, Z_i) \geq \gamma > 0$, then the standard perceptron bound can be extended to show (see [5]) that the number of mistakes that the perceptron method makes is no more than $(R/\gamma)^2$, where $R \geq \sup_{i,Y} \|X_i[Y_i] - X_i[Y]\|_2$.

Theorem 7. *Consider a linear separator $w_*(Z_1^n)$ for the training data $Z_1, \ldots, Z_n$ such that $\inf_i \gamma(w_*, Z_i) \geq \gamma(Z_1^n) > 0$. For all $R(Z_1^n) \geq \sup_{i,Y} \|X_i[Y_i] - X_i[Y]\|_2$. We have with probability of at most e^{-nt}:*

$$M = \frac{R(Z_1^n)^2}{\gamma(Z_1^n)^2} \leq n, \quad \frac{1}{n}\sum_{i=1}^n \mathrm{err}(\hat{h}_i) \geq \mathrm{KL}_2^{-1}(n^{-1}M||2n^{-1}\ln(M+2)+t).$$

Again, Proposition 1 can be used to obtain a more intuitive bound. If the mistaken bound $R^2/\gamma^2 = O(1)$, then the generalization performance in Theorem 7 is $O(1/n)$ at constant probability $t = O(1/n)$. This can be compared to well-known batch margin bounds in the literature, which (to the author's knowledge) do not achieve the $O(1/n)$ rate under the same assumptions.

Assume further that there is a linear separator w_* that does not only separate the training data, but also all the test data. Let R and γ be defined with respect to all data, then by the data independent bound in Theorem 2, we have with probability no more than e^{-nt}:

$$\frac{1}{n}\sum_{i=1}^n \mathrm{err}(\hat{h}_i) \geq \mathrm{KL}_2^{-1}(n^{-1}R^2/\gamma^2||t).$$

3.2 Regression

It is known that some learning problems have loss functions that satisfy the following self-bounding condition: $b\mathbf{E}L(h(X), Y) \geq \mathbf{E}L(h(X), Y)^2$ for some $b > 0$. For such problems, the Bennett inequality in Section 2.2 should be applied.

To illustrate the idea, in the following, we shall consider the least squares regression problem:

$$L(h(X), Y) = (h(X) - Y)^2,$$

where $Y \in [0, 1]$. Let S be a closed convex set of functions such that $h(X) \in S$ implies that $h(X) \in [0, 1]$. Let h_S be the optimal predictor in S:

$$\mathbf{E}_{X,Y}(h_S(X) - Y)^2 = \inf_{h \in S} \mathbf{E}_{X,Y}(h(X) - Y)^2.$$

We have the following inequality

Lemma 7. *Let $Z = (X, Y)$, and $\Delta L_S(h, Z) = (h(X) - Y)^2 - (h_S(X) - Y)^2$. For all $h \in S$:*

$$4\mathbf{E}_Z \Delta L_S(h, Z) \geq \mathbf{E}_Z \Delta L_S(h, Z)^2.$$

Proof. The convexity of S and optimality of h_S implies that $\forall h \in S$, the derivative of $\mathbf{E}_Z(h_S(X) + t(h(X) - h_S(X)) - Y)^2$ as a function of t is non-negative at $t = 0$. That is, $\mathbf{E}_Z(h_S(X) - Y)(h(X) - h_S(X)) \geq 0$. Therefore

$$\begin{aligned}\mathbf{E}_Z\ \Delta L_S(h, Z)^2 =& \mathbf{E}_Z\ (h(X) - h_S(X))^2(h(X) + h_S(X) - 2Y)^2 \\ \leq& 4\mathbf{E}_Z\ (h(X) - h_S(X))^2 \\ \leq& 4\mathbf{E}_Z\ [(h(X) - h_S(X))^2 + 2(h(X) - h_S(X))(h_S(X) - Y)] \\ =& 4\mathbf{E}_Z \Delta L_S(h, Z).\end{aligned}$$

Let $\mathcal{A}$ be a learning algorithm, and S be a set of convex functions. Assume that all hypothesis learned by $\mathcal{A}$ belong to S, we can obtain the following theorem from Theorem 5.

