Statistical learning under a non-iid data generating process

Ayoub El Hanchi

December 8, 2018

1 Introduction

Probably the most important assumption used in the derivation of learning bounds and in the design of algorithms in the statistical learning framework is that of independence and stationarity of the data generating process. Many learning problems of interest do not satisfy this assumption. In this report, we study bounds on the generalization error in the setting where data is generated from a non-stationary dependent process. The primary assumption we make is that the process generating this data can be modeled as a stochastic process, and that the samples received by the learner are ordered by the time they were collected at.

We start by fixing our notation and making some general assumptions. We then derive a general bound on the representativeness of a given sample in the non-stationary dependent setting. Finally, we discuss the implications of the derived bound on learnability using the ERM rule with the average risk, and argue in favor of alternatives to this paradigm. The appendix contains a formal definition of stochastic processes and other probabilistic concepts used throughout the text. All the results presented here are from [1] with minor modifications.

2 Notation and general assumptions

- $\mathcal{X}$: Domain set.
- $\mathcal{Y}$: Target set.
- $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ a Polish space.
- $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), P)$ a probability space with the Borel sigma algebra on $\mathcal{Z}$.
- $Z_{-\infty}^{\infty}$: the stochastic process modeling the data generating process.
- Z_a^b : the vector of random variables $(Z_a, ..., Z_b)$.
- P_a^b : the distribution of Z_a^b

- z_1^T : samples in the original order in which they were drawn. Realizations of the random variables Z_1^T.
- $\mathcal{H}$: Hypothesis set.
- $l(h, z) : \mathcal{H} \times \mathcal{Z} \rightarrow R^+$: Loss function. Assumed to be Borel in its second argument and bounded by $M \in R^+$.

3 Representativeness bound

In this section we develop the main result of this report, which is a bound on the representativeness of a given a sample. Informally, our goal is to quantify how much we can rely on our sample to distinguish between the hypotheses in $\mathcal{H}$ in terms of their ability to make correct predictions. Assuming all our prior knowledge is put in the choice of the hypotheses set, this is the only measure we can rely on to choose the best hypothesis.

In our current setting where we are modeling the data generating process as a stochastic process, choosing the best hypothesis could mean two different things. We may only be interested in making predictions on the current process, or we may be interested in making predictions about both the current process and some other version of it. It is worth noting that this distinction does not exist in the i.i.d case (and more generally when the random variables modeling the training set are independent of the random variables modeling the testing set).

This motivates us to define two distinct risk functions:

$$\mathcal{L}_{T+s}(h) := \mathop{\mathbb{E}}_{Z_{T+s} \sim P_{T+s}} \left[l(h, Z_{T+s}) \mid Z_1^T = z_1^T \right]$$

$$\overline{\mathcal{L}}_{T+s}(h) := \mathop{\mathbb{E}}_{Z_{T+s} \sim P_{T+s}} \left[l(h, Z_{T+s}) \right]$$

for some fixed $s \in \mathbb{N}$. We refer to the first quantity by the path-dependent error, and refer to the second by the average error. Note that the first quantity is defined in terms of the random variable $l(h, Z_{T+s}) \mid Z_1^T = z_1^T$ as defined in definition 4 (or, equivalently, the usual conditional expectation).

Our aim will be to bound the representativeness of our sample by a function of the number of collected samples, which we formally define as:

$$\Phi(z_1^T) := \sup_{h \in \mathcal{H}} \left| \mathcal{L}_{T+s}(h) - \frac{1}{T} \sum_{i=1}^{T} l(h, z_i) \right| = \max \left\{ \Phi^+(z_1^T), \Phi^-(z_1^T) \right\}$$

where in $\Phi^+(z_1^T)$ the argument of the supremum is without the absolute sign, and the order of the substraction is the same as the one appearing in $\Phi(z_1^T)$, whereas in $\Phi^-(z_1^T)$ the order of the substraction is reversed. We define $\overline{\Phi}(z_1^T)$ similarly. We will be mainly presenting results for $\overline{\Phi}(z_1^T)$, but we will discuss the importance of studying $\Phi(z_1^T)$ in our discussion.

All the bounds that we present here are in terms of a slightly more general version of Rademacher complexity which we will define in the next section. We will closely follow the approach taken in the i.i.d. case:

1. Bound $\underset{Z_1^T \sim P_1^T}{\mathbb{E}}\left[\overline{\Phi}(Z_1^T)\right]$ using Rademacher complexity.
2. Bound the deviation of $\overline{\Phi}(Z_1^T)$ from its mean using Hoeffding's inequality.

