Journal of Machine Learning Research 10 (2009) 977-996 Submitted 2/08; Revised 9/08; Published 4/09

On Uniform Deviations of General Empirical Risks with Unboundedness, Dependence, and High Dimensionality

Wenxin Jiang WJIANG@NORTHWESTERN.EDU
Department of Statistics
Northwestern University
Evanston, IL 60208, USA

Editor: Gábor Lugosi

Abstract

The statistical learning theory of risk minimization depends heavily on probability bounds for uniform deviations of the empirical risks. Classical probability bounds using Hoeffding's inequality cannot accommodate more general situations with unbounded loss and dependent data. The current paper introduces an inequality that extends Hoeffding's inequality to handle these more general situations. We will apply this inequality to provide probability bounds for uniform deviations in a very general framework, which can involve discrete decision rules, unbounded loss, and a dependence structure that can be more general than either martingale or strong mixing. We will consider two examples with high dimensional predictors: autoregression (AR) with ℓ_1-loss, and ARX model with variable selection for sign classification, which uses both lagged responses and exogenous predictors.

Keywords: dependence, empirical risk, probability bound, unbounded loss, uniform deviation

1. Introduction

In machine learning, a problem of central importance is to bound a probability of uniform deviation $P[\sup_{b\in B}|n^{-1}\sum_{t=1}^n \rho(\omega_t,b)-n^{-1}\sum_{t=1}^n E\rho(\omega_t,b)|>\delta]$, where $\delta>0$ is a positive deviation (which can be allowed to depend on n and characterize a convergence rate), b is a parameter in a parameter space B (which is typically a Borel measurable subset of an Euclidean space), $D=(\omega_1,...,\omega_n)$ form the data set of n random observations, $\rho(\cdot,\cdot)$ is a loss function (measurable to a certain product σ-field), $\hat{R}(b)=n^{-1}\sum_{t=1}^n \rho(\omega_t,b)$ is an empirical risk, and $R(b)=n^{-1}\sum_{t=1}^n E\rho(\omega_t,b)$ is its expectation.[1]

Such a probability is of interest since it is well known to bound the performance $R(\hat{b})$ of an empirical risk minimizer $\hat{b}=\arg\min_{b\in B}\hat{R}(b)$, relative to the optimal performance $\inf_{b\in B}R(b)$ over B:

$$P[R(\hat{b})-\inf_{b\in B}R(b)>2\delta]\le P[\sup_{b\in B}|\hat{R}(b)-R(b)|>\delta],$$

due to, for example, Lemma 8.2 (Devroye, Györfi and Lugosi, 1996). Recently, Jiang and Tanner (2007, 2008) indicate that the probability of uniform deviation is also of central importance in

[1] In this paper, we will not be concerned about the measurability problem that may be involved in quantities such as $\sup_{b\in B}|\hat{R}(b)-R(b)|$. Works on 'universal measurability' described in, for example, Yu (1994, Appendix) and Davidson (1994, Section 21.1) imply that this is not a problem for all our examples later where B is a Borel measurable subset of a compact metric space. Alternatively we could consider P as the 'outer probability' as in Newey (1991).

studying the performance of a Bayesian approach of empirical risk minimization considered in, for example, Zhang (2006), when b is generated randomly according to a Gibbs posterior $\pi_{b|D}(db) \propto e^{-n\psi\hat{R}(b)}\pi_b(db)$ where $\pi_b(db)$ is a prior distribution on B and $\psi^{-1} > 0$ is a 'temperature parameter'. This includes the usual Bayesian posterior as a special case when $\psi = 1$ and when $-n\hat{R}$ is the log-likelihood function. A straightforward application of Jiang and Tanner (2008, Proposition 6) renders

$$P[R(b) - \inf_{b\in B} R(b) > 5\delta] \leq P[\sup_{b\in B} |\hat{R}(b) - R(b)| > \delta] + e^{-2n\psi\delta}/\pi_b[R(b) - \inf_{b\in B} R(b) < \delta],$$

when $b|D \sim \pi_{b|D}$ and D is generated from a true distribution. This again shows the dependence of the risk performance on the probability of uniform deviation.

The probability of uniform deviation is treated in the standard machine learning text such as Devroye, Györfi and Lugosi (1996) by the Vapnik-Chervonenkis theory using a Hoeffding's inequality on the probability of pointwise deviation $P[|n^{-1}\sum_{t=1}^n \rho(\omega_t, b) - n^{-1}\sum_{t=1}^n E\rho(\omega_t, b)| > \delta]$, which typically assumes that ω_t's are iid (independent and identically distributed), and that the loss function ρ is bounded. The goal of this paper is to generalize in several directions, so that ρ can be unbounded and ω_t's can be dependent. In addition, we will allow b to have a possibly high dimension that can increase with n in certain ways. When ρ has sufficiently thin tail in the distribution and when ω_t's have certain kind of decaying dependence over t, we derive bounds of the form

$$P[\sup_{b\in B} |n^{-1}\sum_{t=1}^{n} \rho(\omega_t, b) - n^{-1}\sum_{t=1}^{n} E\rho(\omega_t, b)| > n^{-0.5+\gamma_1}] = O(e^{-c_1 n^{c_2}}),$$

for any small positive γ_1, where c_1, c_2 are some positive constants depending on γ_1. Such a result indicates uniform convergence of the empirical risk at a near 'parametric' rate (close to $O_P(n^{-0.5})$) despite high dimensionality in b, dependence in ω_t, and unbounded loss function ρ.

Such results are obtained using a very general 'pointwise' inequality that generalizes Hoeffding's inequality, which will be introduced in Section 2. This allows unbounded loss and a framework of dependence that is more general than strong mixing, which is therefore more general than previous works using strong mixing (e.g., Vidyasagar, 2005; Zou and Li, 2007) or β-mixing (e.g., Yu, 1994; Lozano, Kulkarni and Schapire, 2006). The 'uniform aspect' is then treated in a very general framework in Section 3 allowing both continuity and discontinuity of ρ in b. Examples of applications of this general framework are given in Sections 4 and 5.

2. An Inequality

We first introduce an inequality that is more general than Hoeffding's inequality (Hoeffding, 1963). This inequality may be called a 'triplex inequality' since its right hand side has three parts. In addition to a term that is of an exponential form as in the Hoeffding's inequality, it also includes a term to gauge the dependence, and a term to control the unboundedness of the random variables. The result is therefore almost assumption free and generally applicable: it does not assume independence or boundedness of the random variables.

Theorem 1 *(A triplex inequality.) let $\{\mathcal{F}_t\}_{-\infty}^{\infty}$ be an increasing sequence of σ-fields and ρ_t be a random variable that is $\mathcal{F}_t$-measurable for each t. Then for any $\varepsilon, C > 0$ and positive integers n, m,* *we have*

$$
\begin{aligned}
&P[|\sum_{t=1}^{n}(\rho_t - E\rho_t)| > n\varepsilon] \leq 2me^{-n\varepsilon^2/(288m^2C^2)}\\
&+(6/\varepsilon)n^{-1}\sum_{t=1}^{n}E|E(\rho_t|\mathcal{F}_{t-m}) - E\rho_t|\\
&+(15/\varepsilon)n^{-1}\sum_{t=1}^{n}E|\rho_t|I(|\rho_t| > C),
\end{aligned}
$$

as long as the right hand side exists and does not exceed one.

2.1 Remarks

1. The bound is not necessarily very tight; the constants appearing in the theorem may be improved. However, these typically do not affect the convergence rates in the later applications of this inequality.

2. In later applications, the choices of m and C can be made to depend on n so that the combination of all three terms converge to zero as $n \to \infty$.

3. The last term will be called the 'tail term' since it is related to the tail behavior of ρ_t. This often can be bounded by techniques similar to the Markov inequalities. Note that for nonnegative $X = |\rho_t|$, $EXI(X > C) \leq EX^{k+1}C^{-k}$ and $EXI(X > C) \leq \sqrt{EX^2}\sqrt{P(X > C)} \leq \sqrt{EX^2}\sqrt{Ee^{\theta X}}e^{-\theta C/2}$, for $k, \theta > 0$. So existence of moments of X will imply a power law and existence of the moment generating function in a neighborhood of zero will imply an exponential law for the decay of the 'tail term' in C.

4. The second term will be called the 'dependence term' since it is related to the dependence described in the framework of L_1-mixingale (see, e.g., Chapter 16, Davidson, 1994), which is more general than either martingale or strong mixing. When $\{\rho_t\}$ is a sequence of martingale differences, the dependence term vanishes. If $\{\rho_t\}$ is strong mixing with coefficients α_m, and has bounded L_q norms ($q > 1$), then Theorem 14.2 of Davidson (1994) would imply that the dependence term decreases according to order $O(\alpha_m^{1-1/q})$ as m increases.

5. The mixingale formulation of the dependence term can also handle a process ρ_t that is *not* strong mixing. We will provide an example below when ρ_t is not strong mixing but is *approximable* to a strong mixing process, where we can still make the dependence term small for large m. Such an extension from 'strong mixing' to 'approximable by strong mixing', although seemingly a small improvement, is very significant. The problem of strong mixing is that a function of a mixing sequence (even an independent sequence) that depends on an infinite number of lags is not generally mixing. This is regarded as a 'serious drawback from the viewpoint of applications in time-series modelling' (Davidson, 1994, p.261), and has led to the 'approximability' framework summarized in Davidson (1994, Chapter 17), which is popular in modern time series study but has not been paid much attention to by the machine learning society. Our work can incorporate this approximability concept and provide a 'bridge' introducing this framework to our field.

