Risk bounds for time series without strong mixing

Daniel J. McDonald
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
danielmc@stat.cmu.edu

Cosma Rohilla Shalizi
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
cshalizi@stat.cmu.edu

Mark Schervish
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
mark@cmu.edu

Version: June 2, 2011

Abstract

We show how to control the generalization error of time series models wherein past values of the outcome are used to predict future values. The results are based on a generalization of standard IID concentration inequalities to dependent data. We show how these concentration inequalities behave under different versions of dependence to provide some intuition for our methods.

1 Introduction

Much of the literature in machine learning focuses on studying the behavior of predictions constructed based on a training set $(X_1, Y_1), \ldots, (X_n, Y_n)$ where one wishes to construct a mapping from X to Y. This training set may consist of n IID draws from a common distribution, or it may have some dependence property such as ergodicity or mixing behavior [8, 4, 7]. It may even be generated by an adversary intent on deceiving us about the relationship [2, 10].

Time series data are different. We observe only a single sequence of random variables $\mathbf{Y}_1^n = (Y_1, \ldots, Y_n)$ taking values in a measurable space $\mathcal{Y}$ and wish to learn a function which takes the past observations as inputs and predicts the future. Suppose, given data from time 1 to time n, we wish to predict time $n + h$ for some $h \in \mathbb{N}$. Then for some loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$, and some predictor $g : \mathcal{Y}^n \to \mathcal{Y}$, we define the *prediction risk*, or *generalization error*, as

$$R(g) := \mathbb{E}[\ell(Y_{n+h}, g(\mathbf{Y}_1^n)]. \tag{1}$$

Here we assume that the data series is stationary, a notion to be defined more precisely later. But this allows us to have some hope of controlling the generalization error defined in (1). Absent this sort of behavior, the past and future could be unrelated.

Since the true distribution is unknown, so is $R(g)$, but we can attempt to estimate it based on only our observed data. In situations with predictors X and responses Y, there is

the obvious estimator

$$\widetilde{R}_n(g) := \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, g(X_i)).$$

However, in this case, we may use some or all of the past to generate predictions, and similarly, it may be that we have not observed Y_{i+h} for some i. To ease notation for the remainder of the paper, assume that we have observed some sequence of data $Y_1, \ldots, Y_{n+j}$ for $j \in \mathbb{N}$ such that it is possible to evaluate the quantity $\ell(Y_{i+h}, g(Y_1, \ldots, Y_i))$ for each $i \in \{1, \ldots, n\}$. For time series prediction, we define the *training error* as

$$\widehat{R}_n(g) := \frac{1}{n}\sum_{i=1}^{n} \ell(Y_{i+h}, g(\mathbf{Y}_1^i). \tag{2}$$

Here g is some function chosen out of a class of possible functions $\mathcal{G}$.

Choosing a particular prediction function $\widehat{g}$ as the minimizer of $\widehat{R}_n$ over $\mathcal{G}$ is "empirical risk minimization" (ERM); this often gives poor results because the choice of $\widehat{g}$ adapts to the training data, causing the training error to be an over-optimistic estimate of the true risk. Additionally, training error must shrink as model complexity grows so that ERM will tend to overfit the data and give poor out-of-sample predictions.

While $\widehat{R}_n(\widehat{g})$ converges to $R(\widehat{g})$ for many algorithms, one can show that when $\widehat{g}$ minimizes (2), $\mathbb{E}[\widehat{R}_n(\widehat{g})] \leq R(\widehat{g})$. There are a number of ways to mitigate this issue. The first is to restrict the class $\mathcal{G}$. The second is to change the optimization problem, penalizing model complexity. Rather than attempting to estimate $R(g)$, we provide bounds on it which hold with high probability across all possible prediction functions $g \in \mathcal{G}$. A typical result in this literature is a confidence bound on the risk which says that with probability at least $1 - \delta$,

$$R(\widehat{g}) \leq \widehat{R}_n(\widehat{g}) + \Gamma(C(\mathcal{G}), n, \delta),$$

where $C(\cdot)$ measures the complexity of the model class $\mathcal{G}$, and $\Gamma(\cdot)$ is a function of the complexity, the confidence level, and the number of observed data points.

In §2, we provide some background material necessary to characterize our results, including some concentration inequalities for dependent data. Section 3 derives risk bounds for time series and gives a novel proof that the standard Rademacher complexity characterizes the flexibility of $\mathcal{G}$. Section 4 supplies some straightforward examples showing how dependence affects the quality of bounds. Section 5 concludes and provides some ideas about the future of these results.