Both steps rely on the i.i.d assumption. Our approach will be to reduce the non-stationary dependent case to the i.i.d case, and use the above steps to recover a bound. Our first step will be to approximate our dependent samples with independent ones. We then introduce a measure of non-stationarity of the process and argue that it is appropriate in our setting. We end this section by combining these ideas to present a bound on $\overline{\Phi}(Z_1^T)$, and we compare it to the standard result in the i.i.d. case.

3.1 Dealing with dependency between samples

We assume that $Z_{-\infty}^{\infty}$ is a β-mixing process. Our goal is to relate β-mixing samples to independent samples. Intuitively, we expect samples collected far apart in time to be less dependent on each other, while we expect successive samples to have a stronger dependence. We make this assumption precise by assuming that the the β-mixing coefficient function is decreasing.

Following our intuition, a natural way of reducing the correlation of our samples is therefore to pick m samples separated by some number of other samples a. The problem with this approach is that only a small fraction of the overall samples gets used. Generalizing this idea, we instead divide the random variables modeling our samples into $2m$ blocks $(Z(1), Z(2), ..., Z(2m))$ (while keeping our samples in order), each of length a_i so that $\sum_{i=1}^{2m} a_i = T$, and think of each block as a sample. Let us denote by P_i the distribution of the i^th block.

We refer to the odd blocks by $Z^o = (Z(1), Z(3), ..., Z(2m-1))$ and to their distribution by P^o, and we refer to the even blocks by Z^e and their distribution by P^e. Denote by I^o the set of indices of elements of Z_1^T in Z^o and let $a^o = (a_1, \dots, a_{2m-1})$ be the vector holding the length of each block contained in Z^o. Define I^e and a^e similarly for Z^e. Again our intuition tells us that picking blocks that are themselves separated by another block helps in reducing the correlation. Now if were to replace the sequence of blocks Z^o by a sequence of independent blocks $\tilde{Z}^o = (\tilde{Z}(1), \tilde{Z}(3), ..., \tilde{Z}(2m-1))$, how different would the distribution $\tilde{P}^o := \otimes_{i \in \{1,3,\dots,2m-1\}} P_i$ of these independent blocks be from P^o ? The following proposition answers this question and makes our intuition precise.

Proposition 1. *Let g be a real-valued Borel function such that $-M_1 \leq g \leq M_2$*

$$\left| \underset{Z^o \sim P^o}{\mathbb{E}}[g(Z^o)] - \underset{\tilde{Z}^o \sim \tilde{P}^o}{\mathbb{E}}[g(\tilde{Z}^o)] \right| \leq (M_1 + M_2) \sum_{i=1}^{m-1} \beta(a_{2i})$$

The proof of proposition 2, although very nice, is quite long, and we omit it here. It originally appeared in [3] but it can also be found in [1]. (It is really where all the probability concepts are used). The statement confirms our intuition: as we increase the distance between blocks (the $a'_{2i}s$), the dependency between blocks decreases and we get a tighter bound. On the other hand, as we increase the number of blocks, which we are treating as samples, the bound becomes looser. The optimal choice of number of blocks/length of blocks will depend on whether we favor a higher precision or a higher confidence in our choice of hypothesis as we will see later when we derive our bound.

Using this result, let us now relate the representativeness of the dependent samples to the representativeness of the independent versions of the odd and even blocks.

Lemma 1. *Let $\epsilon > 0$, then:*

$$P(\overline{\Phi}(Z_1^T) > \epsilon) \leq P(\overline{\Phi}(\tilde{Z}^o) > \epsilon) + P(\overline{\Phi}(\tilde{Z}^e) > \epsilon) + \sum_{i=2}^{2m-1} \beta(a_i)$$

Proof. Using convexity of the supremum we have:

$$\begin{aligned}
\overline{\Phi}(Z_1^T) &= \sup_{h\in\mathcal{H}} \left| \overline{\mathcal{L}}_{T+s}(h) - \frac{1}{T}\sum_{i=1}^{T} l(h, Z_i) \right| \\
&= \sup_{h\in\mathcal{H}} \left| \overline{\mathcal{L}}_{T+s}(h) - \frac{|I^o|}{T}\frac{1}{|I^o|}\sum_{i\in I^o} l(h, Z_i) - \frac{|I^e|}{T}\frac{1}{|I^e|}\sum_{i\in I^e} l(h, Z_i) \right| \\
&\leq \frac{|I^o|}{T} \sup_{h\in\mathcal{H}} \left| \overline{\mathcal{L}}_{T+s}(h) - \frac{1}{|I^o|}\sum_{i\in I^o} l(h, Z_i) \right| + \frac{|I^e|}{T} \sup_{h\in\mathcal{H}} \left| \overline{\mathcal{L}}_{T+s}(h) - \frac{1}{|I^e|}\sum_{i\in I^e} l(h, Z_i) \right| \\
&= \frac{|I^o|}{T}\overline{\Phi}(Z^o) + \frac{|I^e|}{T}\overline{\Phi}(Z^e)
\end{aligned}$$