2.2 An Example for the Dependence Term

Suppose that $\{\rho_t\}_{t=-\infty}^{\infty}$ can be approximated in an L_1-sense by a strong mixing sequence $\{\rho_{t,k}\}_{t=-\infty}^{\infty}$ as k increases (where $\rho_{t,k}$ is measurable -$\mathcal{F}_t$ for each t):

$$n^{-1}\sum_{t=1}^{n} E|\rho_t - \rho_{t,k}| \le c_1 \nu_k,$$

where $c_1 > 0$ and $\nu_k > 0$ are nonstochastic and ν_k decreases to zero as $k \to \infty$. Suppose the qth moment $||\rho_{t,k}||_q \equiv (E|\rho_{t,k}|^q)^{1/q} \le c_2$ for some constants $q > 1$, $c_2 > 0$. Then Theorem 14.2 of Davidson (1994) implies that

$$n^{-1}\sum_{t=1}^{n} E|E(\rho_{t,k}|\mathcal{F}_{t-m}) - E\rho_{t,k}| \le 6c_2\alpha_m(\{\rho_{t,k}\}_{t=-\infty}^{\infty})^{1-1/q}.$$

Then apply the triangular inequality and note that the dependence term is proportional to

$$\begin{aligned} n^{-1}\sum_{t=1}^{n} E|E(\rho_t|\mathcal{F}_{t-m}) - E\rho_t| &\le n^{-1}\sum_{t=1}^{n} E|E(\rho_{t,k}|\mathcal{F}_{t-m}) - E\rho_{t,k}| + 2n^{-1}\sum_{t=1}^{n} E|\rho_t - \rho_{t,k}| \\ &\le 6c_2\alpha_m(\{\rho_{t,k}\}_{t=-\infty}^{\infty})^{1-1/q} + 2c_1\nu_k. \end{aligned}$$

We may be able to choose $k = k(m)$ to increase with m somehow so that both terms above are small for large m. Such a choice $k(m)$ depends on the mechanism of approximation. When k indicates the number of lags involved as in the following example, one can choose $k(m) \approx m/2$.

For example, consider an $MA(\infty)$ process $\rho_t = \sum_{j=0}^{\infty}\theta_j V_{t-j}$ where $\{V_t\}_{-\infty}^{\infty}$ is a zero-mean, L_q-bounded sequence (i.e., $\sup_t ||V_t||_q < \infty$) for some $q > 1$. (We can take $\mathcal{F}_t$ to be the σ-field generated by $\{V_s\}_{s=-\infty}^{t}$.) Then ρ_t is not necessarily strong mixing even when V_t's are independent innovations, even when $|\theta_j|$ decreases very rapidly, due to the infinitely many lags involved (see, e.g., Section 14.3, Davidson 1994). On the other hand, when $|\theta|_1 \equiv \sum_1^{\infty}|\theta_j| < \infty$, we can define 'finite-lag' approximators $\rho_{t,k} = \sum_{j=0}^{k}\theta_j V_{t-j}$ so that $E|\rho_t - \rho_{t,k}| = E|\sum_{k+1}^{\infty}\theta_j V_{t-j}| \le \sup_t ||V_t||_1 \sum_{k+1}^{\infty}|\theta_j|$ which is of the form $c_1\nu_k$ where $\nu_k = \sum_{k+1}^{\infty}|\theta_j| \to 0$ as $k \to \infty$.

Suppose V_t is strong mixing (e.g., when innovations are independent) with mixing coefficient $\alpha_m(\{V_t\}_{-\infty}^{\infty})$. Then the strong mixing coefficient of $\rho_{t,k}$ satisfies

$$\alpha_m(\{\rho_{t,k}\}_{t=-\infty}^{\infty}) \le \alpha_{m-k}(\{V_t\}_{-\infty}^{\infty}),$$

since the $\rho_{t,k}$ depends on lags $V_t, V_{t-1}, ..., V_{t-k}$. Note that $||\rho_{t,k}||_q \le \sum_{j=0}^{k}|\theta_j| \sup_t ||V_{t-k}||_q \le |\theta|_1 \sup_t ||V_t||_q$ which can be taken as the constant c_2.

Now note that the dependence term of interest is proportional to

$$\begin{aligned} &n^{-1}\sum_{t=1}^{n} E|E(\rho_t|\mathcal{F}_{t-m}) - E\rho_t| \\ &\le 6c_2\alpha_m(\{\rho_{t,k}\}_{t=-\infty}^{\infty})^{1-1/q} + 2c_1\nu_k \\ &\le 6c_2\alpha_{m-k}(\{V_t\}_{-\infty}^{\infty})^{1-1/q} + 2c_1\nu_k. \end{aligned}$$

Then one can take, for example, $k = \lceil m/2 \rceil$ (the integer part of $m/2$) and make the upperbound small for large m.

This shows that the current formulation of the inequality can handle dependence that is more general than strong mixing.

2.3 Proving the Triplex Inequality

The idea behind an upperbound with a decomposition into such three terms has appeared in econometric literature. For example, de Jong and Woutersen (2004) used this idea to treat an unbounded sum appearing in the binary choice models. The idea of our proof is related to a mixingale treatment seen in, for example, Chapter 16 of Davidson (1994). Since the specific form of the current inequality is not seen in these literatures, we will provide a proof below for completeness. The following Lemma will be used in the proof.

Lemma 1 *let $\{\mathcal{F}_t\}_{-\infty}^{\infty}$ be an increasing sequence of σ-fields. Let X_t be a random variable that is $\mathcal{F}_t$-measurable and is bounded so that $|X_t| \le C$ for some constant C for each t. Then for any $\varepsilon > 0$ and positive integers n, m, we have*

$$P[|\sum_{t=1}^{n} X_t - E\sum_{t=1}^{n} X_t| > n\varepsilon] \le 2me^{-n\varepsilon^2/(32m^2C^2)} + (2/\varepsilon)n^{-1}\sum_{t=1}^{n} E|E(X_t|\mathcal{F}_{t-m}) - EX_t|, \qquad (1)$$

as long as the right hand side exists.

Proof for Lemma 1 Consider $U_n \equiv \sum_{t=1}^n X_t - E\sum_{t=1}^n X_t$, which can be 'telescoped' into $U_n = U_{1,n} + U_{2,n} + ... + U_{m,n} + V_n$ where $U_{1,n} = \{X_1 - E(X_1|\mathcal{F}_{1-1})\} + ... + \{X_n - E(X_n|\mathcal{F}_{n-1})\}$, $U_{2,n} = \{E(X_1|\mathcal{F}_{1-1}) - E(X_1|\mathcal{F}_{1-2})\} + ... + \{E(X_n|\mathcal{F}_{n-1}) - E(X_n|\mathcal{F}_{n-2})\}$,..., $U_{m,n} = \{E(X_1|\mathcal{F}_{1-(m-1)}) - E(X_1|\mathcal{F}_{1-m})\} + ... + \{E(X_n|\mathcal{F}_{n-(m-1)}) - E(X_n|\mathcal{F}_{n-m})\}$, $V_n = \{E(X_1|\mathcal{F}_{1-m}) - EX_1\} + ... + \{E(X_n|\mathcal{F}_{n-m}) - EX_n\}$. Then a union bound leads to

$$\begin{aligned}
&P[|U_n| > n\varepsilon] \\
&\le P[|U_{1,n}| > n\varepsilon/(2m)] + P[|U_{2,n}| > n\varepsilon/(2m)] + ... + P[|U_{m,n}| > n\varepsilon/(2m)] \\
&+ P[|V_n| > n\varepsilon/2]. \qquad (2)
\end{aligned}$$

Note that $U_{1,n}$ is a sum of n martingale differences each bounded in magnitude by $2C$. So $P[|U_{1,n}| > n\varepsilon/(2m)] \le 2e^{-n\varepsilon^2/(32m^2C^2)}$ by applying a generalization of the Hoeffding's inequality to the martigale differences (see, e.g., Theorem 15.20, Davidson, 1994, or Theorem 9.1, Devroye, Györfi and Lugosi, 1996). Similarly is $P[|U_{j,n}| > n\varepsilon/(2m)] \le 2e^{-n\varepsilon^2/(32m^2C^2)}$ for all $j = 1, ..., m$. Now $P[|V_n| > n\varepsilon/2] \le (2/\varepsilon)n^{-1}E|V_n| \le (2/\varepsilon)n^{-1}\sum_{t=1}^n E|E(X_t|\mathcal{F}_{t-m}) - EX_t|$ using the Markov inequality and the triangular inequalities. Combining these upperbounds for the terms on the right hand side of (2) leads to the proof. Q.E.D.

The inequality appearing in the current lemma holds without assumption of a dependence structure. It still assumes a bounded X_t. Next we remove the boundedness assumption by incorporating a term related to the 'tail behavior' of a random variable ρ_t, which is now possibly unbounded.

Proof for Theorem 1 We will decompose $\rho_t = X_t + Y_t$ where $X_t = \rho_t I[|\rho_t| \le C]$ and $Y_t = \rho_t I[|\rho_t| > C]$. Then $|\sum(\rho_t - E\rho_t)| \le |\sum(X_t - EX_t)| + \sum|Y_t| + \sum E|Y_t|$ by using triangular inequalities. Then $P[\sum_1^n(\rho_t - E\rho_t)| > n\varepsilon] \le P[|\sum_1^n(X_t - EX_t)| > n\varepsilon/3] + P[\sum_1^n|Y_t| > n\varepsilon/3] + P[\sum_1^n E|Y_t| > n\varepsilon/3]$ The first term is bounded by the preceding lemma by

$$2me^{-n\varepsilon^2/(288m^2C^2)} + (6/\varepsilon)n^{-1}\sum_{t=1}^{n} E|E(X_t|\mathcal{F}_{t-m}) - EX_t|.$$

The second term is bounded above by $(3/\varepsilon)n^{-1}\sum_1^n E|Y_t|$. The third term is a probability of a deterministic event, which is zero if the righthand side of the triplex inequality in Theorem 1 does not exceed one. Therefore $P[\sum_1^n(\rho_t - E\rho_t)| > n\varepsilon] \le 2me^{-n\varepsilon^2/(288m^2C^2)} + (6/\varepsilon)n^{-1}\sum_{t=1}^n E|E(X_t|\mathcal{F}_{t-m}) - EX_t| + (3/\varepsilon)n^{-1}\sum_1^n E|Y_t|$. Now note that $|E|E(X_t|\mathcal{F}_{t-m}) - EX_t| - E|E(\rho_t|\mathcal{F}_{t-m}) - E\rho_t|| \le E|E(Y_t|\mathcal{F}_{t-m}) - EY_t| \le 2E|Y_t|$ using the triangular inequalities and the Jensen's inequality. Then $E|E(X_t|\mathcal{F}_{t-m}) - EX_t| \le E|E(\rho_t|\mathcal{F}_{t-m}) - E\rho_t| + 2E|Y_t|$ and $P[\sum_1^n(\rho_t - E\rho_t)| > n\varepsilon] \le 2me^{-n\varepsilon^2/(288m^2C^2)}$ $+(6/\varepsilon)n^{-1}\sum_{t=1}^n\{E|E(\rho_t|\mathcal{F}_{t-m}) - E\rho_t| + 2E|Y_t|\} + (3/\varepsilon)n^{-1}\sum_1^n E|Y_t|$ which leads to the proof of the theorem. Q.E.D.