2 Time series, complexity, and concentration of measure

In this section, we introduce some of the math necessary to develop our results: stationarity is a prerequisite for control of generalization error; Rademacher complexity measures the flexibility of the model space $\mathcal{G}$; dependence modifies concentration inequalities.

Throughout what follows, $\mathbf{Y} = \{Y_t\}_{t=-\infty}^{\infty}$ will be a sequence of random variables, i.e., each Y_t is a measurable mapping from some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ into a measurable space $\mathcal{Y}$. A block of the random sequence will be written $\mathbf{Y}_i^j \equiv \{Y_t\}_{t=i}^{j}$, where either limit may go to infinity. The σ-field generated by a particular block $\mathbf{Y}_i^j$ will be given by $\mathcal{F}_i^j$.

2.1 Time series

The dependent data setting we investigate is based on stationary time series input data. We first remind the reader of the notion of (strict or strong) stationarity.

Definition 2.1 (Stationarity)**.** *A sequence of random variables* $\mathbf{Y}$ *is stationary when all its finite-dimensional distributions are invariant over time: for all* t *and all non-negative integers* i *and* j*, the random vectors* $\mathbf{Y}_t^{t+i}$ *and* $\mathbf{Y}_{t+j}^{t+i+j}$ *have the same distribution.*

Stationarity does not imply that the random variables Y_t are independent across time t, only that the distribution of Y_t is constant over time.

2.2 Rademacher complexity

Statistical learning theory provides several ways of measuring the complexity of a class of predictive models. The results we use rely on Rademacher complexity (see, e.g., [1]), which measures how well the model can (seem to) fit white noise.

Definition 2.2 (Rademacher Complexity)**.** *Let* $\mathbf{Y}_1^n$ *be a time series drawn according to a joint distribution* ν*. The* empirical Rademacher complexity *is*

$$\widehat{\mathfrak{R}}_n(\mathcal{G}) := 2\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(\mathbf{Y}_1^i)\right| \mid \mathbf{Y}_1^n\right],$$

where σ_i *are a sequence of random variables, independent of each other and everything else, and equal to* $+1$ *or* -1 *with equal probability. The* Rademacher complexity *is*

$$\mathfrak{R}_n(\mathcal{G}) := \mathbb{E}_\nu\left[\widehat{\mathfrak{R}}_n(\mathcal{G})\right]$$

where the expectation is over sample paths $\mathbf{Y}_1^n$ *generated by* ν*.*

The term inside the supremum, $\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(\mathbf{Y}_1^i)\right|$, is the sample covariance between the noise σ and the predictions of a particular model g. The Rademacher complexity takes the largest value of this sample covariance over all models in the class (mimicking empirical risk minimization), then averages over realizations of the noise.

Intuitively, Rademacher complexity measures how well our models could seem to fit outcomes which were really just noise, giving a baseline against which to assess the risk of over-fitting or failing to generalize. As the sample size n grows, for any given g the sample covariance $\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(\mathbf{Y}_1^i)\right| \to 0$, by the ergodic theorem; the overall Rademacher complexity should also shrink, though more slowly, unless the model class is so flexible that it can fit absolutely anything, in which case one can conclude nothing about how well it will predict in the future from the fact that it performed well in the past.

2.3 Concentration inequalities

For IID data, the main tools for developing risk bounds are the inequalities of Hoeffding [3] and McDiarmid [6]. Instead, we will use dependent versions of each which generalize the IID results. These inequalities are derived in van de Geer [12]. They rely on constructing predictable bounds for random variables based on past behavior, rather than assuming *a priori* knowledge of the distribution.

Theorem 2.3 (van de Geer [12] Theorem 2.5)**.** *Consider a random sequence* $\mathbf{Y}_1^n$ *where*

$$L_i \leq Y_i \leq U_i \text{ a.s. for all } i \geq 1,$$

where $L_i < U_i$ *are* $\mathcal{F}_1^{i-1}$*-measurable random variables,* $i \geq 1$*. Define*

$$C_n^2 = \sum_{i=1}^{n} (U_i - L_i)^2,$$

with the convention $C_0^2 = 0$*. Then for all* $\epsilon > 0$*,* $c > 0$*,*

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i \geq \epsilon \text{ and } C_n^2 \leq c^2 \text{ for some } n \right) \leq \exp\left\{ -\frac{2\epsilon^2}{c^2} \right\}.$$

Of course if L_i and U_i are non-random, this returns the usual Hoeffding inequality. Here however, they must only be forecastable given past values of the random sequence.

