Journal of Machine Learning Research 16 (2016) pp Submitted 02/15; Revised 09/15; Published MM/16

Stability and Generalization in Structured Prediction

Ben London BLONDON@CS.UMD.EDU
University of Maryland
Bert Huang BHUANG@VT.EDU
Virginia Tech
Lise Getoor GETOOR@SOE.UCSC.EDU
University of California, Santa Cruz

Editor: X

Abstract

Structured prediction models have been found to learn effectively from a few large examples—sometimes even just one. Despite empirical evidence, canonical learning theory cannot guarantee generalization in this setting because the error bounds decrease as a function of the number of examples. We therefore propose new PAC-Bayesian generalization bounds for structured prediction that decrease as a function of both the number of examples and the size of each example. Our analysis hinges on the stability of joint inference and the smoothness of the data distribution. We apply our bounds to several common learning scenarios, including max-margin and soft-max training of Markov random fields. Under certain conditions, the resulting error bounds can be far more optimistic than previous results and can even guarantee generalization from a single large example.

Keywords: structured prediction, learning theory, PAC-Bayes, generalization bounds

1. Introduction

Many important applications of machine learning require making multiple interdependent predictions whose dependence relationships form a graph. In some cases, the number of inputs and outputs can be enormous. For instance, in natural language processing, a document may contain thousands of words to be assigned a part-of-speech tag; in computer vision, a digital image may contain millions of pixels to be segmented; and in social network analysis, a relational graph may contain millions of users to be categorized. Obtaining fully annotated examples can be time-consuming and expensive, due to the number of variables. It is therefore common to train a structured predictor on far fewer examples than are used in the unstructured setting. In the extreme (yet not atypical) case, the training set consists of a single example, with large internal structure. A central question in statistical learning theory is *generalization*; that is, whether the expected error at test time will be reasonably close to the empirical error measured during training. Canonical learning-theoretic results for structured prediction (e.g., Taskar et al., 2004; Bartlett et al., 2005; McAllester, 2007) only guarantee generalization when the number of training examples is high. Yet, this pessimism contradicts a wealth of experimental results (e.g., Taskar et al., 2002; Tsochantaridis et al., 2005), which indicate that training on a few large examples is sufficient. In this work, we address the question of when generalization is possible in this setting. We derive new

generalization bounds for structured prediction that are far more optimistic than previous results. When sufficient conditions hold, our bounds guarantee generalization from a few large examples—even just one.

The intuition behind our analysis is motivated by a common practice known alternatively as *templating* or *parameter-tying*. At a high level, templating shares parameters across substructures (e.g., nodes, edges, etc.) with identical local structure. (Templating is explained in detail in Section 2.2.2.) Originally proposed for relational learning as a way of dealing with non-uniformly-structured examples, templating has an additional benefit in that it effectively limits the complexity of the hypothesis class by reducing the number of parameters to be learned. Each instance of a substructure within an example acts as a kind of "micro example" of a template. Since each example may contain many micro examples, it is plausible that generalization could occur from even a single example.

Part of the difficulty when formalizing this intuition is that the micro examples are interdependent. Like all statistical arguments, generalization bounds must show that the empirical error concentrates around the expected error, and analyzing the concentration of functions of dependent random variables is nontrivial. Moreover, inference in a structured predictor is typically formulated as a global optimization over all outputs simultaneously. Due to model-induced dependencies, changes to one input may affect many of the outputs, which affects the loss differently than in binary or multiclass prediction. Thus, this problem cannot be viewed as simply learning from interdependent data, which has been studied extensively (e.g., Usunier et al., 2006; Mohri and Rostamizadeh, 2010; Ralaivola et al., 2010).

We therefore have two obstacles: the dependence in the data distribution and the dependence induced by the the predictor. We characterize the former dependence using concepts from measure concentration theory (Kontorovich and Ramanan, 2008), and we view the latter dependence through the lens of algorithmic *stability*. Unlike previous literature (e.g., Bousquet and Elisseeff, 2002), we are not interested in the stability of the learning algorithm; rather, we examine the stability of inference (more specifically, a functional of the predictions) with respect to perturbations of the input. In prior work (London et al., 2013, 2014), we used the term *collective stability* to describe the stability of the predictions, guaranteeing collective stability for predictors whose inference objectives are strongly convex. In this work, we propose a form of stability that generalizes collective stability by analyzing the loss function directly. Our new definition accommodates a broader range of loss functions and predictors, and eliminates our previous reliance on strong convexity. Moreover, we support functions that are *locally* stable over some subset of their domain, and random functions that are stable with high probability.

This probabilistic notion of stability lends itself nicely to the *PAC-Bayes* framework, in which prediction proceeds by drawing a random hypothesis from a distribution on the hypothesis class. For this and other technical reasons, we use PAC-Bayesian analysis to derive our generalization bounds. When certain conditions are met by the distributions on the data and hypothesis class, our bounds can be as tight as $\tilde{\mathrm{O}}\left(1/\sqrt{mn}\right)$, where m is the number of examples, and n is the size of each example. Note that this expression decreases as either m or n increase. This rate is much tighter than previous results, which only guarantee $\tilde{\mathrm{O}}\left(1/\sqrt{m}\right)$. From our bounds, we conclude that it is indeed possible to generalize from a few large examples—potentially even just one.

1.1 Related Work

One of the earliest explorations of generalization in structured prediction is by Collins (2001), who developed risk bounds for language parsers using various classical tools, such as the Vapnik-Chervonenkis dimension and margin theory. In Taskar et al.'s (2004) landmark paper on max-margin Markov networks, the authors use covering numbers to derive risk bounds for their proposed class of models. Bartlett et al. (2005) improved this result using PAC-Bayesian analysis.[1] McAllester (2007; 2011) provided a comprehensive PAC-Bayesian study of various structured losses and learning algorithms. Recently, Hazan et al. (2013) proposed a PAC-Bayes bound with a form often attributed to Catoni (2007), which can be minimized directly using gradient descent. Giguère et al. (2013) used PAC-Bayesian analysis to derive risk bounds for the kernel regression approach to structured prediction. In a similar vein as the above literature, yet taking a significantly different approach, Bradley and Guestrin (2012) derived finite sample complexity bounds for learning conditional random fields using the composite likelihood estimator.

All of the above works have approached the problem from the traditional viewpoint, that the generalization error should decrease proportionally to the number of examples. In a previous publication (London et al., 2013), we proposed the first bounds that decrease with both the number of examples and the size of each example (given suitably weak dependence within each example). We later refined these results using PAC-Bayesian analysis (London et al., 2014). Our current work builds upon this foundation to derive similarly optimistic generalization bounds, while accommodating a broader range of loss functions and hypothesis classes.

From a certain perspective, our work fits into a large body of literature on learning from various types of interdependent data. Most of this is devoted to "unstructured" prediction. Usunier et al. (2006) and Ralaivola et al. (2010) used concepts from graph coloring to analyze generalization in learning problems that induce a dependency graph, such as bipartite ranking. In this case, the training data contains dependencies, but prediction is localized to each input-output pair. Similarly, Mohri and Rostamizadeh (2009, 2010) derived risk bounds for ϕ-mixing and β-mixing temporal data, using an "independent blocking" technique due to Yu (1994). The hypotheses they consider predict each time step independently, which makes independent blocking possible. Since we are interested in hypotheses (and loss functions) that perform joint inference, which may not decompose over the outputs, we cannot employ techniques such as graph coloring and independent blocking.

A related area of research is learning to forecast time series data. In this setting, the goal is to predict the next (or, some future) value in the series, given (a moving window of) previous observations. The generalization error of time series forecasting has been studied extensively by McDonald et al. (e.g., 2012) in the β-mixing regime. Similarly, Alquier and Wintenburger (2012) derived oracle inequalities for ϕ-mixing conditions.

The idea of learning from one example is related to the "one-network" learning paradigm, in which data is generated by a (possibly infinite) random field, with certain labels observed for training. The underlying model is estimated from the partially observed network, and

1. PAC-Bayesian analysis is often accredited to McAllester (1998, 1999), and has been refined by a number of authors (e.g., Herbrich and Graepel, 2001; Langford and Shawe-Taylor, 2002; Seeger, 2002; Ambroladze et al., 2006; Catoni, 2007; Germain et al., 2009; Lever et al., 2010; Seldin et al., 2012).

the learned model is used to predict the missing labels, typically with some form of joint inference. Xiang and Neville (2011) examined maximum likelihood and pseudo-likelihood estimation in this setting, proving that are asymptotically consistent. Note that this is a *transductive* setting, in that the network data is fixed (i.e., realized), so the learned hypothesis is not expected to generalize to other network data. In contrast, we analyze *inductive* learning, wherein the model is applied to future draws from a distribution over network data.

Connections between stability and generalization have been explored in various forms. Bousquet and Elisseeff (2002) proposed the stability of a learning algorithm as a tool for analyzing generalization error. Wainwright (2006) analyzed the stability of marginal probabilities in variational inference, identifying the relationship between stability and strong convexity (similar to our work in London et al., 2013, 2014). He used this result to show that an *inconsistent* estimator, which uses approximate inference during training, can asymptotically yield lower regret (relative to the optimal Bayes least squares estimator) than using the *true* model with approximate inference. Honorio (2011) showed that the Bayes error rate of various graphical models is related to the stability of their log-likelihood functions with respect to changes in the model parameters.

1.2 Our Contributions

Our primary contribution is a new PAC-Bayesian analysis of structured prediction, producing generalization bounds that decrease when either the number of examples, m, or the size of each example, n, increase. Under suitable conditions, our bounds can be as tight as $\tilde{\mathrm{O}}\left(1/\sqrt{mn}\right)$. Our results apply to any composition of loss function and hypothesis class that satisfies our local stability conditions, which includes a broad range of modeling regimes used in practice. We also propose a novel view of PAC-Bayesian "derandomization," based on the principle of stability, which provides a general proof technique for converting a generalization bound for a randomized structured predictor into a bound for a deterministic structured predictor.

As part of our analysis, we derive a new bound on the moment-generating function of a locally stable functional. The tightness of this bound (hence, our generalization bounds) hinges on a measure of the aggregate dependence between the random variables within each example. Our bounds are meaningful when the dependence is sub-logarithmic in the number of variables. We provide two examples of stochastic processes for which this condition holds. These results, and their implications for measure concentration, are of independent interest.

We apply our PAC-Bayes bounds to several common learning scenarios, including *max-margin* and *soft-max* training of (conditional) Markov random fields. To demonstrate the benefit of local stability analysis, we also consider a specific generative process that induces unbounded stability in certain predictors, given certain inputs. These examples suggest several factors to be considered when modeling structured data, in order to obtain the fast generalization rate: (1) templating is crucial; (2) the norm of the parameters contributes to the stability of inference, and should be controlled via regularization; and (3) limiting local interactions in the model can improve stability, hence, generalization. All of these considerations can be summarized by the classic tension between representational power and overfitting, applied to the structured setting. Most importantly, these examples confirm

that generalization from limited training examples is indeed possible for many structured prediction techniques used in practice.

1.3 Organization

The remainder of this paper is organized as follows. Section 2 introduces the notation used throughout the paper and reviews some background in structured prediction, templated Markov random fields, generalization error and PAC-Bayesian analysis. In Section 3, we propose general properties that characterize the local stability of a generic functional (e.g., the composition of a loss function and hypothesis). In Section 4, we introduce the statistical quantities and inequalities used in our analysis, as well as some examples of "nice" dependence conditions. Section 5 presents our main results: new PAC-Bayes bounds for structured prediction. We also propose a general proof technique for derandomizing the bounds using stability. In Section 6, we apply our bounds to a number of common learning scenarios. Specifically, we examine learning templated Markov random fields in the max-margin and soft-max frameworks, under various assumptions about the data distribution. Section 7 concludes our study with a discussion of the results and their implications for practitioners of structured prediction.

2. Preliminaries

This section introduces the notation and background used in this paper. We begin with notational conventions. We then formally define structured prediction and review some background on templated Markov random fields, a general class of probabilistic graphical models commonly used in structured prediction. Finally, we review the concept of generalization and discuss the PAC-Bayes framework, which we use to state our main results.

2.1 Notational Conventions

Let $\mathcal{X} \subseteq \mathbb{R}^k$ denote a domain of observations, and let $\mathcal{Y}$ denote a finite set of discrete labels. Let $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$ denote the cross product of the two, representing input-output pairs.

Let $\mathbf{Z} \triangleq (Z_i)_{i=1}^n$ denote a set of n random variables, with joint distribution $\mathbb{D}$ on a sample space $\mathcal{Z}^n$. We denote *realizations* of $\mathbf{Z}$ by $\mathbf{z} \in \mathcal{Z}^n$. We use $\Pr_{\mathbf{Z}\sim\mathbb{D}}\{\,\cdot\,\}$ to denote the probability of an event over realizations of $\mathbf{Z}$, distributed according to $\mathbb{D}$. Similarly, we use $\mathbb{E}_{\mathbf{Z}\sim\mathbb{D}}[\,\cdot\,]$ to specify an expectation over $\mathbf{Z}$. When it is clear from context which variable(s) and distribution the probability (or expectation) is taken over, we may omit the subscript notation. We will occasionally employ the shorthand $\mathbb{D}(\mathcal{S})$ to denote the measure of a subset $\mathcal{S} \subseteq \mathcal{Z}^n$ under $\mathbb{D}$; i.e., $\mathbb{D}(\mathcal{S}) = \Pr_{\mathbf{Z}\sim\mathbb{D}}\{\mathbf{Z} \in \mathcal{S}\}$. With a slight abuse of notation, which should be clear from context, we also use $\mathbb{D}(\mathbf{Z}_{i:j} \,|\, E)$ to denote the distribution of some subset of the variables, $(Z_i, \ldots, Z_j)$, conditioned on an event, E.

For a graph $G \triangleq (\mathcal{V}, \mathcal{E})$, with nodes $\mathcal{V}$ and edges $\mathcal{E}$, we use $|G| \triangleq |\mathcal{V}| + |\mathcal{E}|$ to denote the total number of nodes and edges in G.

2.2 Structured Prediction

At its core, *structured prediction* (sometimes referred to as *structured output prediction* or *structured learning*) is about learning concepts that have a natural internal structure. In

the framework we consider, each example of a concept contains n interdependent random variables, $\mathbf{Z} \triangleq (Z_i)_{i=1}^n$, with joint distribution $\mathbb{D}$. Each $Z_i \triangleq (X_i, Y_i)$ is an input-output pair, taking values in $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.[2] Each example is associated with an implicit dependency graph, $G \triangleq (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} \triangleq \{1, \ldots, n\}$ indexes $\mathbf{Z}$, and $\mathcal{E}$ captures the dependencies in $\mathbf{Z}$. Unless otherwise stated, assume that the edge structure is given *a priori*. The edge structure may be obvious from context, or may be inferred beforehand. To simplify our analysis, we assume that each example uses the same structure.

The prediction task is to infer $\mathbf{Y} \triangleq (Y_i)_{i=1}^n$, conditioned on $\mathbf{X} \triangleq (X_i)_{i=1}^n$. A *hypothesis*, h, maps $\mathcal{X}^n$ to $\mathcal{Y}^n$, using some internal parametric representation that incorporates the structure of the problem (an example of which is given in Section 2.2.1). We are interested in hypotheses that perform joint reasoning over all variables simultaneously. We therefore assume that computing $h(\mathbf{X})$ implicitly involves a global optimization that does not decompose over the outputs, due to dependencies.

2.2.1 Markov Random Fields

To better understand structured prediction, it will help to consider a specific hypothesis class. One popular class is that of *Markov random fields* (MRFs), a broad family of undirected graphical models that generalizes many models used in practice, such as relational Markov networks (Taskar et al., 2002), conditional random fields (Lafferty et al., 2001), and Markov logic networks (Richardson and Domingos, 2006). In this section, we review some background on MRFs.

Recall that each example is associated with a dependency graph, $G \triangleq (\mathcal{V}, \mathcal{E})$. We assume that the edge set is undirected. This does not limit the applicability of our analysis, since there exists a straightforward conversion from directed models (Koller and Friedman, 2009). The parameters of an MRF are organized according to the *cliques* (i.e., complete subgraphs), $\mathcal{C}$, contained in G. For each clique, $c \in \mathcal{C}$, we associate a real-valued *potential* function, $\theta_c(\mathbf{y} \,|\, \mathbf{x}; \mathbf{w})$, parameterized by a vector of weights, $\mathbf{w} \in \mathbb{R}^d$, for some $d \geq 1$. This function indicates the score for $\mathbf{Y}_c$ being in state $\mathbf{y}_c$, conditioned on the observation $\mathbf{X} = \mathbf{x}$. The potentials define a log-linear conditional probability distribution,

$$p\left(\mathbf{Y} = \mathbf{y} \,|\, \mathbf{X} = \mathbf{x}; \mathbf{w}\right) \triangleq \exp\left(\sum_{c \in \mathcal{C}} \theta_c(\mathbf{y} \,|\, \mathbf{x}; \mathbf{w}) - \Phi(\mathbf{x}; \mathbf{w})\right),$$

where

$$\Phi(\mathbf{x}; \mathbf{w}) \triangleq \ln \sum_{\mathbf{y}' \in \mathcal{Y}^n} \exp\left(\sum_{c \in \mathcal{C}} \theta_c(\mathbf{y}' \,|\, \mathbf{x}; \mathbf{w})\right)$$

is a normalizing function known as the *log-partition*.

For convenience, we represent the label space, $\mathcal{Y}$, by the set of $|\mathcal{Y}|$ standard basis (i.e., "one-hot") vectors, $\mathbf{e}_1, \ldots, \mathbf{e}_{|\mathcal{Y}|}$. Thus, the joint state of a clique, c, is represented by a vector, $\mathbf{y}_c = \bigotimes_{i \in c} y_i$, of length $|\mathcal{Y}_c| \triangleq |\mathcal{Y}|^{|c|}$. With a slight abuse of notation, we overload the potential functions so that $\boldsymbol{\theta}_c(\mathbf{x}; \mathbf{w}) \in \mathbb{R}^{|\mathcal{Y}_c|}$ denotes a vector of potentials, and

$$\theta_c(\mathbf{y} \,|\, \mathbf{x}; \mathbf{w}) = \boldsymbol{\theta}_c(\mathbf{x}; \mathbf{w}) \cdot \mathbf{y}_c.$$

2. To minimize bookkeeping, we have assumed a one-to-one correspondence between input and output variables, and that the Z_i variables have identical domains, but these assumptions can be relaxed.

Thus, with

$$\boldsymbol{\theta}(\mathbf{x}; \mathbf{w}) \triangleq (\boldsymbol{\theta}_c(\mathbf{x}; \mathbf{w}))_{c \in \mathcal{C}} \quad \text{and} \quad \hat{\mathbf{y}} \triangleq (\mathbf{y}_c)_{c \in \mathcal{C}}\,,$$

we have that

$$\sum_{c \in \mathcal{C}} \theta_c(\mathbf{y} \,|\, \mathbf{x}; \mathbf{w}) = \boldsymbol{\theta}(\mathbf{x}; \mathbf{w}) \cdot \hat{\mathbf{y}}.$$

We refer to $\hat{\mathbf{y}}$ as the *full* representation of $\mathbf{y}$.

The canonical inference problems for MRFs are *maximum a posteriori* (MAP) inference, which computes the mode of the distribution,

$$\arg\max_{\mathbf{y} \in \mathcal{Y}^n} p\left(\mathbf{Y} = \mathbf{y} \,|\, \mathbf{X} = \mathbf{x}; \mathbf{w}\right),$$

and *marginal* inference, which computes the marginal distribution of a subset of the variables. In general, both tasks are intractable—MAP inference is NP-hard and marginal inference is #P-hard (Roth, 1996)—though there are some useful special cases for which inference is tractable, and many approximation algorithms for the general case. In this work, we assume that an efficient (approximate) inference algorithm is given.

2.2.2 Templating

An important property of the above construction is that the same vector of weights, $\mathbf{w}$, is used to parameterize all of the potential functions. One could imagine that $\mathbf{w}$ contains a unique subvector, $\mathbf{w}_c$, for every clique. However, one could also bin the cliques by a set of *templates*—such as singletons (nodes), pairs (edges) or triangles (hyperedges)—then use the same weights for each template. This technique is alternatively referred to as *templating* or *parameter-tying*.

With templating, one can define general inductive rules to reason about datasets of arbitrary size and structure. Because of this flexibility, templating is used in many *relational* models, such as relational Markov networks (Taskar et al., 2002), relational dependency networks (Neville and Jensen, 2004), and Markov logic networks (Richardson and Domingos, 2006).

A templated model implicitly assumes that all *groundings* (i.e., instances) of a template should be modeled identically, meaning location within the graph is irrelevant. A non-templated model is location-aware and therefore has higher representational power. However, without templating, the dimensionality of $\mathbf{w}$ scales with the number of cliques; whereas, with templating, the dimensionality of $\mathbf{w}$ is constant. Thus, we find the classic tension between representational power and overfitting. To mitigate overfitting, one must restrict model complexity. Yet, too little expressivity will hamper predictive performance. This consideration is critical to the application of our generalization bounds.

In practice, templated models typically consist of unary and pairwise templates. We refer to these as pairwise models. Higher-order templates (i.e., cliques of three or more) can capture certain inductive rules that pairwise models cannot. For example, for a binary relation r, the transitive closure $r(A, B) \wedge r(B, C) \implies r(A, C)$ requires triadic templates. Rules like this are sometimes used for link prediction and entity resolution. Of course, this additional expressivity comes at a cost, as will become apparent later.

2.2.3 Defining the Potential Functions

In many applications of MRFs, the potentials are defined as multilinear functions of $(\mathbf{w}, \mathbf{x}, \mathbf{y})$. For example, assuming each node i has local observations $x_i \in \mathcal{X}$ and label $y_i \in \mathcal{Y}$, we can define a vector of local *features*,

$$f_i(\mathbf{x}, \mathbf{y}) \triangleq x_i \otimes y_i,$$

using the Kronecker product (since y_i is a standard basis vector). Similarly, for each edge $\{i, j\} \in \mathcal{E}$, let

$$f_{ij}(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{2} \begin{bmatrix} x_i \\ x_j \end{bmatrix} \otimes (y_i \otimes y_j).$$

Here, we have defined the edge features using a concatenation of the local observations, though this need not be the case. In general, the edge features can be arbitrary functions of the observations, such as kernels or similarity functions. Or, we could eschew the observations altogether and just use $y_i \otimes y_j$, which is typical in practice.

The potential functions are then defined as weighted feature functions. For the following, we will assume that the weights are templated, as described in Section 2.2.2. For each node, we associate a set of singleton weights, $\mathbf{w}_s \in \mathbb{R}^{d_s}$, and for each edge, a set of pairwise weights, $\mathbf{w}_p \in \mathbb{R}^{d_p}$, where d_s and d_p denote the respective lengths of the node and edge features. Then,

$$\theta_i(\mathbf{y} \,|\, \mathbf{x}; \mathbf{w}) \triangleq \mathbf{w}_s \cdot f_i(\mathbf{x}, \mathbf{y}) \quad \text{and} \quad \theta_{ij}(\mathbf{y} \,|\, \mathbf{x}; \mathbf{w}) \triangleq \mathbf{w}_p \cdot f_{ij}(\mathbf{x}, \mathbf{y});$$

and, with

$$\mathbf{w} \triangleq \begin{bmatrix} \mathbf{w}_s \\ \mathbf{w}_p \end{bmatrix} \quad \text{and} \quad \mathbf{f}(\mathbf{x}, \mathbf{y}) \triangleq \begin{bmatrix} \sum_{i \in \mathcal{V}} f_i(\mathbf{x}, \mathbf{y}) \\ \sum_{\{i,j\} \in \mathcal{E}} f_{ij}(\mathbf{x}, \mathbf{y}) \end{bmatrix},$$

we have that

$$\boldsymbol{\theta}(\mathbf{x}; \mathbf{w}) \cdot \hat{\mathbf{y}