3. Uniform Deviation

The above inequality (in Theorem 1) can be used to bound the probability of a large (pointwise) deviation $T_n(b) = \hat{R}(b) - R(b)$, where b is a parameter, $\hat{R}(b)$ is a sample average $\hat{R}(b) = n^{-1}\sum_{t=1}^n \rho(\omega_t, b)$ (where for each t, ω_t is measurable -$\mathcal{F}_t$ from an increasing sequence of σ-fields), and $R(b)$ is its expectation $R(b) = E\hat{R}(b)$. It is often of interest to bound the probability of a large *uniform* deviation $\sup_{b\in B}|T_n(b)|$ over a parameter space B for b.

The connection between the pointwise and uniform deviations can be obtained by covering B with many (say, Γ) smaller sets B_i's, so that $B \subset \cup_{i=1}^{\Gamma} B_i$. Choose b_i to be some parameter located in B_i for each i. Note that $\sup_{b\in B}|T_n(b)| \le \max_{i=1}^{\Gamma}|T_n(b_i)| + \max_{i=1}^{\Gamma}\sup_{b\in B_i}|T_n(b) - T_n(b_i)|$. Then a union bound leads to:

Proposition 1 *For any nonstochastic* $\delta > 0$ *and any positive integer n,*

$$P[\sup_{b\in B}|T_n(b)| > 2\delta] \le \sum_{i=1}^{\Gamma} P[|T_n(b_i)| > \delta] + \sum_{i=1}^{\Gamma} P[\sup_{b\in B_i}|T_n(b) - T_n(b_i)| > \delta]. \tag{3}$$

This is the basis for us to bound the probability of uniform large deviation. The first term involves pointwise deviations and can be bounded by the inequality derived before. The second term can be bounded when $T_n(b)$ 'often changes little' in a small set B_i. This can often achieved by assuming a Lipshitz condition for the summand $\rho(\omega_t, b)$ in argument b (see, e.g., Newey, 1991).

In machine learning, however, we often encounter summand $\rho(\omega_t, b)$ that is discontinuous in b. For example, the classification error can be written as $\rho(\omega_t, b) = |y_t - I[x_t'b > 0]|$, which is discontinuous in b, when a linear boundary (in predictor x_t) is used to classify a $\{0,1\}$ valued label y_t. (Here $\omega_t = (y_t, x_t)$.)

We will use a quite general framework that allows some continuous cases and some discontinuous cases as well as some 'mixed' cases. Let $\rho(\omega_t, b)$ be of the form $\rho(\omega_t, b) = f_t(b, A_t(b))$ where $f_t(\cdot,\cdot)$ is continuous in the first argument, but the second argument $A_t(b) = I[g(\omega_t, b) > 0]$ for some fixed function g that determines a decision boundary. The function f_t depends on t through observation ω_t. This framework can then include the following examples:

- (continuous) L_1-loss: $\rho = |y_t - x_t'b|$ (when $f_t(b,\cdot)$ is constant);

- (discontinuous) classification loss: $\rho = |y_t - I[x_t'b > 0]|$ (when $f_t(\cdot, A_t(b))$ is constant);

- 'mixed' loss such as $\rho = (1 - y_t)\alpha(x_t'b)I[\alpha(x_t'b) > 0]$, which may result from a loan decision of lending out amount $\alpha(x_t'b)$ (according to a continuous parametric model α) when $100y_t\%$ of the loan is paid back.

Under this framework we will bound the deviation $\sup_{b\in B_i}|T_n(b)-T_n(b_i)|$. This is summarized in the following Proposition, the proof of which is included in the Appendix.

Proposition 2 *For each parameter b in a convex set B_i that contains b_i, denote $T_n(b)=\hat{R}(b)-E\hat{R}(b)$, where $\hat{R}(b)=n^{-1}\sum_{t=1}^{n}\rho(\omega_t,b)$ and for each t, ω_t is measurable -$\mathcal{F}_t$ (from an increasing sequence of σ-fields). Assume that ρ has the form $\rho(\omega_t,b)=f_t(b,A_t(b))$ where $A_t(b)=I[g(\omega_t,b)>0]$ for some fixed function g that determines a decision boundary.*

Define S_i as the 'boundary set' $S_i=\cup_{b\in B_i}\{\omega_t: g(\omega_t,b)=0\}$. Assume that:

(A1): $g(\omega_t,b)$ is continuous in b and measurable in ω_t;

(A2): (Lipshitz condition) $\sup_{a=0,1}|f_t(b,a)-f_t(b^*,a))|\le N_{it}|b-b^*|_q$ for some $q>0$, for any $b,b^*\in B_i$;[2]

(A3): (Small boundary condition) The boundary set S_i is measurable and $n^{-1}\sum_{t=1}^{n}EI(\omega_t\in S_i)\le\delta/(12C)$ for some constants $\delta,C>0$.

Denote $\lambda=\sup_{b,b^*\in B_i}|b-b^*|_q$ and $M_{it}=|f_t(b_i,1)-f_t(b_i,0)|$.

For any constants $\delta,C>0$ and positive integers n,m, if (A3) holds, then we have:

$$\begin{aligned}
&P[\sup_{b\in B_i}|T_n(b)-T_n(b_i)|>\delta]\\
&\le(6\lambda/\delta)n^{-1}\sum_{t=1}^{n}E(N_{it}+EN_{it})I[N_{it}+EN_{it}>\delta/(6\lambda)]\\
&+(6/\delta)n^{-1}\sum_{t=1}^{n}EM_{it}I(M_{it}>C)\\
&+2me^{-n\delta^2/(1152m^2C^2)}+(12C/\delta)n^{-1}\sum_{t=1}^{n}E|E(I(\omega_t\in S_i)|\mathcal{F}_{t-m})-EI(\omega_t\in S_i)|,
\end{aligned}$$

as long as the right hand side exists.

3.1 Remarks

6. In the 'continuous case', $f_t(b,a)$ is constant in $a\in\{0,1\}$. We can then drop the last three terms in the above bound. This is because $M_{it}=0$ in this case and we can take $C\to 0$. In the 'discontinuous case', $f_t(b,a)$ is constant in b and we can drop the first term in the above bound, since the Lipshitz constant N_{it} can be taken as 0. The result above holds also for the more general 'mixed case' when $f_t(b,a)$ varies with both b and a.

7. Assumption A2 can often be validated by bounding the partial derivative of $f_t(b,a)$ on the first argument. In a later example with L_1 loss we will use a triangular inequality to validate this assumption.

8. Assumption A3 is related to $P(\omega_t\in S_i)$, the probability of an observation falling in the 'boundary set' S_i corresponding to a parameter set B_i. When B_i is small enough, we expect that the 'boundary set' will have small probability and A3 can be satisfied. The situation is clarified when $\omega_t=(y_t,x_t)$, $g(\omega_t,b)$ depends on ω_t only through predictor x_t, and $x_t=(w_t,v_t')'$ has a

[2] For a vector v with component v_j's, define the ℓ_q norm as $|v|_q=(\sum_{j=1}^{\dim(v)}|v_j|^q)^{1/q}$ for $q\in(0,\infty)$, and $|v|_\infty=\sup_{j=1}^{\dim(v)}|v_j|$. We will also formally denote $|v|_0=\sum_{j=1}^{\dim(v)}I[|v_j|>0]$.

scalar component w_t and other components v_t, so that the decision boundary $[g(\omega_t,b)=0]$ 'can be solved' as $[w_t = w(v_t,b)]$ for some fixed function w. In this case the boundary set $S_i = \{(w_t,v_t): w_t = w(v_t,b), b\in B_i\} = [\inf_{b\in B_i} w(v_t,b) \le w_t \le \sup_{b\in B_i} w(v_t,b)]$ if B_i is compact and $w(v_t,b)$ is continuous in b. Suppose we use the ℓ_∞ norm and define $\lambda = \sup_{b,b^*\in B_i}|b-b^*|_\infty$. Denote $d_0 = \sup_{b,b^*\in B_i}|b-b^*|_0$. Then in the Appendix we will show that

$$P(\omega_t \in S_i) \le E_{\mathcal{V}_t}\{\sup_{w_t} p(w_t|\mathcal{V}_t)\sup_{b\in B_i}|\partial_b w(v_t,b)|_\infty\}\lambda d_0. \tag{4}$$

Here $\partial_b w(v_t,b)$ denotes a partial derivative of w, $\mathcal{V}_t$ is some σ-field such that v_t is measurable -$\mathcal{V}_t$ for each t, and $p(w_t|\mathcal{V}_t)$ denotes the conditional density. In the 'linear case'[3] assuming $\lambda < 2$, $g(\omega_t,b) = \pm w_t + v_t' b_v$ (so A1 is satisfied), we can take $w(v_t,b) = \mp v_t' b_v$ and $\sup_{b\in B_i}|\partial_b w(v_t,b)|_\infty = |v_t|_\infty$. If the conditional density is bounded above by constant c, then (4) becomes $P(\omega_t\in S_i) \le cE|v_t|_\infty \lambda d_0$. The assumption A3 will be satisfied for choosing $\lambda \le \delta/(12Cc\sup_t E|v_t|_\infty d_0)$, which restricts the size of B_i.