Theorem 8. *Let $\hat{h}_k \in S$ be a function learned from an algorithm after seeing training data $Z_1, \ldots, Z_k$. Let*

$$\hat{M}_n = \max\left(0, \sum_{i=1}^{n} (\hat{h}_{i-1}(X_i) - Y_i)^2 - \inf_{h \in S} \sum_{i=1}^{n} (h(X_i) - Y_i)^2\right).$$

Then with probability at most e^{-nt},

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{E}_Z\ (\hat{h}_i(X) - Y)^2 \geq \mathbf{E}_Z\ (\hat{h}_S(X) - Y)^2 + \frac{\hat{M}_n}{n} + \sqrt{8n^{-1}\lceil \hat{M}_n \rceil \hat{t}} + 9\hat{t},$$

where $\hat{t} = t + 2n^{-1} \ln\lceil \hat{M}_n + 2\rceil$.

Proof. We apply Theorem 5 with $\xi_k = \Delta L_S(\hat{h}_k, Z)$ and $b = 4$. We only need to note that $ns_n \leq \hat{M}_n$ and $c_b \leq 9$.

Bounds that relate the performance of an learning algorithm to the best possible performance (within a function class) is often referred to as oracle inequality in the learning theory literature. Theorem 8 can be viewed as a data-dependent oracle inequality. If $\hat{M}_n$ is small, then the total generalization performance (compared with h_S) can be faster than $O(1/\sqrt{n})$. In particular, if $\hat{M}_n = O(\ln n)$, then the performance can be as fast as $O(\ln n/n)$ with constant probability $t = O(1/n)$. The theorem can be applied to any online algorithm for least squares regression based on the observed loss. For some algorithms, it is also possible to prove data-dependent or data-independent mistake bounds for $\hat{M}_n$. Similar to Theorem 7, we may derive more specific performance bounds for these specific algorithms, assuming $\hat{M}_n$ can be bounded using some quantities that depend on the data. We shall skip the details here.

4 Expert Aggregation Algorithms

We consider learning algorithms $\mathcal{A}_\theta$ parameterized by $\theta \in \Gamma$. For notation simplicity, given a sample $Z = (X, Y)$, we let

$$L_\theta(Z_1^k; Z) = L(\mathcal{A}_\theta(Z_1^k; X), Y).$$

That is, $L_\theta(Z_1^k; \cdot)$ is the loss of the function learned by $\mathcal{A}_\theta$ using the first k training data.

In the expert framework, as in [12], we consider a prior distribution on Γ, which we denote as $d\pi_0(\theta)$. We maintain and update a distribution on Γ, and the update rule depends on a parameter η. With each sample Z_k, the distribution imposed on the experts are updated as follows:

$$d\pi_k^\eta(\theta) \sim e^{-\eta L_\theta(Z_1^{k-1}; Z_k)}\, d\pi_{k-1}^\eta(\theta),$$

where $\eta > 0$ is a fixed learning rate. We also let $\pi_0^\eta(\theta) = \pi_0(\theta)$. It follows that the distribution after seeing the first k samples Z_1^k, is the Gibbs distribution

$$d\pi_k^\eta(\theta) \sim e^{-\eta \sum_{i=1}^k L_\theta(Z_1^{i-1};Z_i)}\, d\pi_0(\theta). \tag{5}$$

For specific expert algorithms devised in the literature, our earlier analysis can be applied to obtain generalization bounds. In the following, we show that it is also natural to study the concentration behavior of expert aggregating algorithms directly, using tools we developed in Section 2.