Theorem 2.4 (van de Geer [12] Theorem 2.6)**.** *Fix* $n \geq 1$*. Let* Z_n *be* $\mathcal{F}_1^n$*-measurable such that*

$$L_i \leq \mathbb{E}[Z_n \mid \mathcal{F}_1^i] \leq U_i, \text{ a.s.}$$

where $L_i < U_i$ *are* $\mathcal{F}_1^{i-1}$*-measurable. Define* C_n^2 *as above. Then for all* $\epsilon > 0$*,* $c > 0$*,*

$$\mathbb{P}\left(Z_n - \mathbb{E}[Z_n] \geq \epsilon \text{ and } C_n^2 \leq c^2 \right) \leq \exp\left\{ -\frac{2\epsilon^2}{c^2} \right\}.$$

To see how this generalizes McDiarmid's inequality, we provide the following corollary.

Corollary 2.5. *Let* $g(Y_1, \ldots, Y_n)$ *be some real valued function on* $\mathcal{Y}^n$ *such that*

$$\left| \mathbb{E}[g(Y_1, \ldots, Y_n) \mid \mathcal{F}_1^i] - \mathbb{E}[g(Y_1, \ldots, Y_n) \mid \mathcal{F}_1^{i-1}] \right| \leq k_i \tag{3}$$

where k_i *is* $\mathcal{F}_1^{i-1}$*-measurable. Then,*

$$\mathbb{P}\left(g(Y_1, \ldots, Y_n) - \mathbb{E}[g(Y_1, \ldots, Y_n)] > \epsilon \text{ and } \sum_i k_i^2 < c^2 \right) < \exp\left\{ -\frac{2\epsilon^2}{c^2} \right\}.$$

In particular, this gives a couple of immediate consequences. Suppose that g is bounded. Then, we have that

$$k_i \leq \sup_{\mathbf{Y}_i^n} \sup_{\mathbf{Y}_i^{n\prime}} |g(Y_1, \ldots, Y_{i-1}, Y_i, \ldots, Y_n) - g(Y_1, \ldots, Y_{i-1}, Y_i', \ldots, Y_n')| = b_i.$$

This contrasts with the bounded differences inequality in the IID case, wherein one only needs to be concerned with one point that is different. For IID data, we have starting from (3),

$$k_i \leq \sup_{Y_i, Y_i'} |g(Y_1, \ldots, Y_{i-1}, Y_i, \ldots, Y_n) - g(Y_1, \ldots, Y_{i-1}, Y_i', \ldots, Y_n)| = d_i,$$

if g satisfies bounded differences with constants d_i. In other words, Theorem 2.4 conflates dependence with nice functional behavior.

3 Risk bounds

Generalization error bounds follow from deriving high probability upper bounds on the quantity

$$Q_n(\mathcal{H}) := \sup_{h \in \mathcal{H}} \big(R(h) - \widehat{R}_n(h) \big),$$

which is the worst case difference between the true risk $R(h)$ and the empirical risk $\widehat{R}_n(h)$ over all functions in the class of losses $\mathcal{H} = \{h = \ell(\cdot, g(\cdot)) : g \in \mathcal{G}\}$ defined over a particular class of prediction functions $\mathcal{G}$. In the case of time series, $Q_n(h)$ is $\mathcal{F}_n$-measurable, so we can get risk bounds from Theorem 2.4 if we can find suitable L_i and U_i sequences.

Theorem 3.1. *Suppose that $Q_n(\mathcal{H})$ satisfies the forecastable boundedness condition of Theorem 2.4. Then,*

$$\mathbb{P}\left(R(h) < \widehat{R}_n(h) + \mathbb{E}[Q_n(\mathcal{H})] + c\sqrt{\frac{\log 1/\delta}{2}} \quad \textit{or} \quad C_n^2 > c \right) \leq 1 - \delta.$$

In many cases (as in the examples below), C_n^2 will be deterministic, in which case, the result above is greatly simplified. Essentially, the theorem says that as long as each new Y_i gives us additional control on the conditional expectation of Q_n, we can ensure that with high probability, our forecasts of the future will have only small losses. The proof is straightforward: simply set the right hand side of Theorem 2.4 to δ and use DeMorgan's law.

