Learning from Weakly Dependent data under Dobrushin's condition

Yuval Dagan DAGAN@MIT.EDU
EECS & CSAIL, MIT

Constantinos Daskalakis COSTIS@CSAIL.MIT.EDU
EECS & CSAIL, MIT

Nishanth Dikkala NISHANTHD@CSAIL.MIT.EDU
EECS & CSAIL, MIT

Siddhartha Jayanti JAYANTI@MIT.EDU
EECS & CSAIL, MIT

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

Statistical learning theory has largely focused on learning and generalization given independent and identically distributed (i.i.d.) samples. Motivated by applications involving time-series data, there has been a growing literature on learning and generalization in settings where data is sampled from an ergodic process. This work has also developed complexity measures, which appropriately extend the notion of Rademacher complexity to bound the generalization error and learning rates of hypothesis classes in this setting. Rather than time-series data, our work is motivated by settings where data is sampled on a network or a spatial domain, and thus do not fit well within the framework of prior work. We provide learning and generalization bounds for data that are complexly dependent, yet their distribution satisfies the standard Dobrushin's condition. Indeed, we show that the standard complexity measures of Gaussian and Rademacher complexities and VC dimension are sufficient measures of complexity for the purposes of bounding the generalization error and learning rates of hypothesis classes in our setting. Moreover, our generalization bounds only degrade by constant factors compared to their i.i.d. analogs, and our learnability bounds degrade by log factors in the size of the training set.

1. Introduction

A main goal in statistical learning theory is understanding whether observations of some phenomenon of interest can be used to make confident predictions about future observations. Usually this question is studied in the setting where a training set $\boldsymbol{S} = (\boldsymbol{x}_i, \boldsymbol{y}_i)_{i=1}^m$, comprising pairs of covariate vectors $\boldsymbol{x}_i \in \mathcal{X}$ and response variables $\boldsymbol{y}_i \in \mathcal{Y}$, are drawn independently from some unknown distribution D, and the goal is to make predictions about a future sample $(\boldsymbol{x}, \boldsymbol{y})$ drawn independently from the same distribution D. That is, we wish to predict $\boldsymbol{y}$ given $\boldsymbol{x}$.

Given some hypothesis class $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$, comprising predictors that map $\mathcal{X}$ to $\mathcal{Y}$ and a loss function $\ell : \mathcal{Y}^2 \to \mathbb{R}$ whose values $\ell(\hat{y}, y)$ express how bad it is to predict $\hat{y}$ instead of y, a wealth of results characterize the relationship between the size m of the training set $\boldsymbol{S}$ and the approximation accuracy that is attainable for choosing some predictor $h \in \mathcal{H}$ whose expected loss, $L_D(h) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim D}\ell(h(\boldsymbol{x}), \boldsymbol{y})$, on a future sample, is as small as possible. A related question is understanding how well the training set $\boldsymbol{S}$ "generalizes," in the sense of minimizing $\sup_{h\in\mathcal{H}} |L_S(h) - L_D(h)|$, where $L_S(h)$ is the average loss of h on the training set $\boldsymbol{S}$. To characterize the learnability and generalization properties of hypotheses classes, standard complexity measures of function classes, such as the VC dimension (Vapnik and Chervonenkis, 2015) and the Rademacher complexity (Bartlett and Mendelson, 2002), have been developed.

The assumption that the training examples $(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)$ as well as the future test sample $(\boldsymbol{x}, \boldsymbol{y})$ are all independently and identically distributed (i.i.d.) is, however, too strong in many applications. Often, training data points are observed on nodes of a network, or some spatial or temporal domain, and are *dependent* both with respect to each other and with respect to future observations. Examples abound in financial and meteorological applications, and dependencies naturally arise in social networks through *peer effects*, whose study has recently exploded in topics as diverse as criminal activity (see e.g. Glaeser et al., 1996), welfare participation (see e.g. Bertrand et al., 2000), school achievement (see e.g. Sacerdote, 2001), participation in retirement plans (see Duflo and Saez, 2003), and obesity (see e.g. Trogdon et al., 2008; Christakis and Fowler, 2013). A prominent dataset where network effects are studied was collected by the National Longitudinal Study of Adolescent Health, a.k.a. AddHealth study (Harris et al., 2009). This was a major national study of students in grades 7-12, who were asked to name their friends—up to 10, so that friendship networks can be constructed, and answer hundreds of questions about their personal and school life, and it also recorded information such as the age, gender, race, socio-economic background, and health of the students. Disentangling individual effects from network effects in such settings is a recognized challenge (see e.g. the discussion by Manski 1993 and Bramoullé et al. 2009, and the discussion of prediction models for network-linked data by Li et al. 2016).

Motivated by such applications, a growing literature has studied learning and generalization in settings where data is non-i.i.d. This work goes back to at least Yu (1994), and has grown quite significantly in the past decade. A central motivation has been settings involving time-series data. As such, this literature has focused on data sampled from an ergodic process. For this type of data, generalization and learnability bounds have been obtained whose quality depends on the mixing properties of the data generation process as well as the complexity of the hypothesis class under consideration, through appropriate generalizations of the Rademacher complexity. We discuss this literature in Section 1.3, and present precise generalization bounds derived from this literature in Section 1.4.

In contrast to prior work, our main motivation is the study of networked data, due to their significance in economy and society, including in the applications discussed above. The starting point of our investigation is that data observed on a network does not fit well the statistical learning frameworks proposed for non-i.i.d. data in prior work, which targets time-series data. In particular, there is no natural ordering of observations collected on a network with respect to which one may postulate a fast-mixing/correlation-decay property, which may be exploited for statistical power. We thus propose a different statistical learning framework that is better suited to networked data.

We propose to study generalization and learnability when the training samples $\boldsymbol{S} = (\boldsymbol{x}_i, \boldsymbol{y}_i)_{i=1}^m$ are complexly dependent but their joint distribution satisfies Dobrushin's condition; see Definition 4. Dobrushin's condition was introduced by Dobrushin (1968) in the study of Gibbs measures, originally in the context of identifying conditions under which the Gibbs distribution has a unique equilibrium / stationary state and has since been well-studied in statistical physics and probability literature (see e.g. Dobrushin and Shlosman, 1987; Stroock and Zegarlinski, 1992) as it implies a number of desirable properties, such as fast mixing of Glauber dynamics (Külske, 2003), concentration of measure (Marton et al., 1996; Külske, 2003; Chatterjee, 2005b; Daskalakis et al., 2018; Gheissari et al., 2017), and correlation decay (Künsch, 1982). For a survey of properties resulting from Dobrushin's condition see Weitz (2005).

1.1. Our Results

Setting: Assuming that our training set $\boldsymbol{S}$ and test sample $(\boldsymbol{x}, \boldsymbol{y})$ are drawn from a distribution $D^{(m)}$ satisfying Dobrushin's condition, as described above, we establish a number of learnability and generalization results. We make the assumption that every example in our training set $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ comes from the same marginal distribution D which is also the distribution from which we draw the test sample. This assumption is made to provide a uniform benchmark to measure the performance of our learning algorithms against.