By the union bound, we get:

$$P(\overline{\Phi}(Z_1^T) > \epsilon) \leq P(\overline{\Phi}(Z^o) > \epsilon) + P(\overline{\Phi}(Z^e) > \epsilon)$$

Applying proposition 2 with $g(X) := \mathbb{1}_{\{X>\epsilon\}}$ and arguments $\overline{\Phi}(Z^o)$ and $\overline{\Phi}(Z^e)$ respectively, we get the desired result. □

In order to be able to treat the independent blocks as independent samples, it is useful to define Rademacher complexity for independent blocks as follows:

$$\mathcal{R}(Z^o) := \frac{1}{|I^o|} \underset{\substack{\tilde{Z}^o\sim\tilde{P}^o \\ \sigma\sim\mathcal{U}(\{-1,1\}^T)}}{\mathbb{E}} \left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{m} \sigma_i l(h, \tilde{Z}(2i-1)) \right]$$

where $l(h, \tilde{Z}(2i-1)) = \sum\limits_{j\in I_i^o} l(h, \tilde{Z}_j)$. And similarly for the odd blocks. Notice that when the size of the blocks is 1, this reduces to the standard definition of Rademacher complexity.

3.2 Quantifying non-stationarity

In order to be able to derive any non-trivial bound on the representativeness of our sample, we need to be able to quantify how much the distribution of the random variables of the generating process changes over time. A pure measure of this change relying only on measure-theoretic concepts such as total variation distance is not very useful in this case since we are only interested in how this change affects our choice of hypothesis. Instead, we introduce the concept of discrepancy:

$$\overline{d}(t_1, t_2) := \sup_{h \in \mathcal{H}} \left| \overline{\mathcal{L}}_{t_1}(h) - \overline{\mathcal{L}}_{t_2}(h) \right|$$

As an example of the usefulness of this measure in our setting, consider the following important special case.

Example 1. *Let $\mathcal{X} = R^n$ and $\mathcal{Y} = R$ and consider the class of linear hypotheses $\mathcal{H} := \{h(w, X_1^T) = w^T X_1^T \mid w \in R^n\}$. Assume that the data generating process is weakly stationary (i.e. $\underset{Z_t \sim P_t}{\mathbb{E}}[Z_t]$ is the same for all $t \in \mathbb{Z}$). Then the discrepancy vanishes at all times (t_1, t_2), even though the process is not stationary in the strict sense. The proof of this statement can be found in [1].*

It will be useful to refer to the average discrepancy for a given set of indices I in $[T]$:

$$\Delta(I) := \frac{1}{|I|} \sum_{t \in I} \overline{d}(t, T+s)$$

Using this measure of non-stationarity, let us bound the expectation of the representativeness of some sample of size T, realizations of random variables $\tilde{Z}_1^T$ of a stochastic process $\tilde{Z}_{-\infty}^{\infty}$ with distribution $\tilde{P}_{-\infty}^{\infty}$, assuming independence between the random variables of the process.

Lemma 2.

$$\underset{\tilde{Z}_1^T \sim \tilde{P}_1^T}{\mathbb{E}} \left[\Phi(\tilde{Z}_1^T) \right] \leq 2\mathcal{R}(\tilde{Z}_1^T) + \Delta([T])$$

Proof. We have:

$$\begin{aligned}
\underset{\tilde{Z}_1^T \sim \tilde{P}_1^T}{\mathbb{E}} \left[\overline{\Phi}^+(\tilde{Z}_1^T) \right] &= \underset{\tilde{Z}_1^T \sim \tilde{P}_1^T}{\mathbb{E}} \left[\sup_{h \in \mathcal{H}} \left(\overline{\mathcal{L}}_{T+s}(h) - \frac{1}{T} \sum_{i=1}^T l(h, \tilde{Z}_i) \right) \right] \\
&\leq \underset{\tilde{Z}_1^T \sim \tilde{P}_1^T}{\mathbb{E}} \left[\sup_{h \in \mathcal{H}} \left(\overline{\mathcal{L}}_{T+s}(h) - \frac{1}{T} \sum_{i=1}^T \overline{\mathcal{L}}_i(h) + \frac{1}{T} \sum_{i=1}^T \overline{\mathcal{L}}_i(h) - \frac{1}{T} \sum_{i=1}^T l(h, \tilde{Z}_i) \right) \right] \\
&\leq \underset{\tilde{Z}_1^T \sim \tilde{P}_1^T}{\mathbb{E}} \left[\sup_{h \in \mathcal{H}} \frac{1}{T} \sum_{i=1}^T \left(\overline{\mathcal{L}}_{T+s}(h) - \overline{\mathcal{L}}_i(h) \right) + \sup_{h \in \mathcal{H}} \frac{1}{T} \sum_{i=1}^T \left(\overline{\mathcal{L}}_i(h) - l(h, \tilde{Z}_i) \right) \right] \\
&\leq \underset{\tilde{Z}_1^T \sim \tilde{P}_1^T}{\mathbb{E}} \left[\frac{1}{T} \sum_{i=1}^T \sup_{h \in \mathcal{H}} \left| \overline{\mathcal{L}}_{T+s}(h) - \overline{\mathcal{L}}_i(h) \right| \right] + \underset{\tilde{Z}_1^T \sim \tilde{P}_1^T}{\mathbb{E}} \left[\sup_{h \in \mathcal{H}} \frac{1}{T} \sum_{i=1}^T \left(\overline{\mathcal{L}}_i(h) - l(h, \tilde{Z}_i) \right) \right]
\end{aligned}$$

The first term is bounded by $\Delta([T])$ by definition. The second term can be bounded using the same approach we used in the i.i.d. case. Notice that by definition we have for all $i \in \mathbb{Z}$:

$$\mathop{\mathbb{E}}_{\tilde{Z}_i \sim \tilde{P}_i}\left[l(h, \tilde{Z}_i)\right] = \overline{\mathcal{L}}_i(h)$$

Now let $\tilde{Z'}_1^T$ be another version of $\tilde{Z}_1^T$. Then we have:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\tilde{Z}_1^T \sim \tilde{P}_1^T}\left[\sup_{h \in \mathcal{H}} \frac{1}{T}\sum_{i=1}^{T}\left(\overline{\mathcal{L}}_i(h) - l(h, \tilde{Z}_i)\right)\right] &= \mathop{\mathbb{E}}_{\tilde{Z}_1^T \sim \tilde{P}_1^T}\left[\frac{1}{T}\sum_{i=1}^{T}\left(\mathop{\mathbb{E}}_{\tilde{Z'}_i \sim P_i}\left[l(h, \tilde{Z}'_i)\right] - l(h, \tilde{Z}_i)\right)\right] \\
&\leq \mathop{\mathbb{E}}_{\tilde{Z'}_1^T, \tilde{Z}_1^T \sim P_1^T}\left[\sup_{h \in \mathcal{H}} \frac{1}{T}\sum_{i=1}^{T}\left(l(h, \tilde{Z}'_i) - l(h, \tilde{Z}_i)\right)\right] \\
&= \frac{1}{T} \mathop{\mathbb{E}}_{\substack{\tilde{Z'}_1^T, \tilde{Z}_1^T \sim P_1^T \\ \sigma \sim \mathcal{U}(\{-1,1\}^T)}}\left[\sup_{h \in \mathcal{H}} \sum_{i=1}^{T} \sigma_i \left(l(h, \tilde{Z}'_i) - l(h, \tilde{Z}_i)\right)\right] \\
&= \frac{2}{T} \mathop{\mathbb{E}}_{\substack{\tilde{Z}_1^T \sim P_1^T \\ \sigma \sim \mathcal{U}(\{-1,1\}^T)}}\left[\sup_{h \in \mathcal{H}} \sum_{i=1}^{T} \sigma_i l(h, \tilde{Z}_i)\right] \\
&= 2\mathcal{R}(\tilde{Z}_1^T)
\end{aligned}$$

Where the second line follows from the independence of Z'^T_1 and Jensen's inequality. The argument is symmetric in the order of the substraction, and therefore the same holds for $\overline{\Phi}^-(Z_1^T)$.

□

Note that the argument above only makes use of the independence of the samples, and therefore applies to independent blocks as well with no modifications (aside from changing T to the number of blocks).