Since $\mathbb{E}[Q_n(\mathcal{H})]$ is a complicated and unintuitive object, we upper bound it with the Rademacher complexity. The standard symmetrization argument for the IID case does not work, but, for time series prediction (as opposed to the more general dependent data case or the online learning case), Rademacher bounds are still available. We provide this result now.

Theorem 3.2. *For a time series prediction problem based on a sequence $\mathbf{Y}_1^n$,*

$$\mathbb{E}[Q_n(\mathcal{H})] \leq \mathfrak{R}_n(\mathcal{H}). \tag{4}$$

The standard way of proving this result in the IID case is through introduction of a "ghost sample" $\widetilde{\mathbf{Y}}_1^n$ which has the same distribution as $\mathbf{Y}_1^n$. Taking empirical expectations over the ghost sample is then the same as taking expectations with respect to the distribution of $\mathbf{Y}_1^n$. Randomly exchanging Y_i with $\widetilde{Y}_i$ by using Rademacher variables allows for control of $\mathbb{E}[Q_n(\mathcal{H})]$ and leads to the factor of 2 in Definition 2.2. However, in the dependent data setting, this is not quite so easy.

For dependent data, both the ghost sample and the introduction of Rademacher variables arise differently. A similar situation also occurs in the more complex cases of online learning with a (perhaps constrained) adversary choosing the data sequence. It is covered in depth in Rakhlin et al. [10, 11]. With dependent data we need a different version of the "ghost sample" than that used in the IID case. First, we rewrite the left side of (4):

$$\begin{aligned}
\mathbb{E}_{\mathbf{Y}}[Q_n(\mathcal{H})] &= \mathbb{E}_{\mathbf{Y}}\left[\sup_{h \in \mathcal{H}} \Big(R_n(h) - \widehat{R}_n(h) \Big) \right] \\
&= \mathbb{E}_{\mathbf{Y}}\left[\sup_{g \in \mathcal{G}} \left(\mathbb{E}_{Y_{n+h}}[\ell(Y_{n+h}, g(\mathbf{Y}_1^n))] - \frac{1}{n} \sum_{i=1}^{n} \ell(Y_{i+1}, g(\mathbf{Y}_1^i)) \right) \right]. \qquad (5)
\end{aligned}$$

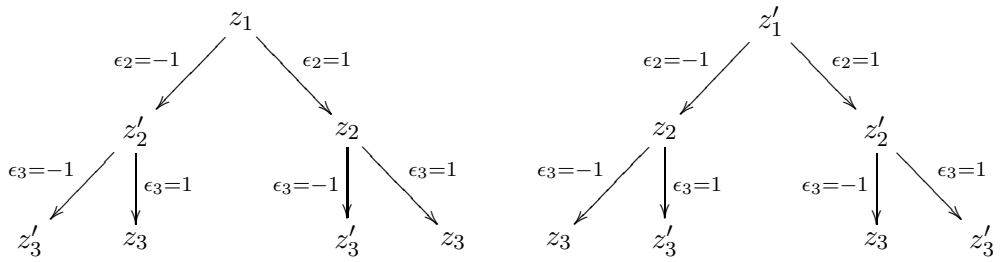

Figure 1: This figure displays the tree structures for $\mathbf{Z}(\boldsymbol{\sigma})$ and $\mathbf{Z}'(\boldsymbol{\sigma})$. The path along each tree is determined by one $\boldsymbol{\epsilon}$ sequence, interleaving the "past" between paths.

Here, we define $z_i = (Y_{i+h}, \mathbf{Y}_1^i)$ so that $h(z_i) = \ell(Y_{i+h}, g(\mathbf{Y}_1^i))$ for some $g \in \mathcal{G}$. At this point, following [10, 11], we introduce a "tangent sequence" $\mathbf{Z}'$ rather than the ghost sample. We construct it recursively as follows. Let,

$$\mathcal{L}(Y_1') = \mathcal{L}(Y_1) \qquad \text{and} \qquad \mathcal{L}(Y_i'|Y_1, \ldots, Y_{i-1}) = \mathcal{L}(Y_i|Y_1, \ldots, Y_{i-1}),$$

where $\mathcal{L}$ denotes the probability law. Then, let $\mathbf{Z} = (z_1, \ldots, z_n)$ and $\mathbf{Z}' = (z_1', \ldots, z_n')$.