Our first main result, presented as Theorem 10, provides an agnostic learnability bound for any hypothesis class that is learnable in the i.i.d. setting, for instance, classes of finite VC dimension (Corollary 11). We provide an informal statement here.

Informal Theorem 1 (Learnability under Dobrushin Dependent Data) *Let $\mathcal{H}$ be a hypothesis class such that $VC(\mathcal{H}) = d$, and let L_D be the expected 0/1 loss function evaluated on a sample from D. Given a training sample $\boldsymbol{S} \sim D^{(m)}$ where $D^{(m)}$ satisfies Dobrushin's condition, there exists a learning algorithm $\mathcal{A}$ such that*

$$\Pr\left[L_D(\mathcal{A}(\boldsymbol{S})) \leq \inf_{h\in\mathcal{H}} L_D(h) + \varepsilon\right] \geq 99/100, \quad \textit{for} \quad m = \widetilde{O}\left(\frac{d}{\varepsilon^2}\right).$$

Our second main result, presented as Theorem 16, provides a generalization bound for hypothesis classes, under stronger conditions on the distribution of $\boldsymbol{S}$, which we term *bounded log-coefficient*, and define in Section 5. We bound the maximal deviation $\sup_{h\in\mathcal{H}} |L_D(h) - L_S(h)|$ in terms of the Gaussian complexity of $\mathcal{H}$, a value which is closely related to the Rademacher complexity. We obtain a bound which is nearly as tight as if the training set $\boldsymbol{S}$ was drawn i.i.d.

Informal Theorem 2 (Uniform Convergence under High Temperature Data) *Let $\mathcal{H}$ be a hypothesis class, and let $L_D(h)$ and $L_S(h)$ denote the training and expected loss, respectively, of a hypothesis h with respect to some arbitrary loss function. Given a training sample $\boldsymbol{S} \sim D^{(m)}$ where $D^{(m)}$ has log-coefficient bounded by 1, the following holds:*

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}} |L_D(h) - L_S(h)|\right] \leq O(\mathfrak{G}_{D^{(m)}}(\mathcal{H})), \tag{1}$$

where $\mathfrak{G}_{D^{(m)}}(\mathcal{H})$ is the Gaussian complexity of $\mathcal{H}$. In particular, if $\mathrm{VC}(\mathcal{H}) = d$, then the left hand side of (1) is bounded by $\sqrt{d/n}$.

1.2. Organization

In Section 1.3 we discuss the related studies along the direction of non i.i.d. generalization and learnability. In Section 1.4 we provide a comparison between our proposed framework and that of prior work on ergodic processes, and the benefits from our framework in terms of sharpness of generalization bounds. In particular, we show an example setting where our bounds are a significant improvement over the bounds implied by prior work. In Section 2 we state some preliminary notation and definitions and some Lemmas from prior work we use throughout the paper. Section 4 contains our learnability results for data satisfying the weaker Dobrushin's condition. Section 5 contains our uniform convergence bound for data satisfying the high temperature condition.

1.3. Related Work

Rademacher and Gaussian compexities for obtaining uniform convergence bounds on generalization of learning algorithms were first introduced in the work of Bartlett and Mendelson (2002) and have since been extensively studied in the literature on learning theory to characterize the sample complexity of learning for a wide range of problems. Extending them beyond i.i.d. settings was mainly studied in the context of ergodic processes and exchangeable sequences. The bounds in the literature on ergodic processes typically depend on the α or β mixing coefficients of these processes. The work on studying learnability for stationary mixing empirical processes started with seminal work of Yu (1994) and was continued by Mohri and Rostamizadeh (2009); Kuznetsov and Mohri (2015); Mohri and Rostamizadeh (2010) and the references therein. Kuznetsov and Mohri (2015) studies non-stationary and non-mixing time series, Kuznetsov and Mohri (2014) and Kuznetsov and Mohri (2017) study non-stationary and mixing time series, McDonald and Shalizi (2017) studies stationary and non-mixing time series, and Mohri and Rostamizadeh (2009) studies stationary mixing time series. Of these works, Mohri and Rostamizadeh (2009) is most relevant to ours, since McDonald and Shalizi (2017)'s work on non-mixing time series solves the forecasting problem, i.e. predicting $\boldsymbol{z}_{m+1}$ given previous data $\{\boldsymbol{z}_i\}_{i=1}^m$, rather than on predicting $\boldsymbol{y}_{m+1}$ given $\boldsymbol{x}_{m+1}$ as in our setting. Moreover, since our work focuses on distributions with identical marginals, the most closely related time-series setting to ours is one where the series is stationary. Hence, we compare our results to previous work on stationary time series in Section 1.4.

Agarwal and Duchi (2013) study the generalization properties of online algorithms in the context of stationary and mixing time series. Another direction in which dependent data have been considered is the setting of exchangeable sequences studied in the works of Berti et al. (2009); Pestov (2010) and references therein. Apart from the above extensions to non i.i.d. data, notions of sequential Rademacher complexity were considered in the literature on online learning (see Rakhlin et al. (2010)). None of these settings capture the type of dependences we handle in our work which can have long-range correlations and no spatial mixing behavior in general.

1.4. Comparison to Related Work on Time Series

Much of the work on non-iid Rademacher complexity has focused on time series. Here, we compare our results with the work most closely related to our setting and show that in certain cases our bounds improve significantly over those implied by prior work. The uniform convergence sample complexity bounds of Mohri and Rostamizadeh (2009) for stationary mixing time series are the most relevant to our setting. Mohri *et. al.*'s approach takes a 'thinning' approach to argue that a sub-sample consisting of well-separated samples is close to being independent. When the time series distributions are pairwise-potential MRFs or satisfy Dobrushin's condition, our sample complexity bounds can be much tighter since we do not need the thinning approach and hence are less wasteful.

To get a generalization gap of at most ε with probability at least $1-\delta$ on a class of hypotheses with VC-dimension d for a specific time series which (a) is stationary and fast-mixing and (b) satisfies Dobrushin's condition.:

$$m_{\text{prior-work}}(\varepsilon, \delta) = \tilde{\Theta}\left(\frac{d^2}{\delta\varepsilon^4}\right) \qquad\qquad m_{\text{this-paper}}(\varepsilon, \delta) = \Theta\left(\frac{d + \log\frac{1}{\delta}}{\varepsilon^2}\right)$$

The quadratic improvement quantifies the ineffciency of the thinning method in this context.