Lemma 8. $\forall \eta \geq 0$*, the following inequality holds:*

$$\Pr\left[-\sum_{i=1}^n \ln \int d\pi_i^\eta(\theta)\mathbf{E}_{Z_i} e^{-\eta L_\theta(Z_1^{i-1};Z_i)} \geq M_n(\eta)\right] \leq e^{-t},$$

where

$$M_n(\eta) = -\ln \int e^{-\eta \sum_{i=1}^n L_\theta(Z_1^{i-1};Z_i)} d\pi_0(\theta).$$

Proof. If we let $\xi_k = -\ln \int d\pi_i^\eta(\theta) e^{-\eta L_\theta(Z_1^{i-1};Z_i)}$, then it is easy to verify that $\sum_{i=1}^n \xi_i = M_n$. We can now apply Lemma 2 with $\rho = 1$, and use the fact that

$$\mathbf{E}_{Z_i} e^{-\xi_i} = \int d\pi_i^\eta(\theta)\mathbf{E}_{Z_i} e^{-\eta L_\theta(Z_1^{i-1};Z_i)}.$$

The desired bound is now a direct consequence of Lemma 2.

Note that η in Lemma 8 has a similar effect of ρ in Lemma 2. If we only have one expert, then we can obtain Lemma 2 from Lemma 8. Similar to the development in Section 2, we may obtain more specific bounds from Lemma 8 using appropriate estimates of logarithmic moment generating functions.

For simplicity, we will only consider Hoeffding inequality for bounded random variables. The aggregating algorithm which we shall investigate is the performance averaged over θ with respect to the distribution π_i^η, and time steps $i = 1, \ldots, n$. This method was referred to as *Hedge* in [8], and is closely related to boosting. Its generalization performance is represented as $\mu_n(\eta)$ in the following lemma, which plays the same role of Lemma 3.

Lemma 9. *If the loss function* $L(\cdot,\cdot) \in [0,1]$*, then*

$$\Pr\left[-\ln\left(1 + \mu_n(\eta)(e^{-\eta} - 1)\right) \geq \frac{1}{n} \ln \int e^{-\eta n s_n(\theta)} d\mu_0(\theta) + t\right] \leq e^{-nt},$$

where

$$\mu_n(\eta) = \frac{1}{n}\sum_{i=1}^n \int d\pi_i^\eta(\theta)\mathbf{E}_{Z_i} L_\theta(Z_1^{i-1};Z_i)$$

and

$$s_n(\theta) = \frac{1}{n}\sum_{i=1}^n L_\theta(Z_1^{i-1}, Z_i).$$

Now, with a fixed learning rate η, similar to the first inequality of Theorem 1, we obtain from Lemma 9 the following generalization bound:

$$\Pr\left[\mu_n(\eta) \geq \frac{-\frac{1}{n}\ln\int e^{-\eta n s_n(\theta)} d\mu_0(\theta) + t}{1-e^{-\eta}}.\right] \leq e^{-nt}.$$

As a simple example, if we only have a finite number of experts: $|\Gamma| < \infty$, then we may take μ_0 to be the uniform distribution. This gives

$$s_n(\eta) \leq \eta \inf_{\theta\in\Gamma} s_n(\theta_\ell) + \frac{1}{n}\ln|\Gamma|.$$

We have

$$\Pr\left[\mu_n(\eta) \geq \frac{\eta \inf_{\theta\in\Gamma} s_n(\theta_\ell) + \frac{1}{n}\ln|\Gamma| + t}{1-e^{-\eta}}\right] \leq e^{-nt}.$$

This generalization bound holds for fixed η, which may not be optimal for the observed data.

An important question is to select η based on the training data Z_1^n so as to achieve a small generalization error $\mu_n(\eta)$. This is essentially a model selection problem, which requires us to develop a data dependent bound (η depends on the training data). In order to do so, we shall use the following result to obtain a simpler representation of $M_n(\eta)$. It is a direct consequence of a convex duality, widely used in the machine learning literature in recent years. For space limitation, we skip the proof.