Proof of Theorem 3.2. Starting from (5) we have

$$\begin{aligned}
\mathbb{E}[Q_n(\mathcal{H})] &= \mathbb{E}_{\mathbf{Z}}\left[\sup_{h\in\mathcal{H}}\left(\mathbb{E}_{\mathbf{Z}}\left[\frac{1}{n}\sum_{i=1}^n h(z_i)\right] - \frac{1}{n}\sum_{i=1}^n h(z_i)\right)\right] \\
&= \mathbb{E}_{\mathbf{Z}}\left[\sup_{h\in\mathcal{H}}\left(\mathbb{E}_{\mathbf{Z}'}\left[\frac{1}{n}\sum_{i=1}^n h(z_i')\right] - \frac{1}{n}\sum_{i=1}^n h(z_i)\right)\right]. \qquad (6)
\end{aligned}$$

Here we have constructed $\mathbf{Z}'$ as a tangent sequence to $\mathbf{Z}$ as discussed above. Then,

$$\begin{aligned}
(6) &\le \mathbb{E}_{\mathbf{Z},\mathbf{Z}'}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n h(z_i') - h(z_i)\right] \qquad \text{(Jensen)} \\
&= \mathbb{E}_{\substack{z_1 \\ z_1'}}\mathbb{E}_{\substack{z_2|z_1 \\ z_2'|z_1'}}\cdots\mathbb{E}_{\substack{z_n|z_{n-1},\ldots,z_1 \\ z_n'|z_{n-1}',\ldots,z_1'}}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n h(z_i') - h(z_i)\right] \qquad (7)
\end{aligned}$$

Now, due to dependence, Rademacher variables must be introduced carefully as in the adversarial case. Rademacher variables create two tree structures, one associated to the $\mathbf{Z}$ sequence, and one associated to the $\mathbf{Z}'$ sequence (see [10, 11] for a thorough treatment). We write these trees as $\mathbf{Z}(\boldsymbol{\sigma})$ and $\mathbf{Z}'(\boldsymbol{\sigma})$, where $\boldsymbol{\sigma}$ is a particular sequence of Rademacher variables (e.g. $(1, -1, -1, 1, \ldots, 1)$) which creates a path along each tree. For example, consider $\boldsymbol{\sigma} = \mathbf{1}$. Then, $\mathbf{Z}(\boldsymbol{\sigma}) = (z_1, \ldots, z_n)$ and $\mathbf{Z}'(\boldsymbol{\sigma}) = (z_1', \ldots, z_n')$, the "right" path of both tree structures. For $\boldsymbol{\sigma} = -\mathbf{1}$. Then, $\mathbf{Z}(\boldsymbol{\sigma}) = (z_1', \ldots, z_n')$ and $\mathbf{Z}'(\boldsymbol{\sigma}) = (z_1, \ldots, z_n)$, the "left" path of both tree structures. Changing ϵ_i from $+1$ to -1 exchanges z_i for z_i' in both trees and chooses the left child of z_{i-1} and z_{i-1}' rather than the right child. Figure 1 displays both trees. In order to talk about the probability of z_i conditional on the "past" in the tree, we need to know the path taken so far. For this, we define a selector function

$$\chi(\sigma) := \chi(\sigma, \rho, \varrho) = \begin{cases} \rho & \sigma = 1 \\ \varrho & \sigma = -1. \end{cases}$$

Distributions over these trees then become the objects of interest.

In the time series case, as opposed to the online learning scenario, the dependence between future and past means the adversary is *not* free to change predictors and responses separately. Once a branch of the tree is chosen, the distribution of future data points is fixed, and depends only on the preceding sequence. Because of this, the joint distribution of any path along the tree is the same as any other path, i.e. for any two paths $\boldsymbol{\sigma}, \boldsymbol{\sigma'}$

$$\mathcal{L}(\mathbf{Z}(\boldsymbol{\sigma})) = \mathcal{L}(\mathbf{Z}(\boldsymbol{\sigma'})) \qquad \text{and} \qquad \mathcal{L}(\mathbf{Z}'(\boldsymbol{\sigma})) = \mathcal{L}(\mathbf{Z}'(\boldsymbol{\sigma'})).$$