2. Preliminaries

Notational Conventions Random variables will be written in a bold font (say $\boldsymbol{x}$), as opposed to elements from the domain set, which are in a normal font (say, x). We will use the notation C, C', C_1, c, c' etc. to denote positive universal constants without explicitly stating it. Given a vector $x = (x_1, \ldots, x_m)$ and $i \in \{1, \ldots, m\}$, x_{-i} denotes the vector x after omitting coordinate i. Given a random variables $\boldsymbol{z}, \boldsymbol{w}$ over $(\Omega, \mathcal{F})$ and $z, w \in \Omega$, denote by $P_{\boldsymbol{z}}(z)$ the probability that $\boldsymbol{z} = z$ if $\boldsymbol{z}$ is discrete and the density of $\boldsymbol{z}$ at z if $\boldsymbol{z}$ is continuous. Additionally, define by $P_{\boldsymbol{z}|\boldsymbol{w}}(z \mid w)$ the probability $\Pr[\boldsymbol{z} = z \mid \boldsymbol{w} = w]$ if $\boldsymbol{z}$ and $\boldsymbol{w}$ are discrete and analogously if they are continuous.[1]

2.1. Learning

Fix some feature set $\mathcal{X}$, label set $\mathcal{Y}$, and a class of hypotheses $\mathcal{H}$, containing functions from $\mathcal{X}$ to $\mathcal{Y}$. Assume a loss function $\ell \colon \mathcal{Y}^2 \to \mathbb{R}$, where $\ell(\hat{y}, y)$ is the loss of predicting $\hat{y}$ when the true label is y. The simplest example of a loss function is the 0-1 loss, $\ell^{01}(\hat{y}, y) = \mathbb{1}_{\hat{y} \neq y}$. For any hypothesis $h \in \mathcal{H}$, one can define the loss function $\ell_h \colon (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ by taking $\ell_h(x, y) = \ell(h(x), y)$. Given some distribution D over $\mathcal{X} \times \mathcal{Y}$, one can define the expected loss of h, namely, $L_D(h) := \mathbb{E}_{(x,y)\sim D}\, \ell_h(x, y)$.

Let $\boldsymbol{S} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_m) \in (\mathcal{X} \times \mathcal{Y})^m$ be a training set of m examples. Usually the coordinates of $\boldsymbol{S}$ are assumed independent and identically distributed (iid) according to D, but we consider more general measures; this will be discussed shortly. The goal of a learning algorithm is to choose a hypothesis $\hat{h} \in \mathcal{H}$ given a sample $\boldsymbol{S}$ to (approximately) minimize the test error, $L_D(\hat{h})$. A common approach for doing so is taking the *empirical risk minimizer* (ERM), namely,

$$\hat{h}_{\mathrm{ERM}} := \arg\min_{h \in \mathcal{H}} L_{\boldsymbol{S}}(h)\ ; \quad \text{where } L_S(h) = \frac{1}{m}\sum_{i=1}^{m} \ell_h(s_i).$$

We say that S is ε-*representative* if for all $h \in \mathcal{H}$, $|L_D(h) - L_S(h)| \leq \varepsilon$. From the triangle inequality, it follows that if S is ε-representative, then

$$L_D(\hat{h}_{\mathrm{ERM}}) \leq \inf_{h \in \mathcal{H}} L_D(h) + 2\varepsilon.$$

Thus, to prove learnability, it suffices to show that $\boldsymbol{S}$ is ε-representative. Note that representativeness is stronger than learnability: it implies that *any* algorithm *generalizes*, namely, that the difference between the training and test errors, $L_D(\cdot)$ and $L_{\boldsymbol{S}}(\cdot)$, is small point-wise.

[1] For general random variables, one can define $P_{\boldsymbol{z}} = d\mu$ where $\boldsymbol{z} \sim \mu$, however, we will ignore this here. Additionally, we will assume that the density is properly defined on all the space (rather than being defined almost everywhere. Also, we assume that the conditional distributions are properly defined.

Learning from dependent samples. Instead of assuming that the samples are iid, we assume that they are drawn from a *dependent* (joint) distribution $D^{(m)}$ over $(\mathcal{X} \times \mathcal{Y})^m$, where all marginals are distributed according to the same distribution D over $\mathcal{X} \times \mathcal{Y}$. Given $\boldsymbol{S} \sim D^{(m)}$, the goal is to (approximately) minimize the test error $L_D(\hat{h})$.

Rademacher, Gaussian and $\boldsymbol{\tau}$ complexities. Given a sample $S = (s_1, \dots, s_m) \in Z^m$, a family $\mathcal{F}$ of functions from Z to $\mathbb{R}$, and a random variable $\boldsymbol{\tau} = (\boldsymbol{\tau}_1, \dots, \boldsymbol{\tau}_m)$ over $\mathbb{R}^m$, define the $\boldsymbol{\tau}$-*complexity* of $\mathcal{F}$ with respect to the sample S by:

$$\widehat{\mathfrak{D}}^{\boldsymbol{\tau}}_S(\mathcal{F}) = \mathop{\mathbb{E}}_{\boldsymbol{\tau}}\left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{\tau}_i f(s_i)\right].$$

Define the *Rademacher complexity* of $\mathcal{F}$ by $\widehat{\mathfrak{R}}_S(\mathcal{F}) := \widehat{\mathfrak{D}}^{\boldsymbol{\sigma}}_S(\mathcal{F})$ where $\boldsymbol{\sigma}$ is uniform over $\{-1, 1\}^m$, and the *Gaussian complexity* of $\mathcal{F}$ by $\widehat{\mathfrak{G}}_S(\mathcal{F}) = \widehat{\mathfrak{D}}^{\boldsymbol{g}}_S(\mathcal{F})$ where $\boldsymbol{g} \sim \mathcal{N}(0, I_m)$. Given a distribution $D^{(m)}$ over $\mathbb{R}^m$, define $\mathfrak{D}^{\boldsymbol{\tau}}_{D^{(m)}}(\mathcal{F}) = \mathbb{E}_{\boldsymbol{S} \sim D^{(m)}}\left[\widehat{\mathfrak{D}}^{\boldsymbol{\tau}}_{\boldsymbol{S}}(\mathcal{F})\right]$, and similarly define $\mathfrak{R}_{D^{(m)}}(\mathcal{F})$ and $\mathfrak{G}_{D^{(m)}}(\mathcal{F})$.

2.2. Weakly dependent distributions

We define two conditions classifying weakly dependent distributions: Dobrushin's condition and high temperature in Markov Random Fields, the first being the weakest and the last being the strongest.

2.2.1. DOBRUSHIN'S CONDITION (DOBRUSHIN, 1968)

First, one defines influences between coordinates of a random variable $\boldsymbol{z} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_m)$. The influence from $\boldsymbol{z}_i$ to $\boldsymbol{z}_j$ captures how strong the value of $\boldsymbol{z}_j$ affects the conditional distribution of $\boldsymbol{z}_i$ when all other coordinates are fixed. Formally:

Definition 3 (Influence in high dimensional distributions) *Let $\boldsymbol{z} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_m)$ be a random variable over Z^m. For $i \neq j \in \{1, \dots, m\}$, define the influence of variable $\boldsymbol{z}_j$ on variable $\boldsymbol{z}_i$ as*

$$I_{j \to i}(\boldsymbol{z}) = \max_{\substack{z_{-i-j} \in Z^{m-2} \\ z_j, z'_j \in Z}} d_{TV}\left(P_{\boldsymbol{z}_i | \boldsymbol{z}_{-i}}(\cdot \mid z_{-i-j} z_j),\ P_{\boldsymbol{z}_i | \boldsymbol{z}_{-i}}(\cdot \mid z_{-i-j} z'_j)\right),$$

where d_{TV} denotes the total variation distance.

