Learning from weakly dependent data under Dobrushin's condition

Yuval Dagan* Constantinos Daskalakis† Nishanth Dikkala‡
Siddhartha Jayanti§

June 24, 2019

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

*Massachusetts Institute of Technology, EE&CS, dagan@mit.edu
†Massachusetts Institute of Technology, EE&CS, costis@csail.mit.edu
‡Massachusetts Institute of Technology, EE&CS, nishanthd@csail.mit.edu
§Massachusetts Institute of Technology, EE&CS, jayanti@mit.edu

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. Christakis and Fowler, 2013, Trogdon et al., 2008). 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, socioeconomic 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 7.

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 2.2. 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 [Chatterjee, 2005b, Daskalakis et al., 2018, Gheissari et al., 2017, Külske, 2003, Marton et al., 1996], 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 4.5, 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 4.6). The dependence on the error and confidence in our bounds match those of the i.i.d. setting up to logarithmic factors in the size of the training set. We focus on hypothesis classes which have small sample compression schemes in our paper. This property is known to imply learnability in the i.i.d. setting and we will show learnability for these classes under Dobrushin distributed data as well. We provide an informal statement of our learnability result applied to finite VC dimension hypothesis classes here.

Informal Theorem 1.1 (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*

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

Our second main result, presented as Theorem 5.3, 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 1.2 (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 $O(\sqrt{d/n})$.*

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 2 we state some preliminary notation and definitions and some lemmas from prior work that we use throughout the paper. Section 3 contains motivating examples for learning from data that satisfies Dobrushin's condition. Section 4 contains the learnability results for data satisfying Dobrushin's condition. Section 5 contains the uniform convergence bound for data satisfying the stronger condition and the proofs for this section appear in Section 6. Section 7 provides 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.

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 Kuznetsov and Mohri [2015], Mohri and Rostamizadeh [2009, 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 7.

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.

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,

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

[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.

If the sample $\boldsymbol{S}$ "represents well" the test distribution D, then the learned hypothesis only suffers low error. To be precise, 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 the ERM is 2ε-optimal with respect to $\mathcal{H}$, namely

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

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.

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

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

Define the *Rademacher complexity* of $\mathcal{F}$ by $\widehat{\mathfrak{R}}_S(\mathcal{F}) := \widehat{\mathfrak{O}}_S^{\boldsymbol{\sigma}}(\mathcal{F})$ where $\boldsymbol{\sigma}$ is uniform over $\{-1, 1\}^m$, and the *Gaussian complexity* of $\mathcal{F}$ by $\widehat{\mathfrak{G}}_S(\mathcal{F}) = \widehat{\mathfrak{O}}_S^{\boldsymbol{g}}(\mathcal{F})$ where $\boldsymbol{g} \sim \mathcal{N}(0, I_m)$. Given a distribution $D^{(m)}$ over $\mathbb{R}^m$, define $\mathfrak{O}_{D^{(m)}}^{\boldsymbol{\tau}}(\mathcal{F}) = \mathbb{E}_{\boldsymbol{S} \sim D^{(m)}} \left[\widehat{\mathfrak{O}}_{\boldsymbol{S}}^{\boldsymbol{\sigma}}(\mathcal{F}) \right]$, 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 $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m)$. 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 2.1 (Influence in high dimensional distributions)**.** *Let $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m)$ be a random variable over Z^m. For $i \neq j \in \{1, \ldots, m\}$, 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 2.2 (Dobrushin's Uniqueness Condition)**.** *Consider a random variable $\boldsymbol{z}$ over Z^m. Define the* Dobrushin coefficient *of $\boldsymbol{z}$ as $\alpha(\boldsymbol{z}) = \max_{1 \leq i \leq m} \sum_{j \neq i} I_{j \to i}(\boldsymbol{z})$. 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 2.3 is from Külske [2003] and Chatterjee [2005a].

Theorem 2.3 (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 $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m) \sim P^{(m)}$ and let $f : Z^m \to \mathbb{R}$ be a real-valued function with the following bounded differences property, with parameters $\lambda_1, \ldots, \lambda_m \geq 0$:*

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

Then, for all $t > 0$,

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

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 2.4 (Markov Random Field (MRF) with pairwise potentials)**.** *The random vector $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m)$ 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 $i \neq j \in \{1, \ldots, m\}$ such that for all $z \in Z^m$,*

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

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 2.5 (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 2.6. *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 2.6 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 following is a natural opinion dynamics for a binary decision (for instance, voting Democrat versus Republican), given the graph of a friend network:

1. each individual starts with an initial preference (Democrat or Republican)
2. and at each time step, a random individual stochastically updates his preference conditioned on the current preferences of his friend

The stationary distribution of these dynamics is a pairwise graphical model, and thus our learnability and generalization results would apply to the model if the influences are not too high.

Another prominent econometric model for peer effects hinges on autoregression. The standard linear regressive stochastic process is described by the equation $Y = XB + E$. This equation, states that the $n \times d$ feature matrix X of n samples with d features each has a linear relationship (given by B) to the response variables coded as the $n \times 1$ response vector Y, up to a small stochastic error E (dimension $n \times 1$). The linear *autoregressive* process is described by the equation $Y = XB + E + AY$, where the vector of responses Y appears on both sides of the equation. Thus,

the model ultimately states that the responses are dependent linearly on the features and other responses. Given a training set satisfying these equations, the task is to predict the response y_{n+1} for a new sample x_{n+1}. This problem falls under our model when the auto-regression matrix A has entries that are sufficiently small. Sacerdote uses this type of autoregressive model to analyze relationships between college roommate assignment and academic achievement [Sacerdote, 2001]. Our results can potentially help in predictive analysis for the same question.

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 a setting of networked data where if certain covariates are weakly correlated across students, then one can obtain generalization bounds using our theory for prediction and regression tasks.

4 Agnostic Learnability of Dobrushin Dependent Data

In this Section, we study learnability under data which is weakly dependent. Qualitatively, learnability implies the existence of a learning algorithm (not necessarily efficient) whose error goes to 0 with confidence going to 1 as the sample size increases. An algorithm $\mathcal{A}$ can be shown to be a learner by showing two properties: (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. 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 using techniques described in Section 4.1.

We first characterize certian properties of the joint distribution of our samples which we show are sufficient to achieve learnability. Then we show that these properties hold under Dobrushin's condition (2.2). For binary hypothesis classes, in the i.i.d. setting, it is well-known that having a finite VC-dimension is equivalent to learnability. For more general hypothesis classes under the 0/1 loss function, Moran and Yehudayoff [2016] show that learnability is equivalent to having a finite size sample compression scheme. In our work, we employ the technique of sample compression to understand learnability, generalizing the results of Moran and Yehudayoff [2016] to the setting with dependent data (Theorem 4.5). This generalization immediately also implies a generalization of the result for binary hypothesis classes (Corollary 4.6).

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, David et al. [2016], Littlestone and Warmuth [1986] 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.

To proceed formalizing the discussion above, we begin with the definition of a sample compression scheme for a certain hypothesis class $\mathcal{H}$ in the general agnostic setting. It consists of two functions: a compressor κ which carefully subsamples the training set and a reconstructor ρ which outputs a hypothesis based on this subsample chosen by the compressor. The compressor's job is to select a subsample of the training set such that it allows the reconstructor to output a hypothesis which attains optimal loss on the entire training set. Intuitively, the 'simpler' the true underlying function of the data, the compressor should be able to compress the training set to a smaller size. We give a formal definition of sample compression schemes below.

Definition 4.1 (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 the following hold:*

- *For all samples S, $\kappa(S) \subseteq S$.*
- *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. $\mathcal{Y} = \{0, 1\}$. As we point out in Theorems 4.2 and 4.3, having a small VC-dimension is equivalent to having a small compression scheme.

Theorem 4.2 (Folklore)**.** *If a class $\mathcal{H}$ has a sample compression scheme of size $d \in \mathbb{N}$ then $\mathrm{VC}(\mathcal{H}) = O(d)$.*

Theorem 4.3 (Theorem 1.3 of Moran and Yehudayoff [2016])**.** *Any class of VC dimension d has a sample compression scheme of size $O\left(d2^{(d+1)}\right)$.*

For many reasonable values of the sample size m, Theorem 4.4 gives a better guarantee than Theorem 4.3.

Theorem 4.4 (Freund [1995])**.** *Any class of VC dimension d has a sample compression scheme of size $O(d \log m)$.*

Our main result of this Section is Theorem 4.5 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 4.5 (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* $C(\alpha)$ *such that,*

$$\Pr\left[L_D(\rho(\kappa(\boldsymbol{S}))) \geq \inf_{h\in\mathcal{H}} L_D(h) + \varepsilon\right] \leq \delta,$$

$$\textit{for} \quad m = \frac{C(\alpha)k\log(k/\varepsilon^2) + \log(1/\delta)}{(1-\alpha)\varepsilon^2}.$$

Given Theorem 4.5, Theorems 4.3 and 4.4 immediately give Corollary 4.6.

Corollary 4.6 (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*

$$\Pr\left[L_D(\rho(\kappa(\boldsymbol{S}))) \geq \inf_{h\in\mathcal{H}} L_D(h) + \varepsilon\right] \leq \delta,$$

$$\textit{for} \quad m = \frac{C(\alpha)k\log(k/\varepsilon^2) + \log(1/\delta)}{(1-\alpha)\varepsilon^2}$$

for some constant $C(\alpha)$ *and for* $k = \min\left(d\log m, d2^{(d+1)}\right)$.

The proof of Theorem 4.5 proceeds in three steps. Our ultimate goal is to bound the quantity $L_D(\rho(\kappa(\boldsymbol{S}))) - \inf_{h\in\mathcal{H}} L_D(h)$.

1. The first step outlines conditions which suffice to show that any compression scheme of size $k = o(m)$ will generalize, i.e. $L_D(\rho(\kappa(\boldsymbol{S}))) - L_{\boldsymbol{S}}(\rho(\kappa(\boldsymbol{S})))$ is small. We believe the identification of sufficient conditions for generalization in Lemma 4.7 could lead to the study of learnability under other types of dependencies.

2. The second step involves showing that Dobrushin's condition implies the pre-conditions of Lemma 4.7 yielding the conclusion that sample compression schemes for Dobrushin distributed data generalize. These two steps are the crucial parts of the proof which differ significantly from the i.i.d. case.

3. The third and final step is showing that $L_{\boldsymbol{S}}(\rho(\kappa(\boldsymbol{S}))) - \inf_{h\in\mathcal{H}} L_D(h)$ is small. This follows from (a) the definition of a valid compression scheme that it must achieve optimal training error: $L_S(\rho(\kappa(\boldsymbol{S}))) \leq \inf_{h\in\mathcal{H}} L_S(h)$; (b) a tail bound on $\inf_{h\in\mathcal{H}} L_S(h) - \inf_{h\in\mathcal{H}} L_D(h)$.

To simplify notation we will refer to the Dobrushin coefficient $\alpha\left(D^{(m)}\right)$ as just α in the proof.

To show the first step, we first present Lemma 4.7.

Lemma 4.7 (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,

1. $\left|\mathbb{E}\left[L_{\boldsymbol{S}}(h)\right] - \mathbb{E}\left[L_{\boldsymbol{S}}(h)|\boldsymbol{S}_I\right]\right| \leq \dfrac{C_2 R|I|}{m}$,
2. $\Pr\left[\left|L_{\boldsymbol{S}}(h) - \mathbb{E}[L_{\boldsymbol{S}}(h)|\boldsymbol{S}_I]\right| \geq t \;\middle|\; \boldsymbol{S}_I\right] \leq 2\exp\left(-\dfrac{t^2 m}{2C_1 R^2}\right)$,

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

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

Proof. Given a sample of size m, any compression scheme (κ, ρ) of size k can select one among $\binom{m}{k}$ choices of subsets of $\boldsymbol{S}$ of size k. The total number of different choices κ can make are at most $\binom{m}{k} \leq m^k$. For any set of indices $I \subseteq [m]$, where $|I| = k$, define $\hat{h}_I$ as follows: given a sample $\boldsymbol{S}$, let $\boldsymbol{S}_I = \{(x_i, y_i) \colon i \in I\}$. Then, $\hat{h}_I = \rho(\boldsymbol{S}_I)$. We will show that for each $I \subset [m]$, $\hat{h}_I$ generalizes. That is, we will show that for any $t > 0$ and any I

$$\Pr\left[\left|L_{\boldsymbol{S}}(\hat{h}_I) - L_D(\hat{h}_I)\right| > \frac{RkC_2 + \sqrt{m}t}{m}\right] \leq 2\exp\left(-\frac{t^2}{2C_1 R^2}\right). \tag{2}$$

Assume without loss of generality that $I = \{1, \ldots, k\}$. Let $\boldsymbol{S}_I = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_k, \boldsymbol{y}_k)\}$. We look at $m\left|L_{\boldsymbol{S}}(\hat{h}_I) - L_D(\hat{h}_I)\right|$.

$$m\left|L_{\boldsymbol{S}}(\hat{h}_I) - L_D(\hat{h}_I)\right| \leq \left|\sum_{i=1}^{m}\left(\ell(\hat{h}_I(\boldsymbol{x}_i), \boldsymbol{y}_i) - \mathbb{E}[\ell(\hat{h}_I(\boldsymbol{x}_i), \boldsymbol{y}_i) \mid \boldsymbol{S}_I]\right)\right| \tag{3}$$

$$+ \left|\sum_{i=1}^{m}\left(\mathbb{E}[\ell(\hat{h}_I(\boldsymbol{x}_i), \boldsymbol{y}_i) \mid \boldsymbol{S}_I] - \mathbb{E}[\ell(\hat{h}_I(\boldsymbol{x}_i), \boldsymbol{y}_i)]\right)\right|. \tag{4}$$

From property 1 of $D^{(m)}$, we get that (4) $\leq C_2 Rk$. It is easy to see that $\mathbb{E}[(3)|\boldsymbol{S}_I] = 0$. From property 2 of $D^{(m)}$, we get that the tail of (3) is bounded as follows:

$$\Pr\left[(3) \geq t \;\middle|\; \boldsymbol{S}_I\right] \leq 2\exp\left(-\frac{t^2}{2C_1 R^2 m}\right).$$

Combining the two we get (2).

Given (2), we can union bound over all the possible choices of indices I our compression scheme could make. The number of such choices, we recall, is upper bounded by m^k. Let $\varepsilon = RkC_2/m + R\sqrt{2C_1\left(k\log m + \log(1/\delta)\right)/m}$. Then,

$$\Pr\left[\left|L_{\boldsymbol{S}}(\rho(\kappa(\boldsymbol{S}))) - L_D(\rho(\kappa(\boldsymbol{S})))\right| \geq \varepsilon\right] \tag{5}$$

$$\leq \Pr\left[\exists\, I \subset [m], |I| = k \colon \left|L_{\boldsymbol{S}}(\hat{h}_I) - L_D(\hat{h}_I)\right| \geq \varepsilon\right] \tag{6}$$

$$\leq \sum_{I \subset [m], |I| = k} \Pr\left[\left|L_{\boldsymbol{S}}(\hat{h}_I) - L_D(\hat{h}_I)\right| \geq \varepsilon\right] \leq m^k \frac{\delta}{m^k} \leq \delta. \tag{7}$$

Since $k = o(m)$ and $(1-\alpha)$ is a constant bounded away from 0, the dominant term in the value of ε is the second one as m grows. Hence we can re-write (7) as

$$\Pr\left[|L_{\boldsymbol{S}}(\rho(\kappa(\boldsymbol{S}))) - L_D(\rho(\kappa(\boldsymbol{S})))| \geq \frac{C(\alpha)R\sqrt{2m\left(k\log m + \log(1/\delta)\right)}}{m}\right] \leq \delta, \tag{8}$$

for a large enough constant C. □

Next, we show that if the data distribution is Dobrushin, the conditions of f Lemma 4.7 are satisfied. We will use two properties of Dobrushin distributions to show this. The first, stated as Lemma 4.8 states that all conditional distributions of a Dobrushin distribution also satisfy Dobrushin's condition with the same coefficient.

Lemma 4.8. *[Conditionining Preserves Low Influence Property] Consider a distribution π defined on Ω^n which satisfies Dobrushin's condition with coefficient α. Let $\boldsymbol{x} \sim \pi$ and let $0 < k \leq n$. Let $(a_1, a_2, \ldots, a_k) \in \Omega^k$ be such that $\Pr_\pi[(x_1, x_2, \ldots, x_k) = (a_1, a_2, \ldots, a_k)] > 0$. Then the conditional probability distribution $\pi_{\vec{a}} := \Pr_\pi[.|(x_1, x_2, \ldots, x_k) = (a_1, a_2, \ldots, a_k)]$ also satisfies Dobrushin's condition with coefficient α.*

We give the proof in Section B. The next property is more technically involved. It considers the Hamming distance between two samples drawn from a conditional distribution under two distinct conditionings. It shows that this quantity can be bounded to be of the order of the size of the set of conditioned variables.

Lemma 4.9 (Bounding Expected Hamming Distance Between two Conditional Measures)**.** *Let $a, a' \in (\mathcal{X} \times \mathcal{Y})^k$ and let $\boldsymbol{U} \sim D^{(m)}\left(.|(z_i)_{i=1}^k = (a_i)_{i=1}^k\right)$ and $\boldsymbol{V} \sim D^{(m)}\left(.|(z_i)_{i=1}^k = (a'_i)_{i=1}^k\right)$. Then, there exists a coupling $(\boldsymbol{Z_1}, \boldsymbol{Z_2})$ such that,*

$$\mathbb{E}[d_H(\boldsymbol{U}, \boldsymbol{V})] \leq \frac{k\alpha}{1-\alpha},$$

where $d_H(x, y) = \sum_{i=k+1}^m \mathbb{1}_{x_i \neq y_i}$ is the Hamming distance between x and y.

The full proof of Lemma 4.9 is given in Section B. We give an outline here. The key technical ingredent is Markov chain coupling. We start by observing that one way to generate a sample from the conditional distribution is to start at an arbitrary configuration and run a Gibbs sampler until mixing. Given the two conditionings, we consider two such Markov chains starting from the same configuration and hence the configurations have a Hamming distance of 0 in the beginning. The chains run in a coupled manner so that the Hamming distance between the configurations can be shown to remain small in expectation after each step of updates. Once we run until both chains have mixed the random configurations at that point correspond to samples from the conditional distributions but at every step along the way the Hamming distance between them has been bounded and hence it will remain bounded after the chains have mixed as well giving the result. To achieve this, it was necessary to find the appropriate coupling. We use what is known as the greedy coupling and is popularly used to show fast mixing of Gibbs sampling for Dobrushin distributions Levin et al. [2009]. This coupling tries to maximize the probability of agreement of the two chains updates' at each step.

With Lemmas 4.8 and 4.9, we are ready to show Lemma 4.10.

Lemma 4.10. *Let $D^{(m)}$ be a distribution over m variables which satisfies Dobrushin's condition with coefficient α such that the marginal of every variable is D. Let $\boldsymbol{S} \sim D^{(m)}$. Then we have*

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

Proof. Let $|I| = k$. Again, we assume without loss of generality that $I = \{1, \ldots, k\}$. The proof idea remains the same for any other set I. We have

$$m\left|\mathbb{E}\left[L_{\boldsymbol{S}}(h)\right] - \mathbb{E}\left[L_{\boldsymbol{S}}(h)|\boldsymbol{S}_I\right]\right| \leq \left|\sum_{i=1}^{k}\Big(\mathbb{E}[\ell(\hat{h}_I(\boldsymbol{x}_i), \boldsymbol{y}_i)] - \mathbb{E}[\ell(\hat{h}_I(\boldsymbol{x}_i), \boldsymbol{y}_i)|\boldsymbol{S}_I]\Big)\right| \tag{9}$$

$$+ \left|\sum_{i=k+1}^{m}\Big(\mathbb{E}[\ell(\hat{h}_I(\boldsymbol{x}_i), \boldsymbol{y}_i)] - \mathbb{E}[\ell(\hat{h}_I(\boldsymbol{x}_i), \boldsymbol{y}_i)|\boldsymbol{S}_I]\Big)\right|. \tag{10}$$

Since $0 \leq \ell(.) \leq R$, we get that (9) $\leq 2Rk$. (10) is bounded using Lemma 4.9 as follows. Let $\boldsymbol{S}_I$ and $\tilde{\boldsymbol{S}}_I$ be two realizations of the sample on the set of indices I. Then,

$$\left|\sum_{i=k+1}^{m}\Big(\mathbb{E}[\ell(\hat{h}_I(\boldsymbol{x}_i), \boldsymbol{y}_i) \mid \boldsymbol{S}_I] - \mathbb{E}[\ell(\hat{h}_I(\boldsymbol{x}_i), \boldsymbol{y}_i)]\Big)\right| \tag{11}$$

$$\leq \sup_{\tilde{\boldsymbol{S}}_I \neq \boldsymbol{S}_I}\sum_{i=k+1}^{m}\Big(\mathbb{E}[\ell(\hat{h}_I(\boldsymbol{x}_i), \boldsymbol{y}_i) \mid \boldsymbol{S}_I] - \mathbb{E}[\ell(\hat{h}_I(\boldsymbol{x}_i), \boldsymbol{y}_i) \mid \tilde{\boldsymbol{S}}_I]\Big) \tag{12}$$

$$\leq \sup_{\tilde{\boldsymbol{S}}_I \neq \boldsymbol{S}_I}\mathbb{E}\left[d_H(\boldsymbol{z}_{i=k+1}^m, \tilde{\boldsymbol{z}}_{i=k+1}^m)L \mid \boldsymbol{S}_I, \tilde{\boldsymbol{S}}_I\right] \leq \frac{Rk\alpha}{1-\alpha}, \tag{13}$$

where (12) utilized the fact that for two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ where $\boldsymbol{x} \geq 0$, $|\mathbb{E}[\boldsymbol{x}] - \mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y} = y_1]| \leq \sup_{y_2}|\mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y} = y_2] - \mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y} = y_1]|$ because $\mathbb{E}[\boldsymbol{x}]$ is a convex combination of values of the form $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y} = y]$. In (13), $\boldsymbol{z}_{i=k+1}^m$ is sampled from the distribution conditioned on $\boldsymbol{S}_I$ and $\tilde{\boldsymbol{z}}_{i=k+1}^m$ is sampled from the distribution conditioned on $\tilde{\boldsymbol{S}}_I$. The first inequality in (13) follows by coupling the two conditional probability spaces together with employing the fact that $|\ell| \leq R$, and the second is proved in Lemma 4.9. Putting the two bounds together, we get the first conclusion of the Lemma.

$$|\mathbb{E}\left[L_{\boldsymbol{S}}(h)\right] - \mathbb{E}\left[L_{\boldsymbol{S}}(h)|\boldsymbol{S}_I\right]| \leq \frac{Rk(2-\alpha)}{m(1-\alpha)}. \tag{14}$$

Next, we show the second conclusion of the Lemma which involves the distribution obtained by conditioning on $\boldsymbol{S}_I$. Recall that $I = \{1, 2, \ldots, k\}$. Firstly, we notice that due to the conditioning

$$m\left(L_{\boldsymbol{S}}(h) - \mathbb{E}[L_{\boldsymbol{S}}(h)|\boldsymbol{S}_I]\right) = \sum_{i=k+1}^{m}\Big(\ell(\hat{h}_I(\boldsymbol{x}_i, \boldsymbol{y}_i)) - \mathbb{E}[\ell(\hat{h}_I(\boldsymbol{x}_i), \boldsymbol{y}_i)|\boldsymbol{S}_I]\Big). \tag{15}$$

Denote $\mu(\boldsymbol{S}_I) = \sum_{i=k+1}^{m} \mathbb{E}[\ell(\hat{h}_I(\boldsymbol{x}_i), \boldsymbol{y}_i)|\boldsymbol{S}_I]$. Note that conditioned on $\boldsymbol{S}_I$, $\hat{h}_I$ is fixed. Next we invoke Lemma 4.8 to argue that since $((\boldsymbol{x}_i, \boldsymbol{y}_i))_{i=1}^m$ comes from a distribution satisfying Dobrushin's condition with coefficient α, the conditional distribution of $((\boldsymbol{x}_i, \boldsymbol{y}_i))_{i=k+1}^m|\boldsymbol{S}_I$ satisfies Dobrushin's condition with coefficient α as well. Hence we get the following concentration bound for (15) by employing Theorem 2.3 for Dobrushin distributions.

$$\Pr\left[\left|\sum_{i=k+1}^{m} \ell(\hat{h}_I(\boldsymbol{x}_i, \boldsymbol{y}_i)) - \mu(\boldsymbol{S}_I)\right| > \left(\sqrt{m-k}\right) t \;\middle|\; \boldsymbol{S}_I\right] \leq 2\exp\left(-\frac{t^2(1-\alpha)}{2R^2}\right).$$

Replacing $\sqrt{m-k}$ with $\sqrt{m}$, one obtains the second conclusion of the Lemma. □

Lemmas 4.7 and 4.10 imply Corollary 4.11.

Corollary 4.11 (Sample Compression Schemes Generalize under Dobrushin)**.** *Consider a sample $\boldsymbol{S} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_m, \boldsymbol{y}_m)\} \sim D^{(m)}$ which satisfies Dobrushin's condition with coefficient $\alpha\left(D^{(m)}\right)$. For any agnostic sample compression scheme (κ, ρ) of size k on a sample $\boldsymbol{S}$, and any loss function ℓ bounded by $R \geq 0$, we have that, for some constant $C(\alpha)$ depending on α,*

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

Lemmas 4.7 and 4.10 together imply Corollary 4.11. Corollary 4.11 together with the property of an agnostic sample compression scheme implies Theorem 4.5.

Proof. **of Theorem 4.5**. Let $\varepsilon = \varepsilon_1 + \varepsilon_2$ where

$$\varepsilon_1 = C(\alpha)R\sqrt{(k\log m + \log(2/\delta))/m},$$
$$\varepsilon_2 = 2R\sqrt{(1-\alpha)^{-1}\log(2/\delta)/m}.$$

Since (κ, ρ) is an agnostic sample compression scheme, we have that $L_{\boldsymbol{S}}(\rho(\kappa(\boldsymbol{S}))) \leq \inf_{h\in\mathcal{H}} L_{\boldsymbol{S}}(h)$. We will show that this implies that $L_{\boldsymbol{S}}(\rho(\kappa(\boldsymbol{S}))) - \inf_{h\in\mathcal{H}} L_D(h) \leq \varepsilon_2$ with probability $\geq 1-\delta/2$. Let $h^* = \arg\min_{h\in\mathcal{H}} L_D(h)$.

$$L_{\boldsymbol{S}}(\rho(\kappa(\boldsymbol{S}))) \leq \inf_{h\in\mathcal{H}} L_{\boldsymbol{S}}(h) \leq L_{\boldsymbol{S}}(h^*).$$

For any $h \in \mathcal{H}$, we have using Theorem 2.3 that

$$\Pr\left[|L_{\boldsymbol{S}}(h) - L_D(h)| \geq \varepsilon_2\right] \leq \delta/2. \tag{16}$$

Hence,

$$\begin{aligned}
&\Pr\left[L_D(\rho(\kappa(\boldsymbol{S}))) \geq \inf_{h\in\mathcal{H}} L_D(h) + \varepsilon\right] \\
=&\Pr\left[L_D(\rho(\kappa(\boldsymbol{S}))) \geq L_{\boldsymbol{S}}(\rho(\kappa(\boldsymbol{S}))) + \varepsilon + \left(\inf_{h\in\mathcal{H}} L_D(h) - L_{\boldsymbol{S}}(\rho(\kappa(\boldsymbol{S})))\right)\right] \\
\leq&\Pr\left[L_D(\rho(\kappa(\boldsymbol{S}))) \geq L_{\boldsymbol{S}}(\rho(\kappa(\boldsymbol{S}))) + \varepsilon_1\right] + \Pr\left[L_{\boldsymbol{S}}(\rho(\kappa(\boldsymbol{S})) \geq \inf_{h\in\mathcal{H}} L_D(h) + \varepsilon_2\right] \\
\leq&\delta/2 + \Pr\left[L_{\boldsymbol{S}}(h^*) \geq L_D(h^*) + \varepsilon_2\right] \leq \delta.
\end{aligned} \tag{17}$$

where the first inequality in (17) follows from Corollary 4.11 and the second inequality follows from (16).

□

4.1 Extending Learnability to Bounded Loss Functions

Learnability under Dobrushin's condition can also be extended for general loss functions which are bounded by some constant L. This can be shown using ε-approximate sample compression schemes defined in Section 4 of David et al. [2016]. Lemma 4.7 continues to hold for these sample compression schemes and the rest of the result follows suit.

5 Uniform Convergence for Weakly Dependent Data

In this section, we obtain uniform convergence bounds for the empirical loss on weakly dependent training data. 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. Hence, we present such bounds for a smaller family of distributions, which contain high temperature Markov Random Fields with pairwise potentials, however, allows 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 5.1 (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 5.2. *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 proof is simple and appears in Section 6.3. 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 5.3. *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 appears in Section 6.4 and is a direct corollary of Theorem 5.4 which is presented below. Note that Lemma 5.2 implies that Theorem 5.3 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 5.3 may hold as well in this regime. However, when $\beta(D^{(m)}) > 1$, Theorem 5.3 is not generally true, since concentration inequalities are not guaranteed to hold.

Applying Theorem 5.3 on any hypothesis class H with finite VC, one obtains the same sample complexity bounds of i.i.d data up to constant factors:

$$O\left(\frac{\mathrm{VC}(H) + \log(1/\delta)}{\varepsilon^2}\right). \tag{18}$$

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. Both notions were introduced to the learning community by Bartlett and Mendelson [2002], and Tomczak-Jaegermann [1989] proved the following:

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

for some universal constants $c, C > 0$. Bednorz and Latala [2014] resolved the long-standing Bernoulli conjecture by Talagrand and gave an exact characterization of the relation between these two notions. The standard techniques for bounding the Rademacher complexity, based on Chernoff-Hoeffding bounds, apply for bounding the Gaussian complexity as well, with the same constants (this includes chaining and covering numbers).

Theorem 5.3 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. Also, the m marginals of the weakly correlated distribution are not assumed to be identical. Hence, one can derive a variant of Theorem 5.3 with non-identical marginals.

Theorem 5.4. *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*

$$\mathop{\mathbb{E}}_{\boldsymbol{S}\sim D^{(m)}}\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{s_i}) - \mathop{\mathbb{E}}_{\boldsymbol{S}}\left[\frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{s_i})\right]\right) \leq \frac{C\mathfrak{G}_{D^{(m)}}(\mathcal{F})}{\sqrt{1 - 2\alpha_{\log}(D^{(m)})}}, \tag{19}$$

where $C > 0$ is a universal constant.

The proof appears in Section 6. Theorem 5.3 follows from Theorem 5.4 simply by applying a McDiarmind-like inequality on weakly correlated data satisfying Dobrushin's condition (Theorem 2.3).

6 Proofs for Section 5

In Section 6.1 we present some preliminaries for the proof. In Section 6.2 we prove the main result, Theorem 5.4. Then, in Section 6.3 we present the proof of Lemma 5.2 and in Section 6.4 the proof of Theorem 5.3.

6.1 Preliminaries: sub Gaussian distributions and stochastic processes

A joint distribution $P^{(m)}$ over $\mathbb{R}^m$ is a K^2-subGaussian if has subGaussian tails in any direction:

Definition 6.1. *A zero-mean distribution $P^{(m)}$ over $\mathbb{R}^m$ is a K^2-subGaussian if for any $\theta \in \mathbb{R}^n$ ($\theta \neq 0$) and any $t > 0$,*

$$\Pr_{\boldsymbol{w} \sim P^{(m)}} \left[\left| \sum_{i=1}^{m} \theta_i \boldsymbol{w}_i \right| > t \right] \leq 2 \exp\left(\frac{-t^2}{2K^2 \sum_{i=1}^{m} \theta_i^2} \right).$$

A *stochastic process* is a collection of joint random variables, $\{\boldsymbol{w}_i\}_{i \in I}$, taking values in $\mathbb{R}$, with some (possibly infinite) index-set I. A basic quantity of interest when talking about stochastic processes is the supremum, and in particular, the expected supremum, $\mathbb{E} \sup_{i \in I} \boldsymbol{w}_i$. We will be focusing on *zero-mean* processes, namely, those which satisfy $\mathbb{E}\, \boldsymbol{w}_i = 0$ for all $i \in I$. We present two important types of stochastic processes: Gaussian and subGaussian processes. A *Gaussian* process is a stochastic process where the variables are jointly Gaussian, namely, for any finite $U \subseteq I$, the collection $\{\boldsymbol{w}_i\}_{i \in U}$ is a multivariate Gaussian variable. A *subGaussian process* is a stochastic process for which $w_i - w_j$ is subGaussian for all $i, j \in I$. The following statement by Talagrand et al. [1996] upper bounds the expected maximum of a subGaussian process by that of a corresponding Gaussian process:

Theorem 6.2 (The majorizing measure theorem)**.** *Fix I to be some index set and let $\{\boldsymbol{w}_i\}_{i \in I}$ and $\{\boldsymbol{g}_i\}_{i \in I}$ be subGaussian and Gaussian zero-mean processes, respectively. For any $i, j \in I$, let σ_{ij}^2 denote the variance of $\boldsymbol{g}_i - \boldsymbol{g}_j$. Assume that for any $i, j \in I$, $\boldsymbol{w}_i - \boldsymbol{w}_j$ is a σ_{ij}^2-subGaussian random variable. Then,*

$$\mathbb{E}\left[\sup_{i \in I} \boldsymbol{w}_i \right] \leq C\, \mathbb{E}\left[\sup_{i \in I} \boldsymbol{g}_i \right],$$

for a universal constant $C > 0$.

6.2 Proof of Theorem 5.4

Here is the proof structure: first, we bound the left hand side of (19) by the $\boldsymbol{\sigma}$-complexity of $\mathcal{F}$ (Eq. (21)), where $\boldsymbol{\sigma}$ does not consist of i.i.d random signs, but rather it is a subGaussian distribution

with zero mean (Lemma 6.3 and the explanation afterwards). 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 6.4 and Lemma 6.5).

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:

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{S}} \sup_{f\in\mathcal{F}} \left(\frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{s}_i) - \mathop{\mathbb{E}}_{\boldsymbol{S}} \frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{s}_i) \right) &= \mathop{\mathbb{E}}_{\boldsymbol{S}} \sup_{f\in\mathcal{F}} \left(\frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{s}_i) - \mathop{\mathbb{E}}_{\boldsymbol{S'}} \frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{s}'_i) \right) \\ &\leq \mathop{\mathbb{E}}_{\boldsymbol{S},\boldsymbol{S'}} \sup_{f\in\mathcal{F}} \left(\frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{s}_i) - \frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{s}'_i) \right). \quad (20)\end{aligned}$$

Indeed, one can verify that the supremum of a sum is at most the sum of supremums. 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, $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_m) \in \{-1, 1\}^m$. Then, $\boldsymbol{T} = (\boldsymbol{t}_1, \ldots, \boldsymbol{t}_m)$ and $\boldsymbol{T}' = (\boldsymbol{t}'_1, \ldots, \boldsymbol{t}'_m)$ are defined as functions of $\boldsymbol{S}$, $\boldsymbol{S}'$ and $\boldsymbol{\sigma}$, as follows: for any $i \in \{1, \ldots, m\}$, 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 $P_{\boldsymbol{\sigma}|\boldsymbol{T}\boldsymbol{T}'}(\cdot \mid T, T')$, the conditional distribution of $\boldsymbol{\sigma}$, conditioned on $\boldsymbol{T} = T$ and $\boldsymbol{T}' = T'$. We bound the right hand side of (20), substituting $\boldsymbol{S}$ and $\boldsymbol{S}'$ with $\boldsymbol{T}$ and $\boldsymbol{T}'$, in a *change of measure* argument:

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{S},\boldsymbol{S'}} \left[\sup_{f\in\mathcal{F}} \left(\frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{s}_i) - \frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{s}'_i) \right) \right] &= \mathop{\mathbb{E}}_{\boldsymbol{T},\boldsymbol{T'},\boldsymbol{\sigma}} \left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\sigma}_i \left(f(\boldsymbol{t}_i) - f(\boldsymbol{t}'_i) \right) \right] \\ &\leq \mathop{\mathbb{E}}_{\boldsymbol{T},\boldsymbol{T'},\boldsymbol{\sigma}} \left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\sigma}_i f(\boldsymbol{t}_i) \right] + \mathop{\mathbb{E}}_{\boldsymbol{T},\boldsymbol{T'},\boldsymbol{\sigma}} \left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\sigma}_i (-f(\boldsymbol{t}'_i)) \right] \\ &=_{(*)} 2 \mathop{\mathbb{E}}_{\boldsymbol{T},\boldsymbol{T'},\boldsymbol{\sigma}} \left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\sigma}_i f(\boldsymbol{t}_i) \right] = 2 \mathop{\mathbb{E}}_{\boldsymbol{T},\boldsymbol{T'}} \widehat{\mathfrak{D}}_{\boldsymbol{T}}^{\boldsymbol{\sigma}_{\boldsymbol{T},\boldsymbol{T'}}}(\mathcal{F}), \quad (21)\end{aligned}$$

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.

Lemma 6.3. *For any T and T', $\boldsymbol{\sigma}_{T,T'}$ is zero-mean and satisfies $I^{\log}_{j,i}(\boldsymbol{\sigma}_{T,T'}) \leq 2I^{\log}_{j,i}(D^{(m)})$ for any $i \neq j$.*

Proof. Fix T and T'. By definition of $\boldsymbol{\sigma}$, for any $\sigma \in \{-1, 1\}^m$, $\Pr[\boldsymbol{T} = T, \boldsymbol{T}' = T', \boldsymbol{\sigma} = \sigma] = \Pr[\boldsymbol{T} = T, \boldsymbol{T}' = T', \boldsymbol{\sigma} = -\sigma]$. This implies that $\Pr[\boldsymbol{\sigma}_{T,T'} = \sigma] = \Pr[\boldsymbol{\sigma}_{T,T'} = -\sigma]$, which implies that $\boldsymbol{\sigma}_{T,T'}$ is zero-mean.

Next, we prove the inequality on the influence. For any $\sigma \in \{-1, 1\}^m$, define $S(\sigma, T, T')$ and $S'(\sigma, T, T')$ as the values that $\boldsymbol{S}$ and $\boldsymbol{S}'$ get when $\boldsymbol{T} = T$, $\boldsymbol{T}' = T'$ and $\boldsymbol{\sigma} = \sigma$.

Fix $i \neq j$ and fix $\sigma_i, \sigma_i', \sigma_j, \sigma_j' \in \{-1,1\}$ and $\sigma_{-i-j} \in \{-1,1\}^{m-2}$. Then,

$$\begin{aligned}
&\frac{1}{4}\log\left(\frac{P_{\boldsymbol{\sigma}_{T,T'}}(\sigma_{-i-j}\sigma_i\sigma_j)P_{\boldsymbol{\sigma}_{T,T'}}(\sigma_{-i-j}\sigma_i'\sigma_j')}{P_{\boldsymbol{\sigma}_{T,T'}}(\sigma_{-i-j}\sigma_i'\sigma_j)P_{\boldsymbol{\sigma}_{T,T'}}(\sigma_{-i-j}\sigma_i\sigma_j')}\right)\\
&= \frac{1}{4}\log\left(\frac{P_{\boldsymbol{\sigma},\boldsymbol{T},\boldsymbol{T'}}(\sigma_{-i-j}\sigma_i\sigma_j,T,T')P_{\boldsymbol{\sigma},\boldsymbol{T},\boldsymbol{T'}}(\sigma_{-i-j}\sigma_i'\sigma_j',T,T')}{P_{\boldsymbol{\sigma},\boldsymbol{T},\boldsymbol{T'}}(\sigma_{-i-j}\sigma_i'\sigma_j,T,T')P_{\boldsymbol{\sigma},\boldsymbol{T},\boldsymbol{T'}}(\sigma_{-i-j}\sigma_i\sigma_j',T,T')}\right)\\
&= \frac{1}{4}\log\left(\frac{P_{\boldsymbol{S}}(S(\sigma_{-i-j}\sigma_i\sigma_j,T,T'))P_{\boldsymbol{S}}(S(\sigma_{-i-j}\sigma_i'\sigma_j',T,T'))}{P_{\boldsymbol{S}}(S(\sigma_{-i-j}\sigma_i'\sigma_j,T,T'))P_{\boldsymbol{S}}(S(\sigma_{-i-j}\sigma_i\sigma_j',T,T'))}\right)\\
&+ \frac{1}{4}\log\left(\frac{P_{\boldsymbol{S'}}(S'(\sigma_{-i-j}\sigma_i\sigma_j,T,T'))P_{\boldsymbol{S'}}(S'(\sigma_{-i-j}\sigma_i'\sigma_j',T,T'))}{P_{\boldsymbol{S'}}(S'(\sigma_{-i-j}\sigma_i'\sigma_j,T,T'))P_{\boldsymbol{S'}}(S'(\sigma_{-i-j}\sigma_i\sigma_j',T,T'))}\right)\\
&\leq 2I_{j,i}^{\log}(D^{(m)}),
\end{aligned}$$

where the last step follows from the definition of the log-influences. The proof follows. □

If follows from Lemma5.2 and Lemma 6.3 that $\alpha(\boldsymbol{\sigma}_{T,T'}) \leq \alpha^{\log}(\boldsymbol{\sigma}_{T,T'}) \leq 2\alpha^{\log}(D^{(m)})$. From Theorem 2.3 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 will 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 (21). The proof follows from the Fernique-Talagrand Majorizing measure theory.

Lemma 6.4. *Fix $z \in Z^m$. If $\boldsymbol{\tau}$ is a K^2-subgaussian, then,*

$$\widehat{\mathfrak{O}}_z^{\boldsymbol{\tau}}(\mathcal{F}) \leq CK\widehat{\mathfrak{G}}_z(\mathcal{F}),$$

(for some universal constant $C > 0$). In particular, $\widehat{\mathfrak{O}}_z^{\boldsymbol{\sigma}_{T,T'}}(\mathcal{F}) \leq C\widehat{\mathfrak{G}}_z(\mathcal{F})/\sqrt{1-2\alpha^{\log}(D^{(m)})}$.

Proof. Define the random process, $\{\boldsymbol{w}_f\}_{f\in\mathcal{F}}$, where each $\boldsymbol{w}_f$ equals $\frac{1}{m}\sum_{i=1}^m f(z_i)\boldsymbol{\tau}_i$. Let $\boldsymbol{g} \sim N(0, \mathrm{I}_m)$, and define the Gaussian process $\{\boldsymbol{g}_f\}_{f\in\mathcal{F}}$ by $\boldsymbol{g}_f = \frac{1}{m}\sum_{i=1}^m \boldsymbol{g}_i f(z_i)$. Note that by the definition of a subGaussian random variable, for any $f, f' \in \mathcal{F}$ and any $t > 0$,

$$\Pr\left[|\boldsymbol{w}_f - \boldsymbol{w}_{f'}| > t\right] = \Pr\left[\left|\frac{1}{m}\sum_{i=1}^m \boldsymbol{\tau}_i\left(f(z_i) - f'(z_i)\right)\right| > t\right] \leq 2\exp\left(\frac{m^2t^2}{2K^2\sum_{i=1}^m\left(f(z_i)-f'(z_i)\right)^2}\right),$$

which implies that $\boldsymbol{w}_f - \boldsymbol{w}_{f'}$ is a $K^2\sum_{i=1}^m\left(f(z_i)-f'(z_i)\right)^2/m^2$-subGaussian. Additionally,

$$\mathrm{Var}(\boldsymbol{g}_f - \boldsymbol{g}_{f'}) = \mathrm{Var}\left(\frac{1}{m}\sum_{i=1}^m \boldsymbol{g}_i\left(f(z_i)-f'(z_i)\right)\right) = \frac{1}{m^2}\sum_{i=1}^m\left(f(z_i)-f'(z_i)\right)^2.$$

Theorem 6.2 implies that

$$\widehat{\mathfrak{O}}_z^{\boldsymbol{\tau}}(\mathcal{F}) = \mathbb{E}\sup_{f\in\mathcal{F}}\boldsymbol{w}_f \leq C\,\mathbb{E}\sup_{f\in\mathcal{F}}K\boldsymbol{g}_f = CK\widehat{\mathfrak{G}}_z(\mathcal{F}).$$

□

Lemma 6.4 implies that the right hand side of (21) is bounded as follows:

$$\mathop{\mathbb{E}}_{\boldsymbol{T},\boldsymbol{T}'} \widehat{\mathfrak{D}}_{\boldsymbol{T}}^{\boldsymbol{\sigma}_{\boldsymbol{T},\boldsymbol{T}'}}(\mathcal{F}) \le \frac{C}{\sqrt{1-2\beta(D^{(m)})}} \mathop{\mathbb{E}}_{\boldsymbol{T},\boldsymbol{T}'} \widehat{\mathfrak{G}}_{\boldsymbol{T}}(\mathcal{F}) = \frac{C}{\sqrt{1-2\beta(D^{(m)})}} \mathfrak{G}_{\boldsymbol{T}}(\mathcal{F}). \tag{22}$$

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

Lemma 6.5. *The following holds:*

$$\mathfrak{G}_{\boldsymbol{T}}(\mathcal{F}) \le 2\mathfrak{G}_{D^{(m)}}(\mathcal{F}).$$

Proof. Recall that $\boldsymbol{T} = T(\boldsymbol{S}, \boldsymbol{S}', \boldsymbol{\sigma})$ is a mixture of two samples $\boldsymbol{S}$ and $\boldsymbol{S}'$ drawn from $D^{(m)}$: $\boldsymbol{t}_i = \boldsymbol{s}_i$ if $\boldsymbol{\sigma}_i = 1$ and otherwise $\boldsymbol{t}_i = \boldsymbol{s}'_i$. Fix $S, S' \in Z^m$ and $\sigma \in \{-1,1\}^m$, and let $T = T(S, S', \sigma)$. Taking expectation over $\boldsymbol{g} \sim \boldsymbol{\mathcal{N}(0, \mathrm{I}_m)}$, one obtains:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{g}\sim\mathcal{N}(0,\mathrm{I}_m)} \left[\sup_{f\in\mathcal{F}} \sum_{i=1}^m \boldsymbol{g}_i f(t_i)\right] = \mathop{\mathbb{E}}_{\boldsymbol{g}\sim\mathcal{N}(0,\mathrm{I}_m)} \left[\sup_{f\in\mathcal{F}} \left(\sum_{i:\,\boldsymbol{\sigma}_i=1} \boldsymbol{g}_i f(s_i) + \sum_{i:\,\boldsymbol{\sigma}_i=-1} \boldsymbol{g}_i f(s'_i)\right)\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{g}\sim\mathcal{N}(0,\mathrm{I}_m)} \left[\sup_{f\in\mathcal{F}} \left(\sum_{i:\,\boldsymbol{\sigma}_i=1} \boldsymbol{g}_i f(s_i) + \sum_{i:\,\boldsymbol{\sigma}_i=-1} \boldsymbol{g}_i f(s'_i) + \mathop{\mathbb{E}}_{\boldsymbol{g}'\sim\mathcal{N}(0,\mathrm{I}_m)} \left[\sum_{i:\,\boldsymbol{\sigma}_i=-1} \boldsymbol{g}'_i f(s_i) + \sum_{i:\,\boldsymbol{\sigma}_i=1} \boldsymbol{g}'_i f(s'_i)\right]\right)\right] \\
&\le \mathop{\mathbb{E}}_{\boldsymbol{g},\boldsymbol{g}'\in\mathcal{N}(0,\mathrm{I}_m)} \left[\sup_{f\in\mathcal{F}} \left(\sum_{i:\,\boldsymbol{\sigma}_i=1} \boldsymbol{g}_i f(s_i) + \sum_{i:\,\boldsymbol{\sigma}_i=-1} \boldsymbol{g}_i f(s'_i) + \sum_{i:\,\boldsymbol{\sigma}_i=-1} \boldsymbol{g}'_i f(s_i) + \sum_{i:\,\boldsymbol{\sigma}_i=1} \boldsymbol{g}'_i f(s'_i)\right)\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{g},\boldsymbol{g}'\sim\mathcal{N}(0,\mathrm{I}_m)} \left[\sup_{f\in\mathcal{F}} \left(\sum_{i=1}^m \boldsymbol{g}_i f(s_i) + \sum_{i=1}^m \boldsymbol{g}'_i f(s'_i)\right)\right] \le 2 \mathop{\mathbb{E}}_{\boldsymbol{g}\sim\mathcal{N}(0,\mathrm{I}_m)} \left[\sup_{f\in\mathcal{F}} \left(\sum_{i=1}^m \boldsymbol{g}_i f(s_i)\right)\right].
\end{aligned}$$

Taking expectation in both sides over $\boldsymbol{T} = T$ and $\boldsymbol{S} = S$, the result follows. □

The proof concludes by equations (20), (21), (22) and Lemma 6.5.

6.3 Proof of Lemma 5.2

First, we bound the log-influences $I^{\log}_{j,i}(\boldsymbol{z})$ by $\beta_{j,i}(\boldsymbol{z})$ for an MRF with pairwise potentials. Assume $\boldsymbol{z}$ has pairwise potentials $\psi_{ij}(z_i, z_j)$ and node-wise potentials $\varphi_i(z_i)$. Fix $i \ne j$, z_i, z'_i, z_j, z'_j and z_{-i-j}, and note that

$$\frac{1}{4}\log\frac{P_{\boldsymbol{z}}[z_i z_j z_{-i-j}] P_{\boldsymbol{z}}[z'_i z'_j z_{-i-j}]}{P_{\boldsymbol{z}}[z'_i z_j z_{-i-j}] P_{\boldsymbol{z}}[z_i z'_j z_{-i-j}]} = \frac{1}{4}\left(\psi_{ij}(z_i z_j) + \psi_{ij}(z'_i z'_j) - \psi_{ij}(z'_i z_j) - \psi_{ij}(z_i z'_j)\right) \le \beta_{ij}(\boldsymbol{z}).$$

This concludes that $I^{\log}_{j,i}(\boldsymbol{z}) \le \beta_{ij}(\boldsymbol{z})$. Next, we bound the Dobrushin influences with respect to the log-influences. We begin with the following auxiliary lemma.

Lemma 6.6. *Let $\{M_{a,b}\}_{a,b\in\{-1,1\}}$ be positive numbers. Then,*

$$\left|\frac{M_{1,1}}{M_{1,1}+M_{-1,1}} - \frac{M_{1,-1}}{M_{1,-1}+M_{-1,-1}}\right| \le \frac{1}{4}\max\left\{\log\frac{M_{1,1}M_{-1,-1}}{M_{1,-1}M_{-1,1}}, \log\frac{M_{1,-1}M_{-1,1}}{M_{1,1}M_{-1,-1}}\right\}. \tag{23}$$

Proof. We can assume that $\sum_{a,b} M_{a,b} = 1$, by scaling. Define a random variable $\boldsymbol{w}$ over $\{-1,1\}^2$ with $\Pr[\boldsymbol{w} = (a,b)] = M_{a,b}$. Any random variable with two binary coordinates can be written as an Ising model. In particular, there exists $\theta \in \mathbb{R}$, and $\varphi_1, \varphi_2 \colon \{-1,1\} \to \mathbb{R}_+$ such that $\Pr[\boldsymbol{w} = (a,b)] = e^{\varphi_1(a)+\varphi_2(b)+ab\theta}$. Note that the left hand side of (23) equals $I_{2\to 1}(\boldsymbol{w})$, and from Lemma 2.6, it is bounded by $|\theta|$. On the other hand, the right hand size of (23) equals $|\theta|$, and the proof follows. □

Fix $z_{-i-j} \in \Omega^{m-2}$ and $z_j, z'_j \in \Omega$ and let F be the event such that

$$d_{TV}(P_{\boldsymbol{z}_i|\boldsymbol{z}_{-i}}(\cdot \mid z_{-i-j}z_j), P_{\boldsymbol{z}_i|\boldsymbol{z}_{-i}}(\cdot \mid z_{-i-j}z'_j)) = P_{\boldsymbol{z}_i|\boldsymbol{z}_{-i}}(F \mid z_{-i-j}z_j) - P_{\boldsymbol{z}_i|\boldsymbol{z}_{-i}}(F \mid z_{-i-j}z'_j).$$

The proof follows by applying Lemma 6.6 with

$$M_{1,1} = \Pr[\boldsymbol{z}_i \in F, \boldsymbol{z}_j = z_j, \boldsymbol{z}_{-i-j} = z_{-i-j}]; \quad M_{1,-1} = \Pr[\boldsymbol{z}_i \in F, \boldsymbol{z}_j = z'_j, \boldsymbol{z}_{-i-j} = z_{-i-j}]$$
$$M_{-1,1} = \Pr[\boldsymbol{z}_i \in F^{\mathsf{c}}, \boldsymbol{z}_j = z_j, \boldsymbol{z}_{-i-j} = z_{-i-j}]; \quad M_{-1,-1} = \Pr[\boldsymbol{z}_i \in F^{\mathsf{c}}, \boldsymbol{z}_j = z'_j, \boldsymbol{z}_{-i-j} = z_{-i-j}].$$

6.4 Proof of Theorem 5.3

We prove a slightly more general result.

Theorem 6.7. *Assume the same setting as in Theorem 5.4 and additionally, that there exists $L > 0$ such that for every $f \in \mathcal{F}$ and $z \in Z$, $|f(z)| \le L$. Then, for any $t > 0$,*

$$\Pr_{\boldsymbol{S}\sim D^{(m)}}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^m f(\boldsymbol{s_i}) - \mathop{\mathbb{E}}_{\boldsymbol{S}}\left[\sum_{i=1}^m f(\boldsymbol{s_i})\right]\right| > \frac{C\mathfrak{G}_{D^{(m)}}(\mathcal{F})}{\sqrt{1-2\alpha_{\log}(D^{(m)})}} + CL\sqrt{m}t\right] \le e^{-t^2/2},$$

for some universal constant $C > 0$.

Proof. We start by bounding the probability that $\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^m f(\boldsymbol{s_i}) - \mathbb{E}_{\boldsymbol{S}}\left[\sum_{i=1}^m f(\boldsymbol{s_i})\right]\right)$ is larger than the corresponding value, removing the absolute value (later, we will argue for the opposite inequality). The proof follows from a McDiarmid-like inequality for dependent distributions. Define the function $M \colon Z^m \to \mathbb{R}$ by

$$M(S) = \sup_{f\in\mathcal{F}}\left(\sum_{i=1}^m f(s_i) - \mathop{\mathbb{E}}_{\boldsymbol{S}}\sum_{i=1}^m f(s_i)\right).$$

For any $S = (s_1,\ldots,s_m)$ and $S' = (s'_1,\ldots,s'_m) \in Z^m$, it holds that $|M(S) - M(S')| \le \sum_{i=1}^m 2L\mathbb{1}_{s_i\neq s'_i}$. Lemma 5.2 and Theorem 2.3 imply that for any $t' > 0$,

$$\Pr_{\boldsymbol{S}\sim D^{(m)}}\left[M(\boldsymbol{S}) - \mathbb{E}\,M(\boldsymbol{S}) > t'\right] \le \exp\left(\frac{-t'^2\left(1-\alpha_{\log}\left(D^{(m)}\right)\right)}{C'L^2m}\right) \le \exp\left(\frac{-t'^2}{2C'L^2m}\right).$$

Substituting $\mathbb{E}\,M(S)$ using Theorem 6.7 and setting $t := t'\sqrt{1/(C'L^2m)}$, the bound concludes.

To bound the opposite inequality, namely, the probability that $\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{\boldsymbol{S}}\left[\sum_{i=1}^m f(\boldsymbol{s_i})\right] - \sum_{i=1}^m f(\boldsymbol{s_i})\right)$ is large, one can apply the same arguments on $-\mathcal{F} := \{-f \colon f \in \mathcal{F}\}$, and note that the Gaussian complexity of $\mathcal{F}$ equals that of $-\mathcal{F}$. □

7 Comparison to Related Work on Time Series

Much of the work on non-iid Rademacher complexity has focused on time series. A *time series* is a distribution $D^{(\infty)}$ on random variables that are indexed by integer times. In our setting, we let the random sequence be $\boldsymbol{z} = \ldots, \boldsymbol{z_{-2}}, \boldsymbol{z_{-1}}, \boldsymbol{z_0}, \boldsymbol{z_1}, \boldsymbol{z_2}, \ldots$, where each $\boldsymbol{z_i} = (\boldsymbol{x_i}, \boldsymbol{y_i})$. Such a series is said to be *stationary* if for any times t and t' and any positive integer k, the joint distribution of $(\boldsymbol{z_t}, \ldots, \boldsymbol{z_{t+k}})$ and $(\boldsymbol{z'_t}, \ldots, \boldsymbol{z_{t'+k}})$ are the same. A time series is said to be *mixing* if events that are farther spread out in time are closer to being independent from one another. More formally, for any $t, t' \in \mathbb{Z} \cup \{-\infty, +\infty\}$, let $\sigma_t^{t'}$ be the σ-algebra generated by $\{\boldsymbol{z_i} | t \leq i \leq t'\}$. The α-function is defined as $\alpha(k) = \sup_t\{|\Pr(A, B) - \Pr(A)\Pr(B)| \mid B \in \sigma_{-\infty}^t, A \in \sigma_{t+k}^{\infty}\}$, and the β-function is defined as $\beta(k) = \sup_t \mathbb{E}_{B \in \sigma_{-\infty}^t} \sup_{A \in \sigma_{t+k}^{\infty}} |\Pr(A|B) - \Pr(A)|$. A time series $D^{(\infty)}$ is said to be α-mixing if $\lim_{k\to\infty} \alpha(k) = 0$ and β-mixing if $\lim_{k\to\infty} \beta(k) = 0$. It is a well known that for all k $\beta(k) > \alpha(k)$, and thus β-mixing implies α-mixing.

Since our work focuses on distributions with identical marginals, we compare it to previous work on stationary time series. Of these works, Mohri and Rostamizadeh [2009] is most relevant to ours, since McDonald et. al.'s work on non-mixing time series solves the forecasting problem, predicting $\boldsymbol{z}_{m+1}$ given previous data, rather than on predicting $\boldsymbol{y}$ given $\boldsymbol{x}$ as in our setting. Mohri and Rostamizadeh [2009] derive the following uniform convergence Rademacher complexity bound for stationary β-mixing time series.

Theorem 7.1 (Theorem 1 in Mohri and Rostamizadeh [2009])**.** *Let H be a hypothesis class, ℓ a loss function bounded by $L \geq 0$, and $\mathcal{F} = \{\ell \circ h \mid h \in H\}$. Then, for any size m sample S from a stationary β-mixing time series $D^{(\infty)}$ with marginal distribution D, and any $\mu, a > 0$ with $2\mu a = m$ and $\delta > 2(\mu - 1)\beta(a)$, then with probability at least $1 - \delta$, the following inequality holds for all $h \in H$:*

$$|L_D(h) - L_S(h)| \leq \mathfrak{R}_{D^\mu}(\mathcal{F}) + L\sqrt{\frac{\log \frac{2}{\delta - 2(\mu-1)\beta(a)}}{2\mu}}$$

Mohri *et. al.* derive their result by exploiting the fact that samples that are sufficiently far apart in the time series are close to independent. They split the sample $(\boldsymbol{z_1}, \ldots, \boldsymbol{z_m})$ into large blocks, and argue that if one data point is taken from each block, then the resultant subsample is close to being iid. Once they "thin" the original sample in this way, they finish the argument by appealing to methods for proving Rademacher bounds for iid distributions.

While Mohri et. al.'s thinning argument makes analysis simple and yields Rademacher bounds, our work significanlty improves the sample complexity bounds if the time series is also an MRF with pairwise potentials, or, if it is Dobrushin - at least for learnability. We demonstrate a super-quadratic improvement in the following example.

7.1 Example: Uniform Convergence Sample Complexity Comparison

Given a symmetric real matrix $\Theta \in \mathbb{R}^{2n+1 \times 2n+1}$, we define the distribution $P_{\Theta,n}$ on the $2n+1$ variables $((\boldsymbol{x_{-n}}, \boldsymbol{y_{-n}}), \ldots, (\boldsymbol{x_n}, \boldsymbol{y_n}))$ as follows. Each $\boldsymbol{x_i}$ takes a value in the set $[-1, 1]$ with the

probability density function $p(x_{-n}, \ldots, x_n) \propto \prod_{i \neq j} e^{\theta_{i,j} x_i x_j}$. Then, we let $\boldsymbol{y_i} = f(\boldsymbol{x_i})$, where $f : [-1, 1] \to \{-1, +1\}$ is any function.

In order for Theorem 5.3 to apply to $P_{\Theta,n}$, we must ensure that the sums of the pairwise potentials including any one node i are less than (and bounded away from) 0.5. So, we let $\theta_{i,j} = \frac{c}{|i-j| \log^2(|i-j|+1)}$ for each $i \neq j$, where c is some constant that ensures the convergent series $\sum_{k=1}^{\infty} \frac{c}{k \log^2(k+1)}$ is bounded by a sufficiently small constant.

While the $\beta(k)$ coefficients are bounded by $o(1)$ for $P_{\Theta,n}$, the distribution is not techincally β-mixing, since a β-mixing distribution must technically be a distribution on a countable collection of variables (not just on $2n+1$ variables). This technicality could be resolved by taking the limiting distribution that arises when n tends to infinity. The resultant distribution would also be stationary by the symmetry of the θs. In order to avoid technicalities of probability theory in favor of more clearly illustrating the sample complexity differences, we just consider a huge (but finite) value of n, and call that resulting distribution $D^{(n)}$.

Let $H \subseteq \{h : [-1, 1] \to \{-1, +1\}\}$ be any hypothesis class of finite VC-dimension d, and consider the 0-1 loss. A sample complexity bound $m(\varepsilon, \delta)$ is a bound on the number of samples needed to ensure that the generalization gap is less than ε with probability at least $1 - \delta$. We compare our sample complexity bound obtained from Equation (18), with Mohri et. al.'s bound, derived from Theorem 7.1. A key to the comparison is the following fact.

Claim 7.2. *For the distribution D, $\beta(k) \geq \alpha(k) = \Omega(\theta_{0,k}) = \Omega\left(\frac{c}{k \log^2(k+1)}\right)$.*

Proof. **Sketch** It is always true that the β mixing coefficient is larger than the α mixing coefficient. In order to show that the α-mixing coefficient is not too small, we study the events $E_i \equiv (\boldsymbol{x_i} > 0)$. In particular, we note that $E_0 \in \sigma_{-\infty}^0$ and $E_k \in \sigma_k^\infty$ and we show that $\Pr(E_0, E_k) - \Pr(E_0)\Pr(E_k) = \Omega(\theta_{0,k})$. The density function is symmetric, i.e. $p((\boldsymbol{x_i})_{i=-n}^n = (a_i)_{i=-n}^n) = p((\boldsymbol{x_i})_{i=-n}^n = (-a_i)_{i=-n}^n)$, so $\Pr(E_0)\Pr(E_k) = 1/4$.

It now suffices to show that $\Pr(E_0, E_k) = 1/4 + \Omega(\theta_{0,k})$. We first observe that since all the $\theta_{i,j}$ coefficients are non-negative in the distribution $P_{\Theta,n}$, $\Pr(E_0, E_k)$ can only be made smaller if all but the coefficient $\theta_{0,k}$ are made zero. Let this new distribution be called P. Under P, the probability

$$\Pr(E_0, E_k) = \frac{\int_0^1 \int_0^1 e^{\theta_{0,k} xy} dx dy}{\int_{-1}^1 \int_{-1}^1 e^{\theta_{0,k} xy} dx dy} \tag{24}$$

$$= \frac{1 + \theta_{0,k}/4 + O(\theta_{0,k}^2)}{4 + O(\theta_{0,k}^2)} \text{ (Taylor expansion)} \tag{25}$$

$$= 1/4 + \theta_{0,k}/16 + O(\theta_{0,k}^2) \tag{26}$$

$$= 1/4 + \Omega(\theta_{0,k}) \tag{27}$$

□

We ignore polylogarithmic-factors for clarity in the following calculations. Since, $\mathcal{H}$ has VC-dimension d, Mohri et. al.'s Theorem 7.1 implies that the generalization gap satisfies $\varepsilon \leq \sqrt{\frac{d}{\mu}} + \sqrt{\frac{1}{\mu}}$. By the conditions of the theorem, the block size a and the number of blocks μ must satisfy $\mu a = m$ and $\mu\beta(a) \leq \delta$. Adding in the constraint $\beta(a) = \tilde{\Omega}(1/a)$ of Lemma 7.2, yields their tightest sample complexity bound

$$m_{\text{prior-work}}(\varepsilon, \delta) = \tilde{\Theta}\left(\frac{d^2}{\delta\varepsilon^4}\right).$$

Alternately, our sample complexity bound from Equation (18) is

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

The set of stationary β-mixing time series studied by Mohri et. al. is not a subset of the Dobrushin and pairwise-potential-MRF distributions we study in this paper. However, in this example, where the time series is also a pairwise-potential-MRF our analysis improves the sample requirement quadratically in ε and d, and *exponentially* in $1/\delta$. The quadratic improvement quantifies the ineffciency of the thinning method in this context, while the exponential improvement in the dependence on $1/\delta$ results from our use of powerful measure concentration inequalities in our analysis.

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.

David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2009.

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.

A The Gibbs Sampling Algorithm

In Section 4 we use the Gibbs sampling algorithm associated with Dobrushin distributions. For completeness we present the algorithm below.

Input: Set of variables V, Configuration $x_0 \in S^{|V|}$, Distribution π
initialization;
for $t = 1$ **to** T **do**
  Sample i uniformly from $\{1, 2, \ldots, n\}$;
  Sample $X_i \sim \Pr_\pi [.|X_{-i} = x_{-i}]$ and set $x_{i,t} = X_i$;
  For all $j \neq i$, set $x_{j,t} = x_{j,t-1}$;
end

Algorithm 1: Gibbs Sampling

B Omitted Details from Section 4

B.1 Lemmas Required for the Proof of Theorem 4.5

Lemma 4.8. *[Conditionining Preserves Low Influence Property] Consider a distribution π defined on Ω^n which satisfies Dobrushin's condition with coefficient α. Let $\boldsymbol{x} \sim \pi$ and let $0 < k \leq n$. Let $(a_1, a_2, \ldots, a_k) \in \Omega^k$ be such that $\Pr_\pi[(x_1, x_2, \ldots, x_k) = (a_1, a_2, \ldots, a_k)] > 0$. Then the conditional probability distribution $\pi_{\vec{a}} := \Pr_\pi[.|(x_1, x_2, \ldots, x_k) = (a_1, a_2, \ldots, a_k)]$ also satisfies Dobrushin's condition with coefficient α.*

Proof. . For any $k+1 \leq i, j \leq n$, such that $i \neq j$, the influence of i on j under the conditional measure is

$$I_{\pi_{\vec{a}}}(i \to j) = \sup_{\substack{(a_{k+1},\ldots,a_n)\setminus(a_i,a_j)\in\Omega^{n-k-2} \\ a_i, a_i' \in \Omega}} \left\|\pi(.|\boldsymbol{x}_{-i,-j} = a_{-i,-j}, x_i = a_i) - \pi(.|\boldsymbol{x}_{-i,-j} = a_{-i,-j}, x_i = a_i')\right\|_1 \tag{28}$$

$$\leq I_\pi(i \to j), \tag{29}$$

where (29) holds because the influence of variable i on j is the supremum total variation over conditionings which are Hamming distance one apart and hence upper bounds the partial conditioning we have in (28) Since π satisfies the property that for each i, $\sum_j I_\pi(i \to j) \leq \alpha$ the above implies that $\sum_j I_{\pi_{\vec{a}}}(i \to j) \leq \alpha$. Hence the Lemma follows. □

Proof. **of Lemma 4.9.** We will use the Gibbs sampling algorithm (1) for our proof. Since $D^{(m)}\left(.|(z_i)_{i=1}^k = (a_i)_{i=1}^k\right)$ and $D^{(m)}\left(.|(z_i)_{i=1}^k = (a_i')_{i=1}^k\right)$ satisfy Dobrushin's condition, they have associated Gibbs sampling Markov chains over the state space $(\mathcal{X} \times \mathcal{Y})^{m-k}$ which are ergodic. Denote the chains by M_U and M_V respectively. Let $(\boldsymbol{U}_t)_{t\geq 0}$ and $(\boldsymbol{V}_t)_{t\geq 0}$ be executions of M_U and M_V respectively such that $\boldsymbol{U}_0 = \boldsymbol{V}_0$. We couple these two executions in the following way. At each time step t, we choose an index $i \in I_k = \{k+1, \ldots, m\}$ uniformly at random and resample $\boldsymbol{U}_{i,t}$ and $\boldsymbol{V}_{i,t}$ according to their conditional distributions. And we update $\boldsymbol{U}_{i,t}$ and $\boldsymbol{V}_{i,t}$ so as to minimize $\Pr[\boldsymbol{U}_{i,t} \neq \boldsymbol{V}_{i,t}]$. Such a coupling is known as the greedy coupling. We now argue via induction that for all t, $\mathbb{E}\left[d_H(\boldsymbol{U}_t, \boldsymbol{V}_t)|d_H(\boldsymbol{U}_0, \boldsymbol{V}_0) = 0\right] \leq \frac{k\alpha}{1-\alpha}$. We have $d_H(\boldsymbol{U}_0, \boldsymbol{V}_0) = 0$. Assume that for some $t > 0$, $\mathbb{E}\left[d_H(\boldsymbol{U}_t, \boldsymbol{V}_t)|d_H(\boldsymbol{U}_0, \boldsymbol{V}_0) = 0\right] \leq \frac{k\alpha}{1-\alpha}$. We will compute a bound on $\mathbb{E}\left[d_H(\boldsymbol{U}_{t+1}, \boldsymbol{V}_{t+1})|(\boldsymbol{U}_t, \boldsymbol{V}_t)\right]$ first. We partition the indices $i \in I_k$ into two sets $I_k^=$ and $I_k^\neq$ as follows.

$$I_k^= = \{i \in I_k \mid \boldsymbol{U}_{i,t} = \boldsymbol{V}_{i,t}\} \tag{30}$$

$$I_k^\neq = \{i \in I_k \mid \boldsymbol{U}_{i,t} \neq \boldsymbol{V}_{i,t}\} \tag{31}$$

To understand whether the Hamming distance goes up or down at time $t+1$, we perform a case analysis. Suppose index i was chosen in time step $t+1$ from the set $I_k^=$. Then $d_H(\boldsymbol{U}_{t+1}.\boldsymbol{V}_{t+1}) -$

$d_H(\boldsymbol{U}_t, \boldsymbol{V}_t) = 1$ or 0.

$$\begin{aligned}
&\Pr\left[d_H(\boldsymbol{U}_{t+1}, \boldsymbol{V}_{t+1}) - d_H(\boldsymbol{U}_t, \boldsymbol{V}_t) = 1 \mid \boldsymbol{U}_t, \boldsymbol{V}_t, i \text{ was chosen for step } t+1 \text{ from } I_k^{=}\right] \\
&= \Pr\left[\boldsymbol{U}_{i,t+1} \neq \boldsymbol{V}_{i,t+1} \mid \boldsymbol{U}_t, \boldsymbol{V}_t, i \text{ was chosen for step } t+1 \text{ from } I_k^{=}\right] \\
&\leq \left\| D^{(m)}(.|(z_i)_{i=1}^k = (a_i)_{i=1}^k, (z_i)_{i=k+1}^m = \boldsymbol{U}_t) - D^{(m)}(.|(z_i)_{i=1}^k = (a_i')_{i=1}^k, (z_i)_{i=k+1}^m = \boldsymbol{V}_t) \right\|_{TV} \qquad (32) \\
&\leq \sum_{j \in [k] \cup I_k^{\neq}} I(j \to i). \qquad (33)
\end{aligned}$$

(32) follows from the definition of Gibbs sampling update probability and the property of total variation distance that it is equal to the worst-case probability of disagreement of a draw from the two distributions over all valid couplings of the two distributions. Hence the greedy coupling should satisfy inequality (32). To get (33), we use the triangle inequality of total variation distance and bound the total variation in the expression with a sum of total variations between where each term is TV between conditional distributions whose conditioned states have Hamming distance ≤ 1. Each of these total variations is then bounded by their corresponding influence terms (since the influence is defined as a supremum over such conditionings).

Now suppose i was chosen from the set $I_k^{\neq}$ instead. Then $d_H(\boldsymbol{U}_{t+1}.\boldsymbol{V}_{t+1}) - d_H(\boldsymbol{U}_t, \boldsymbol{V}_t) = -1$ or 0.

$$\begin{aligned}
&\Pr\left[d_H(\boldsymbol{U}_{t+1}, \boldsymbol{V}_{t+1}) - d_H(\boldsymbol{U}_t, \boldsymbol{V}_t) = -1 \mid \boldsymbol{U}_t, \boldsymbol{V}_t, i \text{ was chosen for step } t+1 \text{ from } I_k^{\neq}\right] \qquad (34) \\
&= \Pr\left[\boldsymbol{U}_{i,t+1} = \boldsymbol{V}_{i,t+1} \mid \boldsymbol{U}_t, \boldsymbol{V}_t, i \text{ was chosen for step } t+1 \text{ from } I_k^{\neq}\right] \qquad (35) \\
&= 1 - \Pr\left[\boldsymbol{U}_{i,t+1} \neq \boldsymbol{V}_{i,t+1} \mid \boldsymbol{U}_t, \boldsymbol{V}_t, i \text{ was chosen for step } t+1 \text{ from } I_k^{\neq}\right] \qquad (36) \\
&\geq 1 - \left\| D^{(m)}(.|(z_i)_{i=1}^k = (a_i)_{i=1}^k, (z_i)_{i=k+1}^m = \boldsymbol{U}_t) - D^{(m)}(.|(z_i)_{i=1}^k = (a_i')_{i=1}^k, (z_i)_{i=k+1}^m = \boldsymbol{V}_t) \right\|_{TV} \qquad (37) \\
&\geq 1 - \sum_{j \in [k] \cup I_k^{\neq}} I(j \to i), \qquad (38)
\end{aligned}$$

using a similar reasoning as above. The expected change in the Hamming distance is

$$\begin{aligned}
&\mathbb{E}\left[d_H(\boldsymbol{U}_{t+1}, \boldsymbol{V}_{t+1}) - d_H(\boldsymbol{U}_t, \boldsymbol{V}_t) \mid \boldsymbol{U}_t, \boldsymbol{V}_t\right] \qquad (39) \\
&\leq \frac{1}{m-k} \mathbb{E}\left[\sum_{i \in I_k^{=}} \sum_{j \in [k] \cup I_k^{\neq}} I(j \to i) - \sum_{i \in I_k^{\neq}} \left(1 - \sum_{j \in [k] \cup I_k^{\neq}} I(j \to i)\right) \mid \boldsymbol{U}_t, \boldsymbol{V}_t\right] \qquad (40) \\
&\leq \frac{1}{m-k} \sum_{j \in [k] \cup I_k^{\neq}} \sum_{i \in [m] \setminus [k]} I(j \to i) - \frac{1}{m-k} d_H(\boldsymbol{U}_t, \boldsymbol{V}_t) \qquad (41) \\
&\leq \frac{(k + d_H(\boldsymbol{U}_t, \boldsymbol{V}_t))\alpha - d_H(\boldsymbol{U}_t, \boldsymbol{V}_t)}{m-k}, \qquad (42)
\end{aligned}$$

where (40) follows from (33) and (38). Now,

$$\mathbb{E}\left[d_H(\boldsymbol{U}_{t+1}, \boldsymbol{V}_{t+1}) | d_H(\boldsymbol{U}_0, \boldsymbol{V}_0) = 0\right] = \mathbb{E}\left[\mathbb{E}\left[d_H(\boldsymbol{U}_{t+1}, \boldsymbol{V}_{t+1}) \mid \boldsymbol{U}_t, \boldsymbol{V}_t\right] \mid d_H(\boldsymbol{U}_0, \boldsymbol{V}_0) = 0\right] \tag{43}$$

$$\leq \mathbb{E}\left[d_H(\boldsymbol{U}_t, \boldsymbol{V}_t) + \frac{(k + d_H(\boldsymbol{U}_t, \boldsymbol{V}_t))\alpha - d_H(\boldsymbol{U}_t, \boldsymbol{V}_t)}{m-k} \mid d_H(\boldsymbol{U}_0, \boldsymbol{V}_0) = 0\right] \tag{44}$$

$$\leq \frac{k\alpha}{1-\alpha}\left(1 - \frac{1-\alpha}{m-k}\right) + \frac{k\alpha}{m-k} = \frac{k\alpha}{1-\alpha}. \tag{45}$$

Hence we have shown that $\mathbb{E}\left[d_H(\boldsymbol{U}_{t+1}, \boldsymbol{V}_{t+1}) | d_H(\boldsymbol{U}_0, \boldsymbol{V}_0) = 0\right] \leq \frac{k\alpha}{1-\alpha}$. Now, we have that

$$\mathbb{E}[d_H(\boldsymbol{U}, \boldsymbol{V})] = \lim_{t\to\infty} \mathbb{E}[d_H(\boldsymbol{U}_t, \boldsymbol{V}_t) \mid \boldsymbol{U}_0, \boldsymbol{V}_0] \tag{46}$$

$$\leq \frac{k\alpha}{1-\alpha}, \tag{47}$$

where (46) follows because the Gibbs sampler we consider is ergodic and (47) follows from (45). □