Similarly, due to the construction of the tangent sequence, we have that $\mathcal{L}(\mathbf{Z}(\boldsymbol{\sigma})) = \mathcal{L}(\mathbf{Z}'(\boldsymbol{\sigma}))$. This equivalence between paths allows us to introduce Rademacher variables swapping z_i for z_i' as well as the ability to combine terms below:

$$\begin{aligned}
(7) &= \mathbb{E}_{\substack{z_1 \\ z_1'}} \mathbb{E}_{\sigma_1} \mathbb{E}_{\substack{z_2 | \chi(\sigma_1, z_1, z_1') \\ z_2' | \chi(\sigma_1, z_1', z_1)}} \mathbb{E}_{\sigma_2} \cdots \mathbb{E}_{\substack{z_n | \chi(\sigma_{n-1}), \ldots, \chi(\sigma_1) \\ z_n' | \chi(\sigma_{n-1}), \ldots, \chi(\sigma_1)}} \mathbb{E}_{\sigma_n} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i (h(z_i') - h(z_i)) \right] \\
&= \mathbb{E}_{\mathbf{Z}, \mathbf{Z}', \sigma} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i (h(z_i') - h(z_i)) \right] \\
&\leq \mathbb{E}_{\mathbf{Z}, \sigma} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(z_i) \right] + \mathbb{E}_{\mathbf{Z}', \sigma} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(z_i') \right] \\
&= 2 \mathbb{E}_{\mathbf{Z}, \sigma} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(z_i) \right] \\
&= \mathfrak{R}_n(\mathcal{H}).
\end{aligned}$$

□

Good control of $\mathbb{E}[Q_n(\mathcal{H})]$ through the Rademacher complexity therefore implies good control of the generalization error. Rademacher complexity is easy to handle for wide ranges of learning algorithms using results in [1] and elsewhere. Support vector machines, kernel methods, and neural networks all have known Rademacher complexities. Furthermore, Lipschitz composition arguments in [5] allow us to deal only with the Rademacher complexity of the function class $\mathcal{G}$ rather than the induced loss class $\mathcal{H}$. For loss functions ℓ which are ϕ-Lipschitz in their second argument, $\mathfrak{R}(\mathcal{H}) \leq 2\phi \mathfrak{R}(\mathcal{G})$.

The main issue then in the application of Theorem 3.1 is the determination of the forecastable bounds L_i and U_i from the data generating process. In the next section, we provide a few simple examples to aid intuition.

4 Examples

We consider three different examples which should aid the reader in understanding the nature of the forecastable bounds. Here we present two extreme cases — independence and complete dependence — as well as an intermediate case. It is important to note that C_n^2 is deterministic in all three cases, though this need not be the case.

4.1 Independence

For IID data, we simply recover IID concentration results. As noted in Corollary 2.5, for IID data, bounded differences yields good control. Similarly, Theorem 2.3 gives the same results as Hoeffding's inequality for IID data. Dependence is more interesting.

4.2 Complete dependence

Let $\mathbf{Y}_1^n$ be generated as follows:

$$Y_1 \sim U(a,b), \quad b > a \qquad\qquad Y_i = Y_{i-1}, \quad i \geq 2.$$

Consider trying to predict the mean $\frac{1}{n}\sum_{i=1}^n Y_i$. Then, given no observations, the almost sure upper bound $U_1 = b$ while the lower bound $L_1 = a$. So $(U_1 - L_1)^2 = (b-a)^2$. For $i > 1$, conditional on $\mathcal{F}_1^{i-1}$ (and therefore $\mathcal{F}_1$), $U_i = L_i$. Thus, $C_n^2 = (b-a)^2$ giving the entirely useless result:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n Y_i - (b+a)/2 \geq \epsilon\right) < \exp\left\{-\frac{2\epsilon^2}{(b-a)^2}\right\}.$$

The right side is independent of n implying that we essentially observed one data point regardless of n.

4.3 Partial dependence

Let $\mathbf{Y}_1^n$ be generated as follows:

$$Y_0 = 0, \qquad\qquad Y_i = \theta Y_{i-1} + \eta_i \quad i \geq 2,$$

where $\theta \in (0,1)$ and $\eta_i \overset{iid}{\sim} U(a,b)$ with $b > a$. Again, consider trying to predict the mean $\frac{1}{n}\sum_{i=1}^n Y_i$. We can define L_i and U_i as follows:

$$L_i = \frac{a}{n}\frac{1-\theta^{n-i}}{1-\theta} + \frac{1}{n}\sum_{k=1}^{i-1} Y_k + \theta Y_{i-1}, \qquad U_i = \frac{b}{n}\frac{1-\theta^{n-i}}{1-\theta} + \frac{1}{n}\sum_{k=1}^{i-1} Y_k + \theta Y_{i-1}.$$