Dobrushin's condition as defined next, certifies that a weakly dependent random vector behaves as i.i.d with respect to some important properties.

Definition 4 (Dobrushin's Uniqueness Condition) *Consider a random variable $\boldsymbol{z}$ over Z^m. Define the* Dobrushin coefficient *of $\boldsymbol{z}$ as $\alpha(\boldsymbol{z}) = \max_{1 \le i \le m} \sum_{j \neq i} I_{j \to i}(\boldsymbol{z})$. The variable $\boldsymbol{z}$ is said to satisfy Dobrushin's uniqueness condition if $\alpha(\boldsymbol{z}) < 1$.*

Note that the constant 1 is important, and for $\varepsilon > 0$ there are examples of vectors which deviate from the bound by ε and are extremely dependent. Distributions satisfying the above condition satisfy McDiarmid-like inequalities and are $O(1/(1-\alpha))$-subGaussians, as presented next.

The following result builds upon the seminal studies on concentration of measure phenomenon for contracting Markov chains by Marton et al. (1996) which is one of the first results on concentration of measure for non-product, non-Haar measures. Theorem 5 is from Külske (2003) and Chatterjee (2005a).

Theorem 5 (Concentration of Measure under Dobrushin's Condition) *Let $P^{(m)}$ be a distribution defined over Z^m satisfying Dobrushin's condition with coefficient α. Let $\boldsymbol{z} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_m) \sim P^{(m)}$ and let $f : Z^m \to \mathbb{R}$ be a real-valued function with the following property,*

$$\forall z, z' \in Z^m\colon \quad |f(z) - f(z')| \le \sum_{i=1}^{m} \mathbb{1}_{z_i \neq z'_i} \lambda_i.$$

Then, for all $t > 0$,

$$\Pr\left[|f(\boldsymbol{z}) - \mathbb{E}[f(\boldsymbol{z})]| \ge t\right] \le 2\exp\left(-\frac{(1-\alpha)t^2}{2\sum_{i=1}^{m} \lambda_i^2}\right).$$

2.2.2. MARKOV RANDOM FIELDS (MRFS) WITH PAIRWISE POTENTIALS

A common way to define a random vector is by a Markov Random Field (MRF). They are defined by potential functions, which are define the correlations between the vector entries. We will be using the definition of an MRF with pairwise potentials, as defined below:

Definition 6 (Markov Random Field (MRF) with pairwise potentials) *The random vector $\boldsymbol{z} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_m)$ over Z^m is an MRF with pairwise potentials if there exist functions $\varphi_i\colon Z \to \mathbb{R}$ and $\psi_{ij}\colon Z^2 \to \mathbb{R}$ for $i \neq j \in \{1, \dots, m\}$ such that for all $z \in Z^m$,*

$$\Pr_{\boldsymbol{z} \sim P^{(m)}}[\boldsymbol{z} = z] = \prod_{i=1}^{m} e^{\varphi_i(z_i)} \prod_{1 \le i < j \le m} e^{\psi_{ij}(z_i, z_j)}.$$

The functions φ_i are called as element-wise potentials *and ψ_{ij} are* pairwise potentials.

Analogous to Dobrushin's coefficient, one can define the inverse temperature of an MRF with pairwise potentials, where low inverse temperature implies weak correlations.

Definition 7 (High Temperature MRFs) *Given an MRF $\boldsymbol{z}$ with potentials $\{\varphi_i\}$ and $\{\psi_{ij}\}$, define*

$$\beta_{i,j}(\boldsymbol{z}) = \sup_{z_i, z_j \in Z} |\psi_{ij}(z_i z_j)|\,; \quad \beta(\boldsymbol{z}) = \max_{1 \le i \le m} \sum_{j \neq i} \beta_{ij}(P^{(m)}).$$

We say that $\boldsymbol{z}$ is high temperature *if the inverse temperature, $\boldsymbol{z}$, is less than* 1.

The inverse temperature is bounded by Dobrushin's coefficient, as presented below. The proof is a simple calculation that can be found in Chatterjee (2005a) after the statement of Theorem 3.8.

Lemma 8 *Given an MRF $\boldsymbol{z}$ with pairwise potentials, for any $i \neq j$, $I_{j \to i}(\boldsymbol{z}) \le \beta_{j,i}(\boldsymbol{z})$. Hence, $\alpha(\boldsymbol{z}) \le \beta(\boldsymbol{z})$.*

Lemma 8 implies that if the inverse temperature is less than 1, then the random variable has i.i.d-like properties. Similarly to the case with Dobrushin's condition, the smallest excess in the inverse temperature over the threshold of 1 may cause the vector to be extremely correlated.

3. Motivation and Examples

In this section, we present some tangible networked data models that would benefit from the learnability results that we prove. Consider the problem of predicting which of many possible choices a person in a social network will make: who will she vote for in a presidential election? what brand of smart phone will he buy? or what major will she study in college? Each individual's choice would, of course, be dependent on her own features; but realistically it would also depend on the choices of her friends and acquaintances. These situations are well studied, and often modeled as opinion dynamics (Montanari and Saberi, 2010), and as autoregressive models (Sacerdote, 2001). The stationary distribution of these dynamics is often a pairwise graphical model, and thus our learnability and generalization results would apply to the model if the influences are not too high.

In addition to the above, our results can potentially be applied to meteorological sensor data, since values from sensors that are geographically close are likely to be correlated. In contexts where the influences are observed to be small enough, there is scope to leverage our results. The AddHealth example cited in the introduction provides another setting of networked data where if certain covariates are weakly correlated across students, our results can be applied.

4. Agnostic Learnability of Dobrushin Dependent Data

In this Section, we study learnability under data which is weakly dependent according to Dobrushin's condition (4). We first characterize what properties of the joint distribution of our samples suffice in order to achieve learnability. Then we show that these properties hold under Dobrushin's condition.