3.3 Main result

Theorem 1. *For all $\delta \geq \sum_{i=2}^{2m-1} \beta(a_i)$, with probability $1 - \delta$ we have:*

$$\overline{\Phi}(Z_1^T) \leq \max\left(2\mathcal{R}(\tilde{Z}^o) + \Delta(I^o), 2\mathcal{R}(\tilde{Z}^e) + \Delta(I^e)\right) + \sqrt{\frac{\log \frac{4}{\delta'}}{2T^2/\max\{\|a^e\|_2^2, \|a^o\|_2^2\}}}$$

where $\delta' = \delta - \sum_{i=2}^{2m-1} \beta(a_i)$

Proof. Following the approach used in the i.i.d case, we start by bounding the expectation of the representativeness. By Lemma 2 we have:

$$\mathop{\mathbb{E}}_{\tilde{Z}^o \sim \tilde{P}^o}\left[\overline{\Phi}(\tilde{Z}^o)\right] \leq 2\mathcal{R}(\tilde{Z}^o) + \Delta(|I^o|)$$

Now if we treat the function $\overline{\Phi}(\tilde{Z}^o) = \overline{\Phi}(\tilde{Z}(1), \tilde{Z}(2), ..., \tilde{Z}(2m-1))$ as a function of the independent blocks, then it is bounded in each of its arguments by $\frac{a_i}{T}M$ (since we assumed the loss function is bounded by M). Applying Hoeffding's inequality, we get for all $\epsilon \geq 0$:

$$P\left(\left|\overline{\Phi}(\tilde{Z}^o) - \mathop{\mathbb{E}}_{\tilde{Z}^o \sim \tilde{P}^o}\left[\overline{\Phi}(\tilde{Z}^o)\right]\right| \geq \epsilon\right) \leq 2\exp\left(\frac{-2T^2\epsilon^2}{\|a^o\|_2^2 M^2}\right)$$

Doing the same thing for $\tilde{Z}^e$, and using Lemma 1 with $\epsilon := K + \epsilon'$ where:

$$K := \max\left(\mathop{\mathbb{E}}_{\tilde{Z}^o \sim \tilde{P}^o}\left[\overline{\Phi}(\tilde{Z}^o)\right], \mathop{\mathbb{E}}_{\tilde{Z}^e \sim \tilde{P}^o}\left[\overline{\Phi}(\tilde{Z}^e)\right]\right)$$

we get:

$$\begin{aligned} P(\overline{\Phi}(Z_1^T) > \epsilon' + K) &\leq P(\overline{\Phi}(\tilde{Z}^o) > \epsilon' + K) + P(\overline{\Phi}(\tilde{Z}^e) > \epsilon' + K) + \sum_{i=2}^{2m-1}\beta(a_i) \\ &\leq 2\exp\left(\frac{-2T^2(\epsilon'+K)^2}{\|a^o\|_2^2 M^2}\right) + 2\exp\left(\frac{-2T^2(\epsilon'+K)^2}{\|a^e\|_2^2 M^2}\right) + \sum_{i=2}^{2m-1}\beta(a_i) \\ &\leq 4\exp\left(\frac{-2T^2(\epsilon'+K)^2}{\max\{\|a^e\|_2^2, \|a^o\|_2^2\}M^2}\right) + \sum_{i=2}^{2m-1}\beta(a_i) \end{aligned}$$

Letting

$$\epsilon' := M\sqrt{\frac{\log\frac{4}{\delta'}}{2T^2/\max\{\|a^e\|_2^2, \|a^o\|_2^2\}}} - K$$

and using the bound on the expectations we recover the desired result. □

The bound above is very similar in form to the one we derived in the i.i.d case with the following differences. To simplify our analysis, let us assume that we divide our sample into $2m$ equal blocks of length a. First, looking at the last term, we see that the effective sample size decreases from T in the i.i.d case to $4m$ due to the dependency between the samples. Similarly, using Massart's lemma on the Rademacher complexity term, we get an $O(\sqrt{\log m}/m)$ bound compared to $O(\sqrt{\log T}/T)$ in the i.i.d case. Our goal initially was to build blocks that act like independent samples, and this result shows that we have achieved just that (up to a factor of 2). Finally the discrepancy term takes into account the non-stationarity of the process.

An attractive feature of this result is that it allows the incorporation of new information about the process to derive tighter bounds. For example, if the process is known to be stationary, then the discrepancy term vanishes and we get the same bound as in [2]. Similarly, if the process is $\boldsymbol{b}$-asymptotically stationary, then the discrepancy term can be bounded using the $\boldsymbol{b}$-stationarity coefficients.

4 Implications for learnability

Our aim from the beginning was to generalize the known learnability results in the i.i.d scenario to a general setting with minimal assumptions. Therefore, a natural question to ask is what is the minimial set of assumptions needed to guarantee learnability in the case where we model the data-generating process as an ordered stochastic process. This question is difficult to answer in full generality, but we can hope to find a set of assumptions that are more general than the i.i.d. in which we can guarantee learnability. The following theorem shows that even the stationarity assumption is not enough to achieve that.