Proposition 2. *Consider all possible distributions π over Γ, we have*

$$-\frac{1}{n}\ln\int e^{-\eta n s_n(\theta,\eta)} d\pi_0(\theta) = \inf_{\pi}\left[\eta\int s_n(\theta)d\pi + \frac{1}{n}\mathrm{KL}(\pi||\pi_0)\right],$$

where $\mathrm{KL}(\pi||\pi_0) = \int \ln\frac{d\pi(\theta)}{d\pi_0(\theta)} d\pi(\theta)$.

We are now ready to present the following bound, which is similar to Theorem 2. A related bound can be found in [11].

Theorem 9. *Under the conditions of Lemma 9. We have $\forall\alpha \in [0,1]$ and $t, \delta \in [0,\infty)$:*

$$\Pr\left[\mu_n(\eta(\alpha, t+\delta)) \geq \mathrm{KL}_2^{-1}(\alpha||t+\delta),\, \exists\pi : \int s_n(\theta)d\pi \leq \alpha, \mathrm{KL}(\pi||\pi_0) \leq n\delta\right] \leq e^{-nt},$$

where $\eta(\alpha,u) = \ln(\mathrm{KL}_2^{-1}(\alpha||u)(1-\alpha)) - \ln(\alpha(1-\mathrm{KL}_2^{-1}(\alpha||u)))$.

Proof. We have with probability of at most e^{-nt}, $\exists\pi$:

$$-\ln\left(1+\mu_n(\eta)(e^{-\eta}-1)\right) \geq \eta\int s_n(\theta)d\pi + \frac{1}{n}\mathrm{KL}(\pi||\pi_0) + t.$$

Now, let $\beta = \mathrm{KL}_2^{-1}(\alpha||t+\delta)$. This implies that with probability at most e^{-nt}, $\exists\pi$ such that:

$$\mu_n(\eta) \geq \beta, \int s_n(\theta)d\pi \leq \alpha, \frac{1}{n}\mathrm{KL}(\pi||\pi_0) \leq \delta,$$
$$-\ln(1+\beta(e^{-\eta}-1)) \geq \eta\alpha + \mathrm{KL}(\alpha||\beta).$$

Note that the last inequality holds trivially with $\eta = \eta(\alpha, t+\delta)$. This leads to the theorem.

Using the standard union bound trick, we can obtain a version of Theorem 9 with data-dependent α and δ. The proof of the following result is a straightforward extension of that of Lemma 4. Again, due to the space limitation, we skip the proof.

Lemma 10. *Using notations of Theorem 9. Consider a set of triples $\{(\alpha_\ell, \delta_\ell, \Delta t_\ell)\}$ indexed by ℓ, where $\alpha_\ell \in [0,1]$, $\delta_\ell \geq 0$, and $\sum_\ell e^{-\Delta t_\ell} \leq 1$. Let $\ell_*(\alpha,\delta)$ be a function such that $\alpha_\ell \geq \alpha$ and $\delta_\ell \geq \delta$. Then we have $\forall t \geq 0$:*

$$\Pr\left[\exists\pi : \ell = \ell_n(\pi), \quad \mu_n(\eta(\alpha_\ell, t_\ell)) \geq \mathrm{KL}_2^{-1}(\alpha_\ell||t_\ell)\right] \leq e^{-nt},$$

where $\ell_n(\pi) = \ell_*(\int s_n(\theta)d\pi, \mathrm{KL}(\pi||\pi_0)/n)$ and $t_\ell = t + \delta_\ell + \Delta t_\ell/n$.

With specific choices of $(\alpha_\ell, \delta_\ell, \Delta t_\ell)$, we can obtain the following result.