Learnability implies the existence of a learning algorithm $\mathcal{A}$ (not necessarily efficient) such that: (a) Given a training set $\boldsymbol{S}$, $\mathcal{A}$ achieves a small training error on the training set, i.e. $L_{\boldsymbol{S}}(\mathcal{A}(\boldsymbol{S})) \leq \inf_{h\in\mathcal{H}} L_{\boldsymbol{S}}(h) + O(\varepsilon)$ and (b) the hypothesis output by $\mathcal{A}$ generalizes, i.e. $|L_{\boldsymbol{S}}(\mathcal{A}(\boldsymbol{S})) - L_D(\mathcal{A}(\boldsymbol{S}))| \leq O(\varepsilon)$ where $\lim_{m\to\infty} \varepsilon = 0$. Here we show that we can achieve the same rates of convergence (up to log factors) for the error of the learning algorithm and the confidence bounds as in the i.i.d. setting. We employ the technique of sample compression to show learnability. For simplifying the exposition of our proof, we focus on the setting where our loss function is 0/1. Our learnability result can be extended to more general loss functions as well.

A sample compression scheme is a specific type of learner which works by first carefully selecting a small subset of the training samples and then returning a hypothesis which depends only on this subset but performs well on the entire training set. The careful selection is to ensure the existence of a hypothesis which depends only on the selected small subset whole loss is minimized over the whole training set. And if the selected subset is of size $o(m)$, then we can show that any hypothesis chosen based solely on this subset will necessarily have a small generalization error. Together we get learnability. In the i.i.d. setting, for multiclass hypotheses and the 0/1 loss function, Littlestone and Warmuth (1986); David et al. (2016) show that agnostic learnability is equivalent to the existence of a sublinear size sample compression scheme. We extend this result to the setting of Dobrushin dependent data achieving nearly the same asymptotic rates as in the i.i.d. setting.

Definition 9 (Agnostic Sample Compression Scheme) *Fix a hypothesis class $\mathcal{H}$, integers $0 < k < m$ and functions $\kappa\colon (\mathcal{X}\times\mathcal{Y})^m \to (\mathcal{X}\times\mathcal{Y})^k$ and $\rho\colon (\mathcal{X}\times\mathcal{Y})^k \to \mathcal{Y}^{\mathcal{X}}$. We say that (κ, ρ) is an agnostic sample compression scheme for $\mathcal{H}$ of size k with respect to a sample-size m if:*

- *For all samples S, $\kappa(S) \subseteq S$ and $|\kappa(S)| \leq k$.*

- *For all samples S, $L_S(\rho(\kappa(S))) \leq \inf_{h \in \mathcal{H}} L_S(h)$.*

To understand what hypothesis classes $\mathcal{H}$ have small sample compression schemes one can look at the instructive setting of binary hypothesis classes, i.e. $\mathcal{Y} = \{0,1\}$. For these classes, it is known that having a small VC-dimension is equivalent to having a small compression scheme Moran and Yehudayoff (2016).

Our main result of this Section is Theorem 10 which states that hypothesis classes with sample compression schemes of size k are agnostic PAC-learnable to error ε from $\widetilde{O}(k/\varepsilon^2)$ samples.

Theorem 10 (Agnostic PAC-Learning for Compressible Hypothesis Classes) *Let $\mathcal{H}$ be a hypothesis class with a sample compression scheme (κ, ρ) of size k and let ℓ denote the 0/1 loss function. Given a sample $\boldsymbol{S} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_m, \boldsymbol{y}_m)\} \sim D^{(m)}$ where $D^{(m)}$ satisfies Dobrushin's condition with coefficient α, there exists a constant C such that,*

$$\Pr\left[L_D(\rho(\kappa(\boldsymbol{S}))) \geq \inf_{h \in \mathcal{H}} L_D(h) + \varepsilon\right] \leq \delta, \quad \textit{for} \quad m = \frac{Ck\log(k/\varepsilon^2) + \log(1/\delta)}{(1-\alpha)\varepsilon^2}.$$

Due to the VC dimension characterizing compressibility for binary hypothesis classes, we get Corollary 11 from Theorem 10.

Corollary 11 (Agnostic PAC-Learning for Finite VC-Dimension Classes) *Let $\mathcal{H}$ be a binary hypothesis class with $VC(\mathcal{H}) = d$, and let ℓ be the 0/1 loss function. Given a sample $\boldsymbol{S} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_m, \boldsymbol{y}_m)\} \sim D^{(m)}$ where $D^{(m)}$ satisfies Dobrushin's condition with Dobrushin coefficient α, we have for some constant C and for $k = \min\left(d\log m, d2^{(d+1)}\right)$,*

$$\Pr\left[L_D(\rho(\kappa(\boldsymbol{S}))) \geq \inf_{h \in \mathcal{H}} L_D(h) + \varepsilon\right] \leq \delta, \quad \textit{for} \quad m = \frac{Ck\log(k/\varepsilon^2) + \log(1/\delta)}{(1-\alpha)\varepsilon^2}.$$

The proof of Theorem 10 proceeds in two steps. The first step is showing that any compression scheme of size $k = o(m)$ will generalize, i.e. $L_S(\rho(\kappa(\boldsymbol{S}))) - L_D(\rho(\kappa(\boldsymbol{S})))$ is small. This is the crucial part of the proof which differs significantly from the i.i.d. case. The second step follows from the definition of a valid compression scheme that it must achieve optimal training error: $L_S(\rho(\kappa(\boldsymbol{S}))) \leq \inf_{h \in \mathcal{H}} L_S(h)$ combined with a tail bound on $\inf_{h \in \mathcal{H}} L_S(h) - \inf_{h \in \mathcal{H}} L_D(h)$.

To show the first step, we first present a general Lemma 12 which states conditions on the data distribution which are sufficient to show that sample compression schemes generalize.

Lemma 12 (Conditions for Generalization of Sample Compression Schemes) *Consider a sample $\boldsymbol{S} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_m, \boldsymbol{y}_m)\} \sim D^{(m)}$ and any loss function ℓ bounded by $R \geq 0$. For any subset of indices $I \subseteq [m]$, let $\boldsymbol{S}_I = \{(x_i, y_i) \colon i \in I\}$. If we have that for any $I \subseteq [m]$, and for constants C_1 and C_2,*

$$\mathit{1.}\ \Pr\left[|L_{\boldsymbol{S}}(h) - L_D(h)| \geq t \;\middle|\; \boldsymbol{S}_I\right] \leq 2\exp\left(-\frac{t^2 m}{2C_1R^2}\right), \quad \mathit{2.}\ |\mathbb{E}\left[L_{\boldsymbol{S}}(h)\right] - \mathbb{E}\left[L_{\boldsymbol{S}}(h)|\boldsymbol{S}_I\right]| \leq \frac{C_2R|I|}{m},$$

then, for any agnostic sample compression scheme (κ, ρ) of size k on $\boldsymbol{S}$, for some constant C,

$$\Pr\left[|L_{\boldsymbol{S}}(\rho(\kappa(\boldsymbol{S}))) - L_D(\rho(\kappa(\boldsymbol{S})))| \geq CR\sqrt{\frac{(k\log m + \log(1/\delta))}{m}}\right] \leq \delta.$$

Then, we have Lemma 13 which shows that if the data distribution is Dobrushin, the conditions of Lemma 12 are satisfied.