Theorem 2. *Assume that $VC(\mathcal{H}) \geq 2$. Then for any algorithm $\mathcal{A}$ that outputs a hypothesis $h_{z_1^T}$ when receiving samples z_1^T, there is a stationary process that is not β-mixing such that for each $T \in \mathbb{N}$, there exists $T' \geq T$ such that:*

$$\mathbb{P}\left(\overline{L}_{T'+1}(h_{Z_1^T}) - \inf_{h \in \mathcal{H}} \overline{L}_{T'+1}(h) \geq \frac{1}{2}\right) \geq \frac{1}{8}$$

The proof is not too difficult but quite lengthy, and we do not include it here for conciseness. It can be found in [1]. The theorem shows that a necessary condition for learnability is stationarity and β-mixing of the data-generating process. This justifies our attention to the β-mixing process setting. Assuming stationarity and β-mixing, and assuming that given a sample of size T, we divide our sample into $2m_T$ equal blocks of length a_T we can see from Theorem 1 that a sufficient condition for learnability is $2(m_T - 1)\beta(a_T) \to 0 \wedge m_T \to \infty$ as $T \to \infty$.

Given Theorem 2 above, it might seem quite hopeless to generalize learning far from the i.i.d assumption. Not only do we need stationarity, we also need high mixing rates, which almost brings us back to the i.i.d assumption. This might indeed be the case if we are measuring our ability to learn using the average error as we have been doing. I think that the issue resides in using this measure. It seems that there are simply too many degrees of freedom in stochastic processes (and some real world processes) for us to expect to get the sort of generalization we require by using this measure from a finite amount of samples.

One alternative to this measure of error is to use the path-dependent error. While learning under this error will not necessarily allow us to generalize to other version of the process, we can at least expect to perform well on the current process of interest, and move past the i.i.d assumption. Path-dependent learning bounds are also given in [1], which show that learning under this error can be done in cases where it is impossible under the average error.

Another advantage of using a path-dependent error is that dependence between samples is not necessarily detrimental to learning as is the case with the average error. We can in fact take advantage of this dependency to improve on our predictions. A simple example that I can think of is if we model our data-generating process as a time-homogoneous Markov chain. Depending on the strength of the dependency between successive elements of the chain, trying to build independent blocks out of this chain might be hopeless. On the other hand, if we focus on a particular instance of the chain, we can hope to learn a good approximation of the expectation function of the transition kernel and perform well in predicting the near future given our samples. Deriving Rademacher complexity bounds for this learning problem under the path-dependent error is I think an interesting problem, although there might already be results on this.

5 Conclusion

Although a little discouraging, I think the results presented in this report are quite insightful, and show us how limited the empirical average risk minimization paradigm is outside the i.i.d. setting. I think this limitation comes from two distinct sources: (i) The choice of the risk function as argued in the previous section. (ii) The failure to incorporate prior knowledge on $\mathcal{H}$.

Using empirical risk minimization on a path-dependent risk instead and introducing prior knowledge on $\mathcal{H}$ may help us overcome these limitations, but I think that ultimately a full probabilistic treatment might offer a better learning paradigm by allowing a natural incorporation of prior knowledge in the form of a prior distribution, and a natural way of updating our beliefs through posterior inference. Developing learning bounds in the Bayesian paradigm is, I think, a very interesting problem, although solving it may require vastly different tools than the ones used to develop bounds for the ERM paradigm.

A Probability and Stochastic processes

For this appendix, let (Ω, Σ, Q) be a probability space, and let $(E, \mathcal{E})$ be a measurable space. All random variables in this section, unless specified otherwise, are functions from Ω to E.

Definition 1. *Let T be an index set. A stochastic process is a collection of random variables $X = \{X_t\}_{t \in T}$. We will assume that T is totally ordered, and write $X = (X_t)_{t \in T}$ to mean the ordered collection of these random variables.*

Proposition 2. *A stochastic process $X = (X_t)_{t \in T}$ is a random variable taking value in $E^T := \prod\limits_{t \in T} E$ with the product σ-algebra*
$\mathcal{E}^T = \otimes_{t \in T} \mathcal{E} := \sigma(\{\prod\limits_{t \in T} A_t \mid \forall t \in T \ A_t \subseteq E \wedge \exists I \subseteq T : (|I| < \infty \wedge \forall t \in T \setminus I \ A_t = E)\})$.