9. It is also noted that in this paper, we will consider 'boundary sets' of the 'solvable' form $S_i = [\inf_{b\in B_i} w(v_t,b) \le w_t \le \sup_{b\in B_i} w(v_t,b)]$ which is assumed to be measurable. In our later examples we will focus on 'linear solvable type' described above, with decision boundary $[w_t = \mp v_t' b_v]$, and B_i being a closed ℓ_∞ ball centered at b_i and with radius $h = \lambda/2 > 0$. Then $S_i = [\mp v_t'(b_i)_v - h|v_t|_1 \le w_t \le \mp v_t'(b_i)_v + h|v_t|_1]$ which is indeed measurable. [More generally, when $\sup_{b\in B_i} w(v_t,b)$ and $\inf_{b\in B_i} w(v_t,b)$ are both continuous in v_t, S_i is measurable.]

We will analyze the bound in Proposition 2 term by term in the later examples. The terms $E(N_{it}+EN_{it})I[N_{it}+EN_{it} > \delta/(6\lambda)]$ and $EM_{it}I(M_{it} > C)$ are tail terms. We can choose sufficiently small λ and sufficiently large C to make them small.

The dependence term $E|E(I(\omega_t\in S_i)|\mathcal{F}_{t-m}) - EI(\omega_t \in S_i)|$ will be small for large m when ω_t can be approximated by strong mixing sequences in some sense.

The exponential term $2me^{-n\delta^2/(1152m^2C^2)}$ can be made small by choosing m and C to depend on n in certain ways. We can allow δ to depend on n also, which will lead to convergence rates.

4. A Continuous Example

Consider $\rho_t = \rho(\omega_t,b) = |Y_t - b_1 Y_{t-1} - ... - b_r Y_{t-r}|$ (where $\omega_t = (Y_t,...,Y_{t-r})$ and $b = (b_1,...,b_r)$), which represents predicting Y_t by r of its own lags under an L_1 loss. We will allow r to increase with n later to allow high dimensionality. We will bound the probability of a large uniform deviation $\sup_{b\in B}|n^{-1}\sum_{t=1}^n \rho_t - n^{-1}\sum_{t=1}^n E\rho_t|$ over an ℓ_∞ ball $B = [|b|_\infty \le C_b]$ with a constant radius $C_b > 0$.

Suppose the true model for Y_t follows an $MA(\infty)$ model $Y_t = \sum_{j=0}^\infty \theta_j Z_{t-j}$, where θ_j's are fixed coefficients with a finite ℓ_1 norm $|\theta|_1 = \sum_0^\infty |\theta_j| < \infty$, and Z_j's are 'innovations', which are assumed to be iid (independent and identically distributed) with zero mean and finite variance. Although we have assumed Y_t to be centered to have mean zero and that there is no intercept term used in the L_1 loss, this is only for convenience and similar results can be obtained without this assumption.

We will consider a case of exponentially decaying θ_j's, but each θ_j can be nonzero:

[3] In the 'linear case' the decision rule $[g(\omega_t,b) > 0]$ has the form $[w_t b_w + v_t' b_v > 0]$. We can always rescale the coefficients $b = (b_w, b_v')'$ by a scalar multiple. One such standardization used in Horowitz (1992) is such that $|b_w| = 1$ or $b_w \in \{-1,+1\}$. Note that for small enough set B_i with $\lambda = \sup_{b,b'\in B_i}|b-b'|_\infty < 2$, b_w is constant in B_i and takes a common sign. We can pick either sign to proceed.

Condition (B1): $\sum_{j=k}^{\infty}|\theta_j| < \nu^k$ *for all large enough k, for some* $\nu \in (0,1)$.

This is a situation when Y_t can have a dependence structure that is not strong mixing, since it involves the infinite past of Z_{t-j}'s (see, e.g., Davidson, 1994, Section 14.3). On the other hand, the dependence term in Theorem 1 can be bounded by L_1 approximation of strong mixing and we have the following result (proved in Appendix, where $\mathcal{F}_t$ is the σ-field generated by $\{Z_s\}_{s=-\infty}^{t}$ for each t):

$$E|E(\rho_t|\mathcal{F}_{t-m}) - E\rho_t| \leq 2(r+1)(C_b+1)E|Z_1|\sum_{j=k+1}^{\infty}|\theta_j| \text{ for any positive integer } k < m-r. \quad (5)$$

In order to bound the tail term in Theorem 1, we use, for some finite constants $u, C_u > 0$,

$$E|\rho_t|I(|\rho_t| > C) \leq (r+1)^2(C_b+1)||Z_1||_2|\theta|_1 C_u e^{-uC(C_b+1)^{-1}(r+1)^{-1}/2}, \quad (6)$$

which is proved in the Appendix assuming an additional condition:
Condition (B2): For innovation Z_1*, the cumulant generating function* $K(u) = \ln Ee^{Z_1 u}$ *is continuously differentiable at* 0*. (E.g.,* Z_1 *can be a Gaussian innovation.)*

Now we apply Proposition 2, where only the first term of the bound is relevant in this continuous case due to Remark 6. The Lipshitz constant N_{it} can be obtained from the triangular inequality $|\rho(\omega_t, b) - \rho(\omega_t, b^*)| \leq |b_1 - b_1^*||Y_{t-1}| + ... + |b_r - b_r^*||Y_{t-r}| \leq (|Y_{t-1}| + ... + |Y_{t-r}|)|b - b^*|_\infty$. So we can take $N_{it} = |Y_{t-1}| + ... + |Y_{t-r}|$ (using ℓ_∞ norm). We have

$$E(N_{it} + EN_{it})I(N_{it} + EN_{it} > \delta/(6\lambda)) \leq (2r^2|\theta|_1||Z_1||_2)C_u e^{-u\psi/2}, \quad (7)$$

for some finite constants $u, C_u > 0$, where $\psi = \delta/(6r\lambda) - E|Z_1||\theta|_1$, which is proved in the Appendix.

Now we apply (3), Theorem 1 and Proposition 2 and combine all terms together (using (5), (6) and (7)) to obtain:

For any positive integers $k < m - r$, m, n, and positive C, δ, we have

$$\begin{aligned} & P[\sup_{b \in B}|n^{-1}\sum_{t=1}^{n}(\rho_n - E\rho_t)| > 2\delta] \\ & \leq \Gamma 2me^{-n\delta^2/(288m^2C^2)} \\ & + \Gamma(6/\delta)2(r+1)(C_b+1)E|Z_1|\sum_{j=k+1}^{\infty}|\theta_j| \\ & + \Gamma(15/\delta)(r+1)^2(C_b+1)||Z_1||_2|\theta|_1 C_u e^{-uC(C_b+1)^{-1}(r+1)^{-1}/2} \\ & + \Gamma(6\lambda/\delta)(2r^2|\theta|_1||Z_1||_2)C_u e^{-u(\delta/(6r\lambda) - E|Z_1||\theta|_1)/2}. \end{aligned} \quad (8)$$

Here $B = [|b|_\infty \leq C_b] = [-C_b, C_b]^r$ for some constant radius $C_b > 0$. We will consider a high dimensional case where the number of lags can increase with sample size n:
Condition (B3): $r = O((\ln n)^M)$ *for some power* $M > 0$.

Note that we can take $B_1, ..., B_\Gamma$ to be Γ closed ℓ_∞ balls of radius $\lambda/2$ to cover B, where $\Gamma \leq (2C_b/\lambda + 1)^r$.

We will let $\delta = n^{-0.5+\gamma_1}/2$ for some small $\gamma_1 > 0$, $m = C = \lceil n^{\gamma_1/4} \rceil$, $\lambda = n^{-1}$, $k = m - 2r$. Then under conditions (B1) to (B3), $\ln\Gamma = O((\ln n)^{M_1})$ for some $M_1 > 0$ and all four terms in the above inequality (8) are $O(e^{-c_1 n^{c_2}})$ for some $c_1, c_2 > 0$ dependent on γ_1. Therefore we have:

Proposition 3 *Under Conditions (B1) to (B3), for any small $\gamma_1 > 0$,*

$$P[\sup_{b\in B}|n^{-1}\sum_{t=1}^{n}(\rho_t - E\rho_t)| > n^{-0.5+\gamma_1}] = O(e^{-c_1 n^{c_2}}).$$

The rate of uniform convergence remains nearly 'parametric' $O_P(n^{-0.5})$ in this case, despite the high dimensionality of set B.

5. A Discontinuous Example

Let $\omega_t = (y_t, x_t)$, where y_t is a real-valued response at t and $x_t = (1, y_{t-1}, ..., y_{t-r}, z_t')'$ is a vector of predictors that can include r lags of y as well as a vector of exogenous variable z_t. Suppose we are interested in predicting the sign of y_t by using a discontinuous loss $\rho_t = |I(y_t > 0) - I(x_t'b > 0)|$. We will bound the probability of a large uniform deviation $\sup_{b\in B}|n^{-1}\sum_{t=1}^{n}(\rho_t - E\rho_t)|$ over a set of 'variable selection' $B = [|b|_0 \le v, |b|_\infty \le C_b, |b_{r+2}| = 1]$, where $|b|_0 \equiv \sum_{j=1}^{\dim(b)} I|b_j| > 0]$ counts the number of selected x-components, $|b|_\infty \le C_b$ bounds the parameter space, and $|b_{r+2}| = 1$ is due to a standardization for the coefficient of the first component of z_t (see Footnote 2). Later we will allow v (maximal number of selected variables), r (number of lags allowed) and $K \equiv \dim(z_t)$ to increase with n in certain ways for high dimensional variable selection.