Lemma 13 *Let $D^{(m)}$ be a distribution over m variables which satisfies Dobrushin's condition with coefficient α such that the marginal of eveey variable is D. Let $\boldsymbol{S} \sim D^{(m)}$. Then we have*

1. $\Pr\left[|L_{\boldsymbol{S}}(h) - L_D(h)| \geq t \;\middle|\; \boldsymbol{S}_I\right] \leq 2\exp\left(-\frac{t^2 m(1-\alpha)}{2R^2}\right),$
2. $|\mathbb{E}\left[L_{\boldsymbol{S}}(h)\right] - \mathbb{E}\left[L_{\boldsymbol{S}}(h)|\boldsymbol{S}_I\right]| \leq \frac{\alpha R|I|}{(1-\alpha)m},$

Lemmas 12 and 13 together with the property of an agnostic sample compression scheme imply Theorem 10.

5. Uniform Convergence for Weakly Dependent Data

In this section, we obtain uniform convergence bounds for weakly dependent distributions. We could not derive such bounds for distributions satisfying Dobruhsin's condition and we do not know if such bounds apply for all classes of finite VC dimension. Instead, we present such bounds for a smaller family of distributions, which contains high temperature Markov Random Fields with pairwise potentials. Moreover this family allows for an arbitrary structure of correlations (as long as they are sufficiently weak). We define the log-influences $I^{\log}_{j,i}$, a notion stronger than Dobrushin's influences $I_{j\to i}$, which replaces the total variation distance appearing in the definition with a stronger bound on the maximal log-ratio of probabilities. Analogously, we obtain the log-coefficient $\alpha_{\log}$, as defined below:

Definition 14 (Log-influence and log-coefficient) *Let $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m)$ be a random variable over Ω^m and let $P_{\boldsymbol{z}}$ denote either its probability distribution if discrete or its density if continuous. Assume that $P_z > 0$ on all Ω^m. For any $i \neq j \in [m]$, define the* log-influence *between j and i as*[2]

$$I^{\log}_{j,i}(\boldsymbol{z}) = \frac{1}{4} \sup_{\substack{z_{-i-j}\in\Omega^{m-2}\\ z_i, z_i', z_j, z_j' \in \Omega}} \log \frac{P_{\boldsymbol{z}}[z_i z_j z_{-i-j}] P_{\boldsymbol{z}}[z_i' z_j' z_{-i-j}]}{P_{\boldsymbol{z}}[z_i' z_j z_{-i-j}] P_{\boldsymbol{z}}[z_i z_j' z_{-i-j}]}.$$

Define the log-coefficient *of $\boldsymbol{z}$ as $\alpha_{\log}(\boldsymbol{z}) = \max_{i\in[m]} \sum_{j\neq i} I^{\log}_{j,i}(\boldsymbol{z})$.*

Note that the log influence is symmetric: $I^{\log}_{j,i} = I^{\log}_{i,j}$. The following relation holds:

Lemma 15 *For any random variable $\boldsymbol{z}$ and $i, j \in [m]$, $I_{j\to i}(\boldsymbol{z}) \leq I^{\log}_{j,i}(\boldsymbol{z}) \leq \beta_{j,i}(\boldsymbol{z})$.*

The main result of this section shows that uniform convergence holds whenever the log-coefficient is less than a half. In that regime, the maximal generalization error of hypotheses from H, $\sup_{h\in H} |L_S(h) - L_D(h)|$, is bounded in terms of the Gaussian complexity of H:

[2] To be more formal, one can define $P_{\boldsymbol{z}} = d\mu$ where $\boldsymbol{z} \sim \mu$ and replace the supremum with an essential supremum.

Theorem 16 *Let $\mathcal{H}$ be a hypothesis class, let $\ell\colon \mathcal{Y}^2 \to [-L, L]$ be a loss function, and let $\mathcal{L}_H = \{\ell_h\colon h \in H\}$. Let $D^{(m)}$ be a distribution over $(X \times Y)^m$, with all m marginals equaling D and $\alpha_{\log}(D^{(m)}) < 1/2$. Then, for all $t > 0$,*

$$\Pr_{\boldsymbol{S}\sim D^{(m)}}\left[\sup_{h\in H}|L_S(h) - L_D(h)| > C\left(\mathfrak{G}_{D^{(m)}}(\mathcal{L}_H) + \frac{Lt}{\sqrt{m}}\right)\right] \leq e^{-t^2/2},$$

(where C is a universal constant whenever $1/2 - \alpha_{\log}\left(D^{(m)}\right)$ is bounded away from zero).

The proof of Theorem 16 is a direct corollary of Theorem 17 which is presented below. Note that Lemma 15 implies that Theorem 16 also holds whenever $D^{(m)}$ is an MRF with pairwise potentials and $\beta(D^{(m)}) < 1/2$. Since the condition $\beta(D^{(m)}) < 1$ sufficient for concentration inequalities to hold, we suspect that Theorem 16 may hold as well in this regime. However,

Applying Theorem 16 on any hypothesis class H with finite VC, one obtains the same sample complexity bounds of i.i.d data up to constant factors:

$$O\left(\frac{\mathrm{VC}(H) + \log(1/\delta)}{\varepsilon^2}\right). \tag{2}$$

This follows from the fact that the Gaussian complexity of $\mathcal{L}_H$ is bounded by $O\left(\sqrt{VC(H)/m}\right)$. The proof is almost identical to the proof bounding the Rademacher complexity by the same quantity (see, for instance, Shalev-Shwartz and Ben-David, 2014, Chapter 27).

Although the Rademacher and Gaussian complexities are not identical, they are almost equivalent. Tomczak-Jaegermann (1989) proved the following for some universal constants $c, C > 0$.

$$c\widehat{\mathfrak{R}}_S(\mathcal{F}) \leq \widehat{\mathfrak{G}}_S(\mathcal{F}) \leq C \ln m\, \widehat{\mathfrak{R}}_S(\mathcal{F}),$$

Theorem 16 is based on a more general result, bounding the expected suprema of empirical processes with respect to the corresponding Gaussian complexity. Here, the supremum is taken over an arbitrary family of unbounded functions, rather than bounded loss functions.

Theorem 17 *Let $D^{(m)}$ be a random vector over some domain Z^m and let $\mathcal{F}$ be a class of functions from Z to $\mathbb{R}$. If $\alpha_{\log}(D^{(m)}) < 1/2$, then*

$$\mathop{\mathbb{E}}_{\boldsymbol{S}\sim D^{(m)}}\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{s_i}) - \mathop{\mathbb{E}}_{\boldsymbol{S}}\left[\frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{s_i})\right]\right) \leq \frac{C\mathfrak{G}_{D^{(m)}}(\mathcal{F})}{\sqrt{1 - 2\alpha_{\log}(D^{(m)})}}, \tag{3}$$

where $C > 0$ is a universal constant.

The proof outline appears in Section 6. Theorem 16 follows from Theorem 17 simply by applying a McDiarmind-like inequality for weakly correlated data (Theorem 5).