Proof. Let $A \in \{\prod\limits_{t \in T} A_t \mid I \subseteq T, |I| < \infty, \forall t \in T \setminus I \ A_t = E\}$, and suppose $\{t_i\}_{i=1}^n \subseteq T$ are the components of A such that $\pi_{t_i}(A) \neq E$ where π_{t_i} is the t_i^{th} projection map. Then:

$$\begin{aligned}
X^{-1}(A) &= \{\omega \in \Omega \mid \forall t \in T \ X_t(w) \in \pi_t(A)\} \\
&= \{\omega \in \Omega \mid \forall i \in [n] \ X_{t_i}(w) \in \pi_{t_i}(A)\} \\
&= \bigcap_{i=1}^{n} \{\omega \in \Omega \mid X_{t_i}(w) \in \pi_{t_i}(A)\} \\
&= \bigcap_{i=1}^{n} X_{t_i}^{-1}(\pi_{t_i}(A)) \\
&\in \Sigma
\end{aligned}$$

Where the last line is true since the X_{t_i} are random variables and Σ is closed under countable intersections. The rest of the proof relies on the fact that the preimage operator is interchangeable with complementation, intersections, and unions of sets. □

Remark 1. *One might think that the restriction on I to be finite is too strong, but this is not the case. It is easily seen that allowing I to be countably infinite produces the same sigma algebra. If I is allowed to be uncountably infinite (assuming that T is), X is no longer a random variable (can we even assign a probability to such a space ?).*

Remark 2. *The distribution of X is fully determined by the joint distribution of all finite subcollections of X. Continuity of probability measure ensures that the joint distribution of countable subcollections of X is well defined.*

Definition 2. *Let (Ω, Σ, Q) and (Ω', Σ', Q') be two probability spaces. We define the product probability measure P on $(\Omega \times \Omega', \Sigma \otimes \Sigma')$ to be the probability measure that satisfies:*

$$\forall A \in \Sigma, B \in \Sigma' \ P(A \times B) = Q(A)Q'(B)$$

The existence and uniqueness of this measure is guaranteed by the Hahn–Kolmogorov theorem. We write $P = Q \otimes Q'$.

Definition 3. *Let P and Q be two set functions on a σ-algebra Σ of a set Ω. The total variation distance between P and Q is defined to be:*

$$\|P - Q\|_{TV} := \sup_{A \in \Sigma} \big|P(A) - Q(A)\big|$$

The total variation distance forms a metric on the set of probability measures on a given measurable space.

Definition 4. *Let X be a random variable. A regular conditional distribution for Q given X is a function $Q(\cdot \mid X = \cdot) : \Sigma \times E \to [0,1]$ such that:*

1. *$\forall x \in E \;\; Q(\cdot \mid X = x)$ is a probability measure on (Ω, Σ).*
2. *$\forall A \in \Sigma \;\; Q(A \mid X = \cdot)$ is a random variable $E \to [0,1]$ with the Borel sigma algebra on $[0,1]$.*
3. *$\forall A \in \Sigma, B \in \mathcal{E} \;\; \int_B Q(A \mid X = x)\; d(Q \circ X^{-1}) = Q(A \cap X^{-1}(B))$*

The existence of a regular conditional distribution depends on the nature of the space $(E, \mathcal{E})$. In particular, It can be shown that if E is a Polish space, and $\mathcal{E}$ is the Borel sigma algebra on E, then such a function exists and $\forall A \in \Sigma$, the random variable $Q(A \mid X = \cdot)$ is uniquely defined on E $(Q \circ X^{-1})$-almost surely.

Let $Y : \Omega \to E'$ be another random variable. Then we define $P_Y(\cdot \mid X = \cdot) : \mathcal{E}' \times E \to [0,1]$ as follows:

$$\forall A' \in \mathcal{E}', x \in E \;\; P_Y(A' \mid X = x) = Q(Y^{-1}(A') \mid X = x)$$

By construction, for each $x \in E$, $P_Y(\cdot \mid X = x)$ is a distribution of Y. We refer to Y distributed according to $P_Y(\cdot \mid X = x)$ by $Y \mid X = x$

For the remaining of this appendix we assume that X is a stochastic process with index set $\mathbb{Z}$. We denote by $(E_a^b, \mathcal{E}_a^b, P_a^b)$ the target space and distribution of $(X_t)_{t=a}^b$ where $a, b \in \mathbb{Z} \cup \{-\infty, \infty\}$ and $b \geq a$. If b is not specified, then we are only referring to the space and distribution of the random variable X_a.