Theorem 10. *Using notations of Theorem 9. We have $\forall t \geq 0$:*

$$\Pr\left[\exists\pi : \mu_n(\eta_n^\pi) \geq \mathrm{KL}_2^{-1}(n^{-1}\lceil ns_n^\pi\rceil||t+t_n^\pi)\right] \leq e^{-nt},$$

where

$$s_n^\pi = \int s_n(\theta)d\pi,$$
$$t_n^\pi = n^{-1}[\lceil\mathrm{KL}(\pi||\pi_0)\rceil + 2\ln(\lceil ns_n^\pi\rceil + 2) + 2\ln(\lceil\mathrm{KL}(\pi||\pi_0)\rceil + 2)],$$
$$\eta_n^\pi = \eta(n^{-1}\lceil ns_n^\pi\rceil, t+t_n^\pi).$$

Proof. In Lemma 10, we let ℓ be represented by a pair of positive integers (p,q) such that $\alpha_{(p,q)} = (p-1)/n, \delta_{(p,q)} = (q-1)/n, \Delta_{(p,q)} = 2\ln(p+1) + 2\ln(q+1)$. It can be easily checked that $\sum e^{-\Delta_{(p,q)}} \leq 1$. Now we can simply let $\ell_*(u,v) = (\lceil nu\rceil, \lceil nv\rceil)$ to obtain the desired result.

With constant probability $t = O(1/n)$, a convergence rate of $O(1/n)$ can be achieved when there exists a π such that $\int s_n^\pi(\theta)d\pi = O(1)$ and $\mathrm{KL}(\pi||\pi_0) = O(1)$. The main part of t_n^π is $n^{-1}\mathrm{KL}(\pi||\pi_0)$. By focusing on the main part, we approximately have the following bound from Theorem 10 and Proposition 1. With probability of at least $1 - e^{-nt}$, $\forall\pi$, we can appropriately chose learning rate η that depends on the data such that

$$\mu_n(\eta) < s_n^\pi + \sqrt{2s_n^\pi(1-s_n^\pi)(t+n^{-1}\mathrm{KL}(\pi||\pi_0))} + 1.5(1-s_n^\pi)(t+n^{-1}\mathrm{KL}(\pi||\pi_0)).$$

As an application, assume there are only a finite number of experts. We may just pick π_0 to be the uniform distribution and π to be concentrated on the expert with the smallest number of empirical loss such that $\mathrm{KL}(\pi||\pi_0) = \ln |\Gamma|$ and $s_n^\pi = \inf_{\theta \in \Gamma} s_n(\theta)$.

5 Conclusion

In this paper, we considered the problem of estimating the total generalization performance of a learning algorithm based on its observed total loss. This is achieved through some newly obtained probability inequalities concerning the concentration of dependent random variables. Consequences of our analysis in classification and regression were discussed. If the observed loss is small, then the estimated generalization performance can be as fast as $O(1/n)$ with constant probability. Moreover, we showed that the technique used to prove probability inequalities for dependent variables can be naturally applied to analyze the generalization behavior of expert aggregating algorithms. In this case, by minimizing the resulting data-dependent bound, we obtain a method of choosing the learning rate η with optimal total generalization performance (according to the bound).

References

1. K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Math. Journal*, 3:357–367, 1967.
2. George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57:33–45, 1962.
3. Avrim Blum, Adam Kalai, and John Langford. Beating the hold-out: Bounds for k-fold and progressive cross-validation. In *COLT' 99*, pages 203–208, 1999.
4. N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, pages 2050–2057, 2004.
5. Michael Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proc. EMNLP'02*, 2002.
6. Victor H. de la Pẽna. A general class of exponential inequalities for martingales and ratios. *The Annals of Probability*, 27:537–564, 1999.
7. David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3:100–118, 1975.
8. Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997.
9. W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.
10. Nick Littlestone. From on-line to batch learning. In *COLT' 89*, pages 269–284, 1989.
11. Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.
12. Volodya Vovk. Aggregating strategies. In *COLT' 90*, pages 371–383, 1990.