The true model of y_t is assumed to be an $MA(\infty)$ transform of a strong mixing process: $y_t = \sum_{j=0}^{\infty}\theta_j f_j(z_{t-j}, \varepsilon_{t-j})$, where f_j is some fixed measurable function for each j so that $\sup_{t,j}||f_j(z_t, \varepsilon_t)||_1 < \infty$, and ε_t is a stochastic sequence independent of z_t called the 'innovation'. We assume that:

Condition (C1): $\{z_t, \varepsilon_t\}$ is strong mixing with mixing coefficient α_m decreasing exponentially fast in m.

Condition (C2): $\sum_{j=k+1}^{\infty}|\theta_j|$ decreases exponentially fast in k.

Note that y_t itself may no longer be strong mixing due to its dependence on the infinite past. These assumptions are satisfied in many situations. For example, in an ARX model $y_t = \varphi y_{t-1} + z_t'\beta + \varepsilon_t$ ($|\varphi| < 1$), an $MA(\infty)$ representation gives $y_t = \sum_{j=0}^{\infty}\varphi^j(z_{t-j}'\beta + \varepsilon_{t-j})$. Here, ε_t does not have to be iid; it can be an 'exponential' strong (or β-) mixing process such as a GARCH process (see, e.g., Francq and Zakoïan, 2006) when y_t follows an ARX-GARCH model.

In the Appendix, we show that under some additional conditions (C3 and C4) on the underlying process $\{z_t, \varepsilon_t\}$, we have, for any positive integer $k < m - r$,

$$E|E(\rho_t|\mathcal{F}_{t-m}) - E\rho_t| \le 6\alpha_{m-r-k} + 8(\sqrt{2M_y} + \sqrt{2M_x r C_b})\sqrt{\sup_{t,l}||f_l(z_t, \varepsilon_t)||_1 \sum_{j>k}|\theta_j|}. \qquad (9)$$

Here M_x, M_y are constants appearing in these additional conditions:

Condition (C3): y_t follows a model of the form $y_t = F_t + \varepsilon_t$ where F_t depends on the history $(\{z_s\}_{\infty}^{t}, \{\varepsilon_s\}_{-\infty}^{t-1})$ and ε_t is an innovation that has a conditional density $p(\varepsilon_t|\{z_s\}_{\infty}^{t}, \{\varepsilon_s\}_{-\infty}^{t-1})$ bounded above by a constant M_y (which is satisfied, for example, by $N(0, \sigma^2)$ innovations).

Condition (C4): The conditional density $p(z_{t,1}|z_{t,2}, ..., z_{t,K}, \{z_s, \varepsilon_s\}_{-\infty}^{t-1})$ is bounded above by a constant M_x.

We can cover B by Γ sets $B_i \subset B$, $i = 1, ..., \Gamma$, with each B_i being a closed ℓ_∞ ball centered at some b_i, with radius $h = \lambda/2 > 0$, and with dimension at most $v - 1$. This is explained in the Appendix, where we also show that we can take

$$\Gamma \le 2v(K + r)^{v-1}(2C_b/\lambda + 1)^{v-1} \qquad (10)$$

as an upperbound obtained from a combinatorial argument. Now we try to apply Proposition 2.

Assumption (A1) is obviously satisfied since $g(\omega_t, b) = x_t'b$. We will show that assumption (A3) holds for all large n in the Appendix, with an additional condition (C5) and with suitable choices of δ, λ (the ℓ_∞ diameter of B_i), r (the number of lags) and v (the maximal number of selected variables) to be specified later. This additional condition is:

Condition (C5): The exogenous variables are bounded above by a finite constant: $|z_t|_\infty \leq C_z$.

Assumption (A2) is satisfied with $N_{it} = 0$ due to this 'discrete' situation (see Remark 6). So the first term of the bound in Proposition 2 is zero. We can take $M_{it} = ||I(y_t > 0) - 1| - |I(y_t > 0) - 0|| = 1$ and $C = 1$ so the second term of the bound is zero also.

In the Appendix we evaluate the last term which is determined by $E|E(I(\omega_t \in S_i)|\mathcal{F}_{t-m}) - EI(\omega_t \in S_i)|$. Assuming (C4), we have, for any positive integer $k < m - r$,

$$E|E(I(\omega_t \in S_i)|\mathcal{F}_{t-m}) - EI(\omega_t \in S_i)| \leq 6\alpha_{m-r-k} + 4\sqrt{r\sum_{j>k}|\theta_j|\sup_{t,l}||f_l(z_t,\varepsilon_t)||_1(2M_x 2(1+C_b+h))}. \tag{11}$$

Now combine the applications of (3), Theorem 1 (with $C = 1$, or just use Lemma 1), and Proposition 2, apply Equations (9) and (11) and we obtain:

For any positive integers $k < m - r$, m, n, and positive δ,

$$\begin{aligned}
&P[\sum_{b\in B}|n^{-1}\sum_{t=1}^{n}(\rho_t - E\rho_t)| > 2\delta] \\
&\leq \Gamma 2me^{-n\delta^2/(32m^2)} \\
&+\Gamma(2/\delta)\left\{6\alpha_{m-r-k} + 8(\sqrt{2M_y} + \sqrt{2M_x rC_b})\sqrt{\sup_{t,l}||f_l(z_t,\varepsilon_t)||_1\sum_{j>k}|\theta_j|}\right\} \\
&+\Gamma 2me^{-n\delta^2/(1152m^2)} \\
&+\Gamma(12/\delta)\left\{6\alpha_{m-r-k} + 4\sqrt{r\sum_{j>k}|\theta_j|\sup_{t,l}||f_l(z_t,\varepsilon_t)||_1(2M_x 2(1+C_b+\lambda/2))}\right\}.
\end{aligned} \tag{12}$$

Now choose parameters to make the bound small. Note that we can take $\Gamma \leq 2v(K+r)^{v-1}(2C_b/\lambda + 1)^{v-1}$. We have assumed exponential decay for α_k and $\sum_{j>k}|\theta_j|$ in k.

Let $C_b > 0$ be a constant in n. Assume the following condition on the various dimension parameters:

Condition (C6): The number of lags $r = O((\ln n)^{M_1})$ for some power $M_1 > 0$; the number of exogenous variables $K = O(n^{M_2})$ for some power $M_2 > 0$, which can form a very-high dimensional candidate predictor, with dimension possibly large than sample size n; the number of selected variables $v = O((\ln n)^{M_3})$ for some power $M_3 > 0$.

We will let $\delta = n^{-0.5+\gamma_1}/2$ for some small $\gamma_1 > 0$, $m = \lceil n^{\gamma_1/4}\rceil$, $\lambda = n^{-1}$, $k = \lceil (m-r)/2\rceil$. Then $\ln\Gamma = O((\ln n)^{M_4})$ for some $M_4 > 0$ and all four terms in the above inequality (12) are $O(e^{-c_1 n^{c_2}})$ (for some $c_1, c_2 > 0$ dependent on γ_1), when Conditions (C1) to (C6) are assumed. Therefore we have:

Proposition 4 *Under conditions (C1) to (C6), for any small $\gamma_1 > 0$,*

$$P[\sup_{b\in B}|n^{-1}\sum_{t=1}^{n}(\rho_n - E\rho_t)| > n^{-0.5+\gamma_1}] = O(e^{-c_1 n^{c_2}}).$$

The rate of uniform convergence remains nearly ‘parametric’ $O_P(n^{-0.5})$ in this case, despite the high dimensionality of set B.

6. Discussion

This paper presents a very general inequality that generalizes Hoeffding’s inequality to dependent and unbounded summands. The inequality may not be very tight, but it involves few assumptions and can be very useful in deriving convergence rates of pointwise and uniform deviations in a number of situations that cannot be dealt with before. We gave two examples here, one with L_1 loss and another on sign classification. There are other examples that may be worked out (e.g., with L_2 loss or with log-likelihood) which are not considered here. We hope that the current work can serve as a probablistic foundation to the theory of empirical risk minimization for many situations with dependent data and unbounded loss.

The current results involve a high dimensional parameter b; near-parametric convergence rates are obtained in examples with exponentially small ‘unboundedness’ (characterized by existence of some moment generating function) and with certain kinds of exponentially decaying temporal dependence. We expect that slower convergence rates may be obtained with more severe ‘unboundedness’, or with a slower decay of temporal dependence, using the same techniques.

Although the number of selected variables is restricted to $O((\log n)^M)$, we can allow these variables to be selected from a much higher number of candidate regressors of dimension up to n^M for any finite positive M, and still maintain a near-parametric convergence rate. This is demonstrated in Section 5 and is also true if we add in regressors and make an ARX model for Section 4. (In fact it is also possible to allow a higher number of selected variables such as n^a for some $a \in (0,1)$, but this will correspond to a slower convergence rate.)

It is noted that there exists much previous work in addressing the problems considered in this paper, in addition to the related work mentioned in the Introduction (we thank the reviewers for bringing our attention to these additional references). In the direction of unbounded loss, Meir and Zhang (2003) consider uniform deviations for iid data using a bound of the Rademaker complexity. Various ratios of empirical processes can also be used to handle unboundedness (see, e.g., Haussler, 1992; Pollard, 1995; Bartlett and Lugosi, 1999; Bercu, Gassiat and Rio, 2002). Bercu, Gassiat and Rio (2002) present a ratio-type result that can relax the assumption on the ‘unboundedness’ while proving exponential concentration for iid data. In the direction of dependence, McDiarmid (1998, e.g., Theorem 3.8) uses the method of bounded differences and includes a term related to a ‘bad set’ of events involving a large variance. Uniform inequalities have also been obtained for martingales (see, e.g., van de Geer, 2000, Theorem 8.13). In the direction of high dimensionality, typical uniform deviation results deal with infinite-dimensional function space (see, e.g, good summaries in Bousquet, Boucheron, and Lugosi, 2003 and van de Geer, 2000).