6. Proof Outline: Uniform Convergence

First, we bound the left hand side of (3) by the $\boldsymbol{\sigma}$-complexity of $\mathcal{F}$ (Eq. (5)), where $\boldsymbol{\sigma}$ does not consist of i.i.d random signs, but rather it is a subGaussian distribution with zero mean. Furthermore, this $\boldsymbol{\sigma}$-complexity is not with respect to $D^{(m)}$ but rather with respect to a different distribution. Then, we bound this $\boldsymbol{\sigma}$-complexity by the Gaussian complexity of $\mathcal{F}$, with respect to $D^{(m)}$ (Lemma 18 and Lemma 19).

Assume that $\boldsymbol{S} = (\boldsymbol{s}_i)_{i\in[m]} \sim D^{(m)}$, and $\boldsymbol{S'} = (\boldsymbol{s}'_i)_{i\in[m]}$ is another i.i.d. random variable drawn from $D^{(m)}$. The following holds:

$$\mathop{\mathbb{E}}_{\boldsymbol{S}} \sup_{f\in\mathcal{F}} \left(\frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{s}_i) - \mathop{\mathbb{E}}_{\boldsymbol{S}} \frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{s}_i) \right) \quad \le \quad \mathop{\mathbb{E}}_{\boldsymbol{S},\boldsymbol{S'}} \sup_{f\in\mathcal{F}} \left(\frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{s}_i) - \frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{s}'_i) \right). \tag{4}$$

We randomly shuffle $\boldsymbol{S}$ and $\boldsymbol{S}'$, to create samples $\boldsymbol{T}$ and $\boldsymbol{T}'$. Formally, m i.i.d and uniform random signs are drawn, $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_m) \in \{-1, 1\}^m$. Then, $\boldsymbol{T} = (\boldsymbol{t}_1, \dots, \boldsymbol{t}_m)$ and $\boldsymbol{T}' = (\boldsymbol{t}'_1, \dots, \boldsymbol{t}'_m)$ are defined as functions of $\boldsymbol{S}$, $\boldsymbol{S}'$ and $\boldsymbol{\sigma}$, as follows: for any $i \in \{1, \dots, m\}$, if $\boldsymbol{\sigma}_i = 1$ then $\boldsymbol{t}_i = \boldsymbol{s}_i$ and $\boldsymbol{t}'_i = \boldsymbol{s}'_i$, and otherwise, $\boldsymbol{t}_i = \boldsymbol{s}'_i$ and $\boldsymbol{t}'_i = \boldsymbol{s}_i$.

For any T and T' denote by $\boldsymbol{\sigma}_{T,T'}$ a random variable sampled from $P_{\boldsymbol{\sigma}|\boldsymbol{TT'}}(\cdot \mid T, T')$, the conditional distribution of $\boldsymbol{\sigma}$, conditioned on $\boldsymbol{T} = T$ and $\boldsymbol{T}' = T'$. We bound the right hand side of (4), substituting $\boldsymbol{S}$ and $\boldsymbol{S'}$ with $\boldsymbol{T}$ and $\boldsymbol{T}'$, in a *change of measure* argument:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{S},\boldsymbol{S'}} \left[\sup_{f\in\mathcal{F}} \left(\frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{s}_i) - \frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{s}'_i) \right) \right] &= \mathop{\mathbb{E}}_{\boldsymbol{T},\boldsymbol{T'},\boldsymbol{\sigma}} \left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\sigma}_i \left(f(\boldsymbol{t}_i) - f(\boldsymbol{t}'_i) \right) \right] \\
&\le \mathop{\mathbb{E}}_{\boldsymbol{T},\boldsymbol{T'},\boldsymbol{\sigma}} \left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\sigma}_i f(\boldsymbol{t}_i) \right] + \mathop{\mathbb{E}}_{\boldsymbol{T},\boldsymbol{T'},\boldsymbol{\sigma}} \left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\sigma}_i (-f(\boldsymbol{t}'_i)) \right] \\
&=_{(*)} 2 \mathop{\mathbb{E}}_{\boldsymbol{T},\boldsymbol{T'},\boldsymbol{\sigma}} \left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\sigma}_i f(\boldsymbol{t}_i) \right] = 2 \mathop{\mathbb{E}}_{\boldsymbol{T},\boldsymbol{T'}} \widehat{\mathfrak{D}}_{\boldsymbol{T}}^{\boldsymbol{\sigma_{T,T'}}}(\mathcal{F}),
\end{aligned} \tag{5}$$

where the equality $(*)$ follows from the fact that the joint distribution of $\boldsymbol{T}$ and $\boldsymbol{\sigma}$ equals the joint distribution of $\boldsymbol{T}'$ and $-\boldsymbol{\sigma}$. Note that $\boldsymbol{\sigma}_{T,T'}$ is generally not a product distribution, however, we can show that it is a zero mean subGaussian.

If follows from Lemma15 that $\alpha(\boldsymbol{\sigma}_{T,T'}) \le \alpha^{\log}(\boldsymbol{\sigma}_{T,T'}) \le 2\alpha^{\log}(D^{(m)})$. From Theorem 5 it follows that $\boldsymbol{\sigma}_{T,T'}$ is a $C/(1-\alpha(\boldsymbol{\sigma}_{T,T'}))$-subGaussian, hence it is a $C/(1-2\alpha^{\log}(D^{(m)}))$-subGaussian. We use this to show that the $\boldsymbol{\sigma}_{T,T'}$ complexity of $\mathcal{F}$ can be bounded in terms of the Gaussian complexity. This will bound the right hand side of (5). The proof follows from Fernique-Talagrand Majorizing measure theory.

Lemma 18 *Fix $z \in Z^m$. If $\boldsymbol{\tau}$ is a K^2-subgaussian, then, $\widehat{\mathfrak{D}}_z^{\boldsymbol{\tau}}(\mathcal{F}) \le CK\widehat{\mathfrak{G}}_z(\mathcal{F})$, (for some universal constant $C > 0$). In particular, $\widehat{\mathfrak{D}}_z^{\boldsymbol{\sigma}_{T,T'}}(\mathcal{F}) \le C\widehat{\mathfrak{G}}_z(\mathcal{F})/\sqrt{1-2\alpha^{\log}(D^{(m)})}$.*

Lemma 18 implies that the right hand side of (5) is bounded as follows:

$$\mathop{\mathbb{E}}_{\boldsymbol{T},\boldsymbol{T'}} \widehat{\mathfrak{D}}_{\boldsymbol{T}}^{\boldsymbol{\sigma_{T,T'}}}(\mathcal{F}) \le \frac{C}{\sqrt{1-2\beta(D^{(m)})}} \mathop{\mathbb{E}}_{\boldsymbol{T},\boldsymbol{T'}} \widehat{\mathfrak{G}}_{\boldsymbol{T}}(\mathcal{F}) = \frac{C}{\sqrt{1-2\beta(D^{(m)})}} \mathfrak{G}_{\boldsymbol{T}}(\mathcal{F}). \tag{6}$$

We will bound this last term by the Gaussian complexity of $D^{(m)}$.