Definition 5. *The β-mixing coefficient of X is the function $\beta : \mathbb{N} \to [0,1]$ defined as:*

$$\forall a \in \mathbb{N} \;\; \beta(a) = \sup_{t \in \mathbb{Z}} \; \mathop{\mathbb{E}}_{X_{-\infty}^t \sim P_{-\infty}^t} \Big[\, \big\|P_{t+a}^{\infty}(\cdot \mid X_{-\infty}^t) - P_{t+a}^{\infty}\big\|_{TV} \Big]$$

A process X is said to be β-mixing if $\lim_{a \to \infty} \beta(a) = 0$.

Definition 6. *The φ-mixing coefficient of X is the function $\varphi : \mathbb{N} \to [0,1]$ defined as:*

$$\forall a \in \mathbb{N} \;\; \varphi(a) = \sup_{t \in \mathbb{Z}} \; \sup_{\substack{A \in \mathcal{E}_{-\infty}^t \\ P_{-\infty}^t(A) \neq 0}} \big\|P_{t+a}^{\infty}(\cdot \mid X_{-\infty}^t \in A) - P_{t+a}^{\infty}\big\|_{TV}$$

where $\forall B \in \mathcal{E}_{t+a}^{\infty} \;\; P_{t+a}^{\infty}(B \mid X_{-\infty}^t \in A) := P_{-\infty}^t \wedge P_{t+a}^{\infty}(A \times B)/P_{-\infty}^t(A)$ is the usual conditional probability ($P_{-\infty}^t \wedge P_{t+a}^{\infty}$ denotes the joint distribution of $(X_i)_{i=-\infty}^t$ and $(X_i)_{i=t+a}^{\infty}$). A process X is said to be φ-mixing if $\lim_{a \to \infty} \varphi(a) = 0$.

Remark 3. *The β-mixing coefficient measures the distance between the conditional and unconditional distribution averaged over all possible conditions weighted by their probabilities, whereas the φ-mixing coefficient measures the maximum distance (not in the formal sense since $P_{t+a}^{\infty}(B \mid X_{-\infty}^{t} \in A)$ is not a probability measure) between them over all possible collections of conditions of non-zero measure. It is clear from the definition that $\forall a \in \mathbb{N}$ $\beta(a) \leq \varphi(a)$ so that an φ-mixing process is necessarily β-mixing.*

Definition 7. *X is said to be strictly stationary if:*

$$\forall t, n, m \in \mathbb{Z} \quad P_{t}^{t+n} = P_{t+m}^{t+m+n}$$

This is equivalent to requiring that the distribution of any finite subcollection $\{X_{t_i}\}_{i=1}^{n}$ is the same as $\{X_{t_i+k}\}_{i=1}^{n}$ for all $k \in \mathbb{Z}$.

Definition 8. *The $\mathfrak{b}$-stationarity coefficient of X with respect to a distribution Π is the function $\mathfrak{b}_{\Pi} : \mathbb{N} \to [0,1]$ defined as:*

$$\forall a \in \mathbb{N} \quad \mathfrak{b}_{\Pi}(a) = \sup_{t \in \mathbb{Z}} \mathop{\mathbb{E}}_{X_{-\infty}^{t} \sim P_{-\infty}^{t}} \left[\left\| P_{t+a}(\cdot \mid X_{-\infty}^{t}) - \Pi \right\|_{TV} \right]$$

A process X is said to be $\mathfrak{b}$-asymptotically stationary if there exists a distribution Π such that $\lim\limits_{a \to \infty} \mathfrak{b}_{\Pi}(a) = 0$. Such a Π is called a $\mathfrak{b}$-stationary distribution of X. (I do not know if such a distribution is unique or not).

Definition 9. *The ϕ-stationarity coefficient of X with respect to a distribution Π is the function $\phi_{\Pi} : \mathbb{N} \to [0,1]$ defined as:*

$$\forall a \in \mathbb{N} \quad \phi_{\Pi}(a) = \sup_{t \in \mathbb{Z}} \sup_{\substack{A \in \mathcal{E}_{-\infty}^{t} \\ P_{-\infty}^{t}(A) \neq 0}} \left\| P_{t+a}(\cdot \mid X_{-\infty}^{t} \in A) - \Pi \right\|_{TV}$$

A process X is said to be ϕ-asymptotically stationary if there exists a distribution Π such that $\lim\limits_{a \to \infty} \mathfrak{b}_{\Pi}(a) = 0$. Such a Π is called a ϕ-stationary distribution of X.

References

[1] Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106(1):93–117, 2017.

[2] Rostamizadeh Mehryar, Mohri; Afshin. Rademacher Complexity Bounds for Non-I.I.D. Processes.

[3] Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability, 1994.*