In comparison, our method addresses the three aspects (unboundedness, dependence, high dimensionality) simultaneously with a relative simple approach. While the additional references can sometimes handle one aspect better, they typically do not address the other aspects in the same time. For example, McDiarmid (1998, e.g., Theorem 3.8) can potentially handle data that are dependent ‘in multi-dimensions’ (such as random graph or spatial dependence), while our method only addresses ‘one-dimensional’ dependence (i.e., a time series). On the other hand, McDiarmid’s method also requires bounded differences which would require some kind of boundedness (e.g., bounded summands in the case of an iid average) while we do not need this assumption. In addition, we note that the ‘bad set’ term bounding the probability of a large variance will often require applying a large deviation inequality again, while our triplex inequality is one in closed form.

Many of these additional references do not treat dependence to the same degree of generality as the current paper. For example, traditional treatments with symmetrization and Rademaker average such as Meir and Zhang (2003) are suitable only for iid data. Ratio-type results such as Bercu, Gassiat, Rio (2002) can relax the assumption on the ‘unboundedness’ while proving exponential concentration for iid data, but it is not clear how this can be extended to general dependent situations. The chaining technique described in, for example, van de Geer (2000, Section 3.2) and Bousquet, Boucheron and Lugosi (2003, Section 5) may be used to improve the convergence rate by a $\log n$ factor in the finite dimensional case. However, most applications of this technique are for independent data or martingales. Our dependence term, on the other hand, is put in the framework of mixingale, which is more general than martingales as discussed in Remark 4.

Typical uniform deviation results deal with infinite-dimensional function spaces. Our formulation is for a somewhat less general situation where the functions are parameterized, and we formulated uniform convergence on a high dimensional parameter space. Although it may be possible to formulate uniform convergence on function spaces directly for dependent data (see, e.g., Yu, 1994 for beta-mixing and Vidyasagar, 2005 for alpha-mixing), we choose the parametric covering framework which is less abstract and easier to understand, and demonstrates the convergence rates more clearly. The result of such a formulation is that a reader with an elementary background on probability and real analysis can follow the development easily, and arrive at such advanced results as the convergence rates with high dimensional variable selection. Although the formulation is deliberately elementary, the results are powerful enough to handle such complicated dependent situations as the sign prediction for an ARX process with GARCH error in Section 5.

Acknowledgments

Supported in part by NSF Grant DMS-0706885. We thank the reviewers and the editors for providing additional references and helpful suggestions.

Appendix A.

Proof of Proposition 2 Note that for any $b \in B_i$, $|T_n(b) - T_n(b_i)| = |n^{-1}\sum_{t=1}^n \{f_t(b, A_t(b)) - f_t(b_i, A_t(b_i))\} - n^{-1}\sum_{t=1}^n E\{f_t(b, A_t(b)) - f_t(b_i, A_t(b_i))\}|$. We therefore investigate the differences of the form $f_t(b, A_t(b)) - f_t(b_i, A_t(b_i)) = \{f_t(b, A_t(b)) - f_t(b_i, A_t(b))\} + \{f_t(b_i, A_t(b)) - f_t(b_i, A_t(b_i))\}$.

Note that $|f_t(b_i, A_t(b)) - f_t(b_i, A_t(b_i))| = M_{it}|A_t(b) - A_t(b_i)|$ where $M_{it} = |f_t(b_i, 1) - f_t(b_i, 0)|$, and $|f_t(b, A_t(b)) - f_t(b_i, A_t(b))| \leq \sup_{b,b^* \in B_i} \sup_{a=0,1} |f_t(b, a) - f_t(b^*, a))| \leq N_{it}\lambda$, where $\lambda \equiv \sup_{b,b^* \in B_i} |b - b^*|_q$, if we assume a Lipshitz condition $\sup_{a=0,1} |f_t(b, a) - f_t(b^*, a))| \leq N_{it}|b - b^*|_q$ under an ℓ_q-norm for some $q > 0$.

We then combine the statements before and apply the triangular inequalities to obtain $|T_n(b) - T_n(b_i)| \leq n^{-1}\sum_{t=1}^n N_{it}\lambda + n^{-1}\sum_{t=1}^n M_{it}|A_t(b) - A_t(b_i)| + n^{-1}\sum_{t=1}^n EN_{it}\lambda$ $+ n^{-1}\sum_{t=1}^n EM_{it}|A_t(b) - A_t(b_i)|$.

Now note that $|A_t(b) - A_t(b_i)| \leq I(\omega_t \in S_i)$ where S_i is the ‘boundary set $S_i = \cup_{b \in B_i}\{\omega_t : g(\omega_t, b) = 0\}$. This is true when we assume that $g(\omega_t, b)$ is continuous in b and B_i is convex.

[The difference $|A_t(b)-A_t(b_i)|$ is $\{0,1\}$ valued and takes value 1 only when only one of $g(\omega_t,b)$ and $g(\omega_t,b_i)$ is positive. This would imply $g(\omega_t,b^*)=0$ at some intermediate point b^* on the line segment between b and b_i, which must fall in B_i due to its convexity. A similar technique is used in Jiang and Tanner (2007) for a binary choice model with $g(\omega_t,b)$ linear in b.]

The above statements hold for any $b\in B_i$. Therefore

$$\begin{aligned}
&\sup_{b\in B_i}|T_n(b)-T_n(b_i)|\\
&\le n^{-1}\sum_{t=1}^n N_{it}\lambda+n^{-1}\sum_{t=1}^n M_{it}I(\omega_t\in S_i)+n^{-1}\sum_{t=1}^n EN_{it}\lambda+n^{-1}\sum_{t=1}^n EM_{it}I(\omega_t\in S_i)\\
&\le n^{-1}\sum_{t=1}^n (N_{it}+EN_{it})\lambda+n^{-1}\sum_{t=1}^n CI(\omega_t\in S_i)+n^{-1}\sum_{t=1}^n CEI(\omega_t\in S_i)\\
&+n^{-1}\sum_{t=1}^n\{M_{it}I(M_{it}>C)+EM_{it}I(M_{it}>C)\},
\end{aligned}$$

by noting that $M_{it}I(\omega_t\in S_i)\le CI(\omega_t\in S_i)+M_{it}I(M_{it}>C)$ for any constant $C>0$.

Then

$$\begin{aligned}
&P[\sup_{b\in B_i}|T_n(b)-T_n(b_i)|>\delta]\\
&\le P[n^{-1}\sum_{t=1}^n (N_{it}+EN_{it})\lambda>\delta/3]\\
&+P[n^{-1}\sum_{t=1}^n CI(\omega_t\in S_i)+n^{-1}\sum_{t=1}^n CEI(\omega_t\in S_i)>\delta/3]\\
&+P[n^{-1}\sum_{t=1}^n\{M_{it}I(M_{it}>C)+EM_{it}I(M_{it}>C)\}>\delta/3]\\
&\le P[n^{-1}\sum_{t=1}^n (N_{it}+EN_{it})\lambda>\delta/3]\\
&+(6/\delta)n^{-1}\sum_{t=1}^n EM_{it}I(M_{it}>C)\\
&+P[n^{-1}\sum_{t=1}^n\{I(\omega_t\in S_i)-EI(\omega_t\in S_i)\}>\delta/(3C)-2n^{-1}\sum_{t=1}^n EI(\omega_t\in S_i)]\\
&\le (6\lambda/\delta)n^{-1}\sum_{t=1}^n E(N_{it}+EN_{it})I[N_{it}+EN_{it}>\delta/(6\lambda)]\\
&+(6/\delta)n^{-1}\sum_{t=1}^n EM_{it}I(M_{it}>C)\\
&+P[n^{-1}\sum_{t=1}^n\{CI(\omega_t\in S_i)-CEI(\omega_t\in S_i)\}>\delta/6],
\end{aligned}$$

where in the last step we assume that $n^{-1}\sum_{t=1}^n EI(\omega_t\in S_i)\le\delta/(12C)$ and have used $P[n^{-1}\sum_{t=1}^n X_t>2Q]\le Q^{-1}n^{-1}\sum_{t=1}^n EX_tI[X_t>Q]$ for nonnegative $X=N_{it}+EN_{it}$ and constant $Q=\delta/(6\lambda)$. [Note that $n^{-1}\sum_{t=1}^n X_t=n^{-1}\sum_{t=1}^n X_tI[X_t>Q]+n^{-1}\sum_{t=1}^n X_tI[X_t\le Q]\le Q+n^{-1}\sum_{t=1}^n X_tI[X_t>Q]$ and use Markov inequality.]

Now apply the pointwise inequality (1) on the last term and we obtain Proposition 2. Q.E.D.

Proof of Equation (4) in Remark 8

$$\begin{aligned}
&P(\omega_t \in S_i) = P[w_t \in [\inf_{b\in B_i} w(v_t,b), \sup_{b\in B_i} w(v_t,b)]] \\
&\leq E_{\mathcal{V}_t} \sup_{w_t} p(w_t|\mathcal{V}_t)|\sup_{b\in B_i} w(v_t,b) - \inf_{b\in B_i} w(v_t,b)| \\
&\leq E_{\mathcal{V}_t} \sup_{w_t} p(w_t|\mathcal{V}_t) \sup_{b,b^*\in B_i} |w(v_t,b) - w(v_t,b^*)| \\
&\leq E_{\mathcal{V}_t}\{\sup_{w_t} p(w_t|\mathcal{V}_t) \sup_{b\in B_i} |\partial_b w(v_t,b)|_\infty\} \sup_{b,b^*\in B_i} |b-b^*|_\infty \sup_{b,b^*\in B_i} |b-b^*|_0 \\
&= E_{\mathcal{V}_t}\{\sup_{w_t} p(w_t|\mathcal{V}_t) \sup_{b\in B_i} |\partial_b w(v_t,b)|_\infty\}\lambda d_0.
\end{aligned}$$

Q.E.D.