Lemma 19 *The following holds: $\mathfrak{G}_{\boldsymbol{T}}(\mathcal{F}) \le 2\mathfrak{G}_{D^{(m)}}(\mathcal{F})$.*

The proof concludes by equations (4), (5), (6) and Lemma 19.

References

Alekh Agarwal and John C Duchi. The generalization ability of online algorithms for dependent data. *IEEE Transactions on Information Theory*, 59(1):573–587, 2013.

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

Witold Bednorz and Rafal Latala. On the boundedness of bernoulli processes. *Annals of Mathematics*, 180(3):1167–1203, 2014.

Patrizia Berti, Irene Crimaldi, Luca Pratelli, Pietro Rigo, et al. Rate of convergence of predictive distributions for dependent data. *Bernoulli*, 15(4):1351–1367, 2009.

Marianne Bertrand, Erzo FP Luttmer, and Sendhil Mullainathan. Network effects and welfare cultures. *The Quarterly Journal of Economics*, 115(3):1019–1055, 2000.

Yann Bramoullé, Habiba Djebbari, and Bernard Fortin. Identification of peer effects through social networks. *Journal of econometrics*, 150(1):41–55, 2009.

Sourav Chatterjee. *Concentration Inequalities with Exchangeable Pairs*. PhD thesis, Stanford University, June 2005a.

Sourav Chatterjee. Concentration inequalities with exchangeable pairs (ph. d. thesis). *arXiv preprint math/0507526*, 2005b.

Nicholas A Christakis and James H Fowler. Social contagion theory: examining dynamic social networks and human behavior. *Statistics in medicine*, 32(4):556–577, 2013.

Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Testing Ising models. In *Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, Philadelphia, PA, USA, 2018. SIAM.

Ofir David, Shay Moran, and Amir Yehudayoff. On statistical learning via the lens of compression. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 2792–2800. Curran Associates Inc., 2016.

PL Dobrushin. The description of a random field by means of conditional probabilities and conditions of its regularity. *Theory of Probability & Its Applications*, 13(2):197–224, 1968.

RL Dobrushin and SB Shlosman. Completely analytical interactions: constructive description. *Journal of Statistical Physics*, 46(5-6):983–1014, 1987.

Esther Duflo and Emmanuel Saez. The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment. *The Quarterly journal of economics*, 118 (3):815–842, 2003.

Yoav Freund. Boosting a weak learning algorithm by majority. *Information and computation*, 121 (2):256–285, 1995.

Reza Gheissari, Eyal Lubetzky, and Yuval Peres. Concentration inequalities for polynomials of contracting Ising models. *arXiv preprint arXiv:1706.00121*, 2017.

Edward L Glaeser, Bruce Sacerdote, and Jose A Scheinkman. Crime and social interactions. *The Quarterly Journal of Economics*, 111(2):507–548, 1996.

Kathleen Mullan Harris, National Longitudinal Study of Adolescent Health, et al. Waves i & ii, 1994–1996; wave iii, 2001–2002; wave iv, 2007–2009 [machine-readable data file and documentation]. *Chapel Hill, NC: Carolina Population Center, University of North Carolina at Chapel Hill*, 10, 2009.

Christof Külske. Concentration inequalities for functions of gibbs fields with application to diffraction and random gibbs measures. *Communications in mathematical physics*, 239(1-2):29–51, 2003.

H Künsch. Decay of correlations under dobrushin's uniqueness condition and its applications. *Communications in Mathematical Physics*, 84(2):207–222, 1982.

Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for time series prediction with non-stationary processes. In Peter Auer, Alexander Clark, Thomas Zeugmann, and Sandra Zilles, editors, *Algorithmic Learning Theory*, pages 260–274, Cham, 2014. Springer International Publishing. ISBN 978-3-319-11662-4.

Vitaly Kuznetsov and Mehryar Mohri. Learning theory and algorithms for forecasting non-stationary time series. In *Advances in neural information processing systems*, pages 541–549, 2015.

Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106(1):93–117, Jan 2017. doi: 10.1007/s10994-016-5588-2. URL https://doi.org/10.1007/s10994-016-5588-2.

Tianxi Li, Elizaveta Levina, and Ji Zhu. Prediction models for network-linked data. *arXiv preprint arXiv:1602.01192*, 2016.

Nick Littlestone and Manfred Warmuth. Relating data compression and learnability. 1986.

Charles F Manski. Identification of endogenous social effects: The reflection problem. *The review of economic studies*, 60(3):531–542, 1993.

Katalin Marton et al. Bounding $\bar{d}$-distance by informational divergence: A method to prove measure concentration. *The Annals of Probability*, 24(2):857–866, 1996.

Daniel J. McDonald and Cosma Rohilla Shalizi. Rademacher complexity of stationary sequences. *arXiv preprint arXiv:1106.0730*, 2017.

Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-iid processes. In *Advances in Neural Information Processing Systems*, pages 1097–1104, 2009.

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

Andrea Montanari and Amin Saberi. The spread of innovations in social networks. *Proceedings of the National Academy of Sciences*, 107(47):20196–20201, 2010. ISSN 0027-8424. doi: 10.1073/pnas.1004098107. URL https://www.pnas.org/content/107/47/20196.

Shay Moran and Amir Yehudayoff. Sample compression schemes for vc classes. *Journal of the ACM (JACM)*, 63(3):21, 2016.

Vladimir Pestov. Predictive pac learnability: A paradigm for learning from exchangeable input data. In *Granular Computing (GrC), 2010 IEEE International Conference on*, pages 387–391. IEEE, 2010.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Random averages, combinatorial parameters, and learnability. In *Advances in Neural Information Processing Systems*, pages 1984–1992, 2010.

Bruce Sacerdote. Peer effects with random assignment: Results for dartmouth roommates. *The Quarterly journal of economics*, 116(2):681–704, 2001.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Daniel W Stroock and Boguslaw Zegarlinski. The logarithmic sobolev inequality for discrete spin systems on a lattice. *Communications in Mathematical Physics*, 149(1):175–193, 1992.

Michel Talagrand et al. Majorizing measures: the generic chaining. *The Annals of Probability*, 24 (3):1049–1103, 1996.

Nicole Tomczak-Jaegermann. *Banach-Mazur distances and finite-dimensional operator ideals*, volume 38. Longman Sc & Tech, 1989.

Justin G Trogdon, James Nonnemaker, and Joanne Pais. Peer effects in adolescent overweight. *Journal of health economics*, 27(5):1388–1399, 2008.

Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer, 2015.

Dror Weitz. Combinatorial criteria for uniqueness of gibbs measures. *Random Structures & Algorithms*, 27(4):445–475, 2005.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994.