From this, we have that

$$\begin{aligned} C_n^2 &= \sum_{i=1}^n \frac{(b-a)^2}{n^2(1-\theta)^2}\left(1-\theta^{n-i}\right)^2 \\ &= \frac{(b-a)^2}{n^2(1-\theta)^2(\theta^2-1)}\left(\theta^{2n} - 2\theta^{n+1} - 2\theta^n + n\theta^2 + 2\theta - n + 1\right) \\ &< \frac{(b-a)^2}{n(1-\theta)^2}. \end{aligned}$$

Therefore, by Theorem 2.4,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n Y_i - (b+a)/2 > \epsilon\right) < \exp\left\{-\frac{2n\epsilon^2(1-\theta)^2}{(b-a)^2}\right\}.$$

For comparison, if everything was IID, Hoeffding's inequality gives

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n Y_i - (b+a)/2 > \epsilon\right) < \exp\left\{-\frac{2n\epsilon^2}{(b-a)^2}\right\}.$$

Therefore, the dependence in $\mathbf{Y}_1^n$ reduces the effective sample size by $(1-\theta)^2$. If $\theta = 1/2$, then each additional datapoint decreases the probability of a bad event by only a $1/4$ relative to the IID scenario.

5 Discussion

In this paper, we have demonstrated how to control the generalization of time series prediction algorithms. These methods use some or all of the observed past to predict future values of the same series. In order to handle the complicated Rademacher complexity bound for the expectation, we have followed the approach used in the online learning case pioneered by Rakhlin et al. [10, 11], but we show that in our particular case, much of the structure needed to deal with the adversary is unnecessary. This results in clean risk bounds which have a form similar to the IID case.

The main issue with risk bounds for dependent data is that they rely on complete knowledge of the dependence for application. This is certainly true in our case in that we need to *know* how to choose U_i and L_i such that we almost surely control $\mathbb{E}[Q_n(\mathcal{H})]$. For the standard case of bounded loss, there are trivial bounds, but these will not give the necessary dependence on n which would imply learnability of good predictors. More knowledge of the dependence structure of the process is required, though this is in some sense undesirable. However, previous results in the dependent data setting, such as those presented in [8, 4, 7, 9], also have this requirement.[1] They rely on precise knowledge of the mixing behavior of the data which is unavailable. At the same time, mixing characterizations are often unintuitive conditions based on infinite dimensional joint distributions. Our version depends only on the ability to forecastably bound expectations given increasing amounts of data.

References

[1] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[2] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge Univ Press, Cambridge, UK, 2006.

[3] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. ISSN 0162-1459.

[4] R. L. Karandikar and M. Vidyasagar. Probably approximately correct learning with beta-mixing input sequences. submitted for publication, 2009.

[5] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. A Series of Modern Surveys in Mathematics. Springer Verlag, Berlin, 1991. ISBN 3540520139.

[6] C. McDiarmid. On the method of bounded differences. In J. Siemons, editor, *Surveys in Combinatorics*, volume 141 of *London Mathematical Society Lecture Note Series*, pages 148–188. Cambridge University Press, 1989.

[7] Ron Meir. Nonparametric time series prediction through adaptive model selection. *Machine Learning*, 39(1):5–34, 2000. URL `http://www.ee.technion.ac.il/~rmeir/Publications/MeirTimeSeries00.pdf`.

[1] IID results have an even more onerous requirement: we must be able to rule out any dependence at all.

[8] Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-iid processes. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, volume 21, pages 1097–1104, 2009.

[9] Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary φ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 11:789–814, February 2010.

[10] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Random averages, combinatorial parameters, and learnability. Technical report, arXiv, 2010. URL `http://arxiv.org/abs/1006.1138v1`.

[11] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Stochastic and constrained adversaries. Technical report, arXiv, 2011. URL `http://arxiv.org/abs/1104.5070`.

[12] Sara van de Geer. On hoeffding's inequality for dependent random variables. In Herold Dehling, Thomas Mikosch, and Michael Sørensen, editors, *Empirical Process Techniques for Dependent Data*, pages 161–169. Birkhäuser, Boston, 2002. URL `http://stat.ethz.ch/~geer/hoeffding2.pdf`.