Proof of Equation (5) Define $Y_{t,k} = \sum_{j=0}^{k} \theta_j Z_{t-j}$ and $\rho_{t,k} = |Y_{t,k} - b_1 Y_{t-1,k} - \ldots - b_r Y_{t-r,k}|$, which are strong mixing due to dependence on finite number of lags. We then have the L_1 approximation error $E|\rho_t - \rho_{t,k}| \leq (C_b+1)E(|Y_t - Y_{t,k}| + \ldots + |Y_{t-r} - Y_{t-r,k}|) \leq (r+1)(C_b+1)E|Z_1|\sum_{j=k+1}^{\infty}|\theta_j|$. Then the technique in Section 2.2 implies that $E|E(\rho_t|\mathcal{F}_{t-m}) - E\rho_t| \leq 6\sup_{t,k}||\rho_{t,k}||_2 \alpha_{m-k-r}(\{Z_t\})^{1/2}$ $+2(r+1)(C_b+1)E|Z_1|\sum_{j=k+1}^{\infty}|\theta_j|$.

Note that $\alpha_{m-k-r}(\{Z_t\}) = 0$ for $m > k+r$, We have (5). Q.E.D.

Proof of Equation (6) We note that $|\rho_t| \leq (C_b+1)(|Y_t| + \ldots + |Y_{t-r}|)$ and therefore $E|\rho_t|I(|\rho_t| > C) \leq (C_b+1)E(|Y_t| + \ldots + |Y_{t-r}|)\sum_{s=0}^{r} I(|Y_{t-s}| > C(C_b+1)^{-1}(r+1)^{-1}) \leq (1+r)^2(C_b+1)\sup_{t,s} E|Y_t|I(|Y_s| > C(C_b+1)^{-1}(r+1)^{-1})$. Note that for $\eta = C(C_b+1)^{-1}(r+1)^{-1}$, we have

$$\begin{aligned}
&E|Y_t|I(|Y_s| > \eta) \\
&\leq ||Y_t||_2||I(|Y_s| > \eta)||_2 \\
&\leq (\sum_{j=0}^{\infty} |\theta_j|||Z_{t-j}||_2)\sqrt{(Ee^{u|Y_s|})e^{-u\eta}} \\
&= ||Z_1||_2|\theta|_1\sqrt{(Ee^{u|Y_1|})}e^{-u\eta/2} \\
&\leq ||Z_1||_2|\theta|_1 C_u e^{-u\eta/2}
\end{aligned}$$

for some positive u and C_u such that $Ee^{u|Y_s|} \leq C_u^2 < \infty$. This is achieved for some small enough u since

$$\begin{aligned}
&Ee^{\pm uY_s} = \exp\{\sum_{j=0}^{\infty} \ln Ee^{\pm u\theta_j Z_1}\} = \exp\{\sum_{j=0}^{\infty} K(\pm u\theta_j)\} \\
&\leq \exp\{\sum_{j=0}^{\infty} \sup_{|v|\leq u|\theta|_1} |K'(v)||u\theta_j|\} \leq \exp\{\sup_{|v|\leq u|\theta|_1} |K'(v)|u|\theta|_1\}.
\end{aligned}$$

Then

$$Ee^{u|Y_s|} \le Ee^{uY_s} + Ee^{-uY_s} \le 2\exp\{\sup_{|v|\le u|\theta|_1} |K'(v)|u|\theta|_1\} \equiv C_u^2 < \infty$$

for some small enough u, due to continuous differentiability of $K(\cdot)$ at 0, which is assumed in Condition (B2). These lead to (6). Q.E.D.

Proof of Equation (7) Note that $EN_{it} = rE|Y_1| \le rE|Z_1||\theta|_1$, and

$$\begin{aligned}
&E(N_{it} + EN_{it})I(N_{it} + EN_{it} > \delta/(6\lambda)) \\
&\le E(|Y_{t-1}| + \ldots + |Y_{t-r}| + rE|Y_1|)I(|Y_{t-1}| + \ldots + |Y_{t-r}| > \delta/(6\lambda) - rE|Z_1||\theta|_1) \\
&\le E(|Y_{t-1}| + \ldots + |Y_{t-r}| + rE|Y_1|)\sum_{s=1}^{r} I(|Y_{t-s}| > \delta/(6r\lambda) - E|Z_1||\theta|_1) \\
&\le |||Y_{t-1}| + \ldots + |Y_{t-r}| + rE|Y_1|||_2||\sum_{s=1}^{r} I(|Y_{t-s}| > \delta/(6r\lambda) - E|Z_1||\theta|_1)||_2 \\
&\le (2r||Y_1||_2)r\sqrt{Ee^{u|Y_1|}}e^{-u\psi/2} \\
&\le (2r^2|\theta|_1||Z_1||_2)C_u e^{-u\psi/2}
\end{aligned}$$

for some small enough $u > 0$, where $\psi = \delta/(6r\lambda) - E|Z_1||\theta|_1$, and C_u is defined in Proof of Equation (6). Q.E.D.

Proof of Equation (9) We notice the process (y_t, x_t) can be approximated by strong mixing processes. This is because if we define $y_{t,k} = \sum_{j=0}^{k}\theta_j f_j(z_{t-j}, \varepsilon_{t-j})$, and $x_{t,k} = (1, y_{t-1,k}, \ldots, y_{t-r,k}, z_t')'$ then $||y_t - y_{t,k}||_1 \le \sum_{j>k}|\theta_j|\sup_{t,l}||f_l(z_t,\varepsilon_t)||_1$ and $||x_t - x_{t,k}||_1 \equiv E|x_t - x_{t,k}|_1 = \sum_{s=t-1}^{t-r}||y_s - y_{s,k}||_1 \le r\sum_{j>k}|\theta_j|\sup_{t,l}||f_l(z_t,\varepsilon_t)||_1$, both of which will decrease exponentially fast with k. On the other hand, $(y_{t,k}, x_{t,k})$ is a measurable transform of $(z_t, \ldots, z_{t-r-k}, \varepsilon_t, \ldots, \varepsilon_{t-r-k})$ and is therefore strong mixing with mixing coefficient α_{m-r-k} for $m > r+k$.

Now define $\rho_{t,k} = |I(y_{t,k} > 0) - I(x_{t,k}'b > 0)|$. The technique in Section 2.2 implies that $E|E(\rho_t|\mathcal{F}_{t-m}) - E\rho_t| \le E|E(\rho_{t,k}|\mathcal{F}_{t-m}) - E\rho_{t,k}| + 2E|\rho_t - \rho_{t,k}|$ where $\mathcal{F}_t$ represents the σ-field generated by $(z_s, \varepsilon_s)_{-\infty}^{t}$. The first term is bounded by $6\alpha_{m-r-k}$ by using Theorem 14.2, Davidson (1994).

Applying the triangular inequalities, the second term is at most $2||I(y_t > 0) - I(y_{t,k} > 0)||_1 + 2||I(x_t'b > 0) - I(x_{t,k}'b > 0)||_1$. We will assume that (†) $P[|y_t| \le \Delta] \le M_y(2\Delta)$ for any small $\Delta > 0$, for some constant $M_y < \infty$. This is true under Condition (C3).

We will also assume that (‡) $P[|x_t'b| \le \Delta] \le M_x(2\Delta)$ for any small $\Delta > 0$, for some constant $M_x < \infty$. Notice that $x_t'b$ is of the form "$\pm z_{t,1}+$ *a linear combination of* $y_{t-1}, \ldots, y_{t-r}, z_{t,2}, \ldots, z_{t,K}$" due to the standardization of the coefficient of $z_{t,1}$. Then it is obvious that (‡) holds under Condition (C4).

Now notice that for two random variables W and W^*,

$$||I(W > 0) - I(W^* > 0)||_1 \le 4\sqrt{2M}\sqrt{||W - W^*||_1}, \tag{13}$$

if $P(|W| \le \Delta) \le M(2\Delta)$ for $\Delta = \sqrt{||W - W^*||_1}/\sqrt{2M}$. This is proved by noting

$$EI(W > 0, W^* \le 0)$$

$$\begin{aligned}
&\le EI(W>0, W^*\le 0, |W-W^*|\le \Delta)+EI(|W-W^*|>\Delta)\\
&\le P[|W|\le \Delta]+E|W-W^*|/\Delta\\
&\le M(2\Delta)+E|W-W^*|/\Delta\\
&= 2\sqrt{2M}\sqrt{||W-W^*||_1}.
\end{aligned}$$

Similarly $EI(W^*>0, W\le 0)\le 2\sqrt{2M}\sqrt{||W-W^*||_1}$. Now $||I(W>0)-I(W^*>0)||_1 = EI(W>0, W^*\le 0)+EI(W^*>0, W\le 0)$ leads to (13).

Now apply (13) and we obtain

$$\begin{aligned}
&2||I(y_t>0)-I(y_{t,k}>0)||_1+2||I(x_t'b>0)-I(x_{t,k}'b>0)||_1\\
&\le 8\sqrt{2M_y}\sqrt{||y_t-y_{t,k}||_1}+8\sqrt{2M_x}\sqrt{||x_t'b-x_{t,k}'b||_1}\\
&\le 8\sqrt{2M_y}\sqrt{||y_t-y_{t,k}||_1}+8\sqrt{2M_x|b|_\infty}\sqrt{||x_t-x_{t,k}||_1}\\
&\le 8\sqrt{2M_y}\sqrt{\sum_{j>k}|\theta_j|\sup_{t,l}||f_l(z_t,\varepsilon_t)||_1}+8\sqrt{2M_xC_b}\sqrt{r\sum_{j>k}|\theta_j|\sup_{t,l}||f_l(z_t,\varepsilon_t)||_1},
\end{aligned}$$

and therefore we have (9). Q.E.D.

Proof of Equation (10) Note that the $B=B_+\cup B_-$ where $B_\pm=\{b\in B: b_{r+1}=\pm 1\}$ represents the two halfs of B with $b_{r+1}=\pm 1$, respectively. Note that $B_\pm$ can be written as a union $B_\pm=\cup_\gamma B_\pm(\gamma)$, where γ represents the subset of 'selected indices' in addition to $r+2$, and the union is taken over all subsets γ of $\{1,\ldots,K+r+1\}\backslash\{r+2\}$ with card$(\gamma)\le v-1$. Here $B_\pm(\gamma)=\{b\in\Re^{K+r+1}: b_{r+2}=\pm 1, |b_j|\le C_b, \forall j\in\gamma; |b_j|=0, \forall j\notin\gamma\cup\{r+2\}\}$. Each $B_\pm(\gamma)$ can be covered by at most $(2C_b/\lambda+1)^{\text{card}(\gamma)}$ sets B_i's of the form $B_i=\{b\in\Re^{K+r+1}: b_{r+2}=\pm 1, |b_j-(b_i)_j|\le\lambda/2, \forall j\in\gamma; |b_j-(b_i)_j|=0, \forall j\notin\gamma\cup\{r+2\}\}$ for some $b_i\in B_\pm(\gamma)$, where card$(\gamma)\le v-1$. So a combinatorial argument leads to upperbound (10). Q.E.D.

Proof of Assumption (A3) for the example in Section 5 Assume Condition (C5), so that $|z_t|_\infty\le C_z<\infty$.

We can apply the arguments in Remark 8 by identifying $w_t=z_{t,1}$, $v_t=(1,y_{t-1},\ldots,y_{t-r},z_{t,2},\ldots,z_{t,K})'$. Note that $d_0\le v$ and we can choose $\mathcal{V}_t$ as the σ-field generated by $\{z_s,\varepsilon_s\}_{-\infty}^{t-1}$ and $z_{t,2},\ldots,z_{t,K}$. Then A3 is satisfied if (C4) holds [the conditional density $p(w_t|\mathcal{V}_t)$ is bounded above by c (=M_x)] and if

$$\lambda\le\delta/(12vCc(1+C_z+r|\theta|_1\sup_{t,l}||f_l(z_t,\varepsilon_t)||_1)), \tag{14}$$

which would then be $\le\delta/(12vCc(1+C_z+r\sup_t||y_t||_1))\le\delta/(12d_0Cc\sup_t E|v_t|_\infty)$. This inequality (14) is satisfied for all large n, when we take $\delta=n^{-0.5+\gamma_1}/2$ for any $\gamma_1>0$, and $\lambda=n^{-1}$ (the ℓ_∞-diameter of B_i), and assume that the number of lags r and the maximal number of selected variables v follow condition (C6). Q.E.D.

Proof of Equation (11) Similar to the approximation argument used before in Proof of Equation (9), we define $\omega_{t,k}=(y_{t,k},x_{t,k})$ and obtain $E|E(I(\omega_t\in S_i)|\mathcal{F}_{t-m})-EI(\omega_t\in S_i)|\le E|E(I(\omega_{t,k}\in S_i)|\mathcal{F}_{t-m})-EI(\omega_{t,k}\in S_i)|+2E|I(\omega_t\in S_i)-I(\omega_{t,k}\in S_i)|$, where the first term is bounded above by

$6\alpha_{m-r-k}$ for $m > r+k$, due to a treatment similar to the one used in Proof of Equation (9). Now we try to bound $E|I(\omega_t \in S_i) - I(\omega_{t,k} \in S_i)| \equiv ED$ in the second term, where $S_i \equiv \cup_{b\in B_i}[x_t'b = 0]$ will be computed below.

For our vector $x_t = (1, y_{t-1}, ..., y_{t-r}, z_{t,1}, z_{t,2}, ..., z_{t,K})'$, we here identify $w_t = z_{t,1} = (x_t)_{r+2}$, and $v_t = (1, y_{t-1}, ..., y_{t-r}, z_{t,2}, ..., z_{t,K})'$ including all $\{(x_t)_j\}_{j\neq r+2}$, for applying Remarks 8 and 9. Note that according to Remark 9, $S_i = [w^-(v_t) \le w_t \le w^+(v_t)]$ where $w^\pm(v_t) = v_t'(b_i)_v \pm h|v_t|_1$. [We have picked a sign $b_{r+2} = -1$ to proceed. The other sign is similar. To be more precise, here $|v_t|_1 = \sum_{j\in\gamma}|(x_t)_j|$ where γ is a subset of ‘selected indices’ in addition to $r+2$, when B_i is as described in Proof of Equation (10).]

Now $D \equiv |I(\omega_t \in S_i) - I(\omega_{t,k} \in S_i)| = 1$ implies that only one of the two points $\{x_t, x_{t,k}\} = \{(v_t', w_t)', \ (v_{t,k}', w_{t,k})'\}$ can lie in S_i, so there must be an intermediate point lying on a boundary of S_i, either on the upper boundary and denoted as $x^+ = ((v^+)', w^+(v^+))'$, or on the lower boundary and denoted as $x^- = ((v^-)', w^-(v^-))'$. [Here we have re-ordered the components of the vectors. For examples, for vector x_t, its component $(x_t)_{r+2} = w_t$ is now placed behind other components (denoted as v_t).]

If in addition, we also have a small distance $|x_t - x_{t,k}|_1 \le \eta$, then one of the following two events must happen (with ± option depending on whether the intermediate point falls on the upper or lower boundary of S_i):

$$\begin{aligned}
&|w_t - w^\pm(v_t)| \\
&\le |w_t - w^\pm(v^\pm)| + |w^\pm(v^\pm) - w^\pm(v_t)| \\
&\le |w_t - w^\pm(v^\pm)| + (C_b + h)|v_t - v^\pm|_1 \\
&\le |(1 + C_b + h)|x_t - x^\pm|_1 \\
&\le (1 + C_b + h)|x_t - x_{t,k}|_1 \le (1 + C_b + h)\eta.
\end{aligned}$$

Now we can bound

$$\begin{aligned}
ED &\le P[D = 1, |x_t - x_{t,k}|_1 \le \eta] + P[|x_t - x_{t,k}|_1 > \eta] \\
&\le P[\cup|w_t - w^\pm(v_t)| \le (1 + C_b + h)\eta] + P[|x_t - x_{t,k}|_1 > \eta] \\
&\overset{(a)}{\le} 2c2(1 + C_b + h)\eta + E|x_t - x_{t,k}|_1/\eta \\
&\le 2c2(1 + C_b + h)\eta + r\sum_{j>k}|\theta_j|\sup_{t,l}||f_l(z_t, \varepsilon_t)||_1/\eta.
\end{aligned}$$

Now take $\eta = \sqrt{r\sum_{j>k}|\theta_j|\sup_{t,l}||f_l(z_t,\varepsilon_t)||_1/(2c2(1+C_b+h))}$ and obtain

$$ED \le 2\sqrt{r\sum_{j>k}|\theta_j|\sup_{t,l}||f_l(z_t, \varepsilon_t)||_1(2c2(1 + C_b + h))}.$$

As in the previous Proof of Assumption (A3), we identify $c = M_x$. We have assumed Condition (C4) for the inequality (a) above. Therefore we have (11). Q.E.D.

References

P. L. Bartlett and G. Lugosi. An inequality for uniform deviations of sample averages from their means. *Statistics and Probability Letters*, 44:55-62, 1999.

B. Bercu, E. Gassiat, and E. Rio. Concentration inequalities, large and moderate deviations for self-normalized empirical processes. *Annals of Probability*, 30:1576-1604, 2002.

O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. *Advanced Lectures on Machine Learning 2003* (O. Bousquet, U. von Luxburg, G. Rätsch, eds.) Springer, Berlin, 169-207, 2003.

J. Davidson. *Stochastic Limit Theory.* Oxford University Press, Oxford, 1994.

R. M. de Jong and T. M. Woutersen. Dynamic time series binary choice. *Manuscript, Ohio State University,* 2004. Downloadable at `http://www.econ.ohio-state.edu/dejong/tiemen45.pdf`.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer, New York, 1996.

C. Francq and J.-M. Zakoïan. Mixing properties of a general class of GARCH(1,1) models without moment assumptions on the observed process. *Econometric Theory*, 22:815-834, 2006.

D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78-150, 1992.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13-30, 1963.

J. L. Horowitz. A smoothed maximum score estimator for the binary response model. *Econometrica*, 60:505-531, 1992.

W. Jiang and M. A. Tanner. Risk minimization for time series binary choice with variable selection. *Technical Report 07-02, Department of Statistics, Northwestern University*, 2007. Downloadable at `http://newton.stats.northwestern.edu/~jiang/tr/choice1.tr.pdf`.

W. Jiang and M. A. Tanner. Gibbs posterior for variable selection in high dimensional classification and data mining. *Annals of Statistics*, 36:2207-2231. 2008. Downloadable at `http://newton.stats.northwestern.edu/~jiang/tr/gibbsone2.tr.pdf`.

A. C. Lozano, S. R. Kulkarni, and R. E. Schapire. Convergence and consistency of regularized boosting algorithms with stationary beta-mixing observations. In *Advances in Neural Information Processing Systems* 18, 2006. Downloadable at `http://www.cs.princeton.edu/~schapire/boost.html`.

C. McDiarmid. Concentration. *Probabilistic Methods for Algorithmic Discrete Mathematics* (M. Habib, C. McDiarmid, J. Ramirez, B. Reed, eds.) 195–248, Springer, Berlin, 1998.

R. Meir and T. Zhang. Generalization error bounds for Bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839-860, 2003.

W. K. Newey. Uniform convergence in probability and stochastic equicontinuity. *Econometrica*, 59:1161-1167, 1991.

D. Pollard. Uniform ratio limit theorems for empirical processes. *Scandinavian Journal of Statistics*, 22:271-278, 1995.

S. Van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, Cambridge, 2000.

M. Vidyasagar. Convergence of empirical means with alpha-mixing input sequences, and an application to PAC learning. *Proceedings of the 44th IEEE Conference on Decision and Control, and the European Control Conference 2005*, pages 560-565, 2005.

B. Yu. Rates of convergence for empirical processes of stationary mixing sequences. *Annals of Probability*, 22:94-114, 1994.

T. Zhang. Information theoretical upper and lower bounds for statistical estimation. *IEEE Transaction on Information Theory*, 52:1307- 1321, 2006.

B. Zou and L. Li. The performance bounds of learning machines based on exponentially strong mixing sequences, *Computers and Mathematics with Applications*, 53:1050-1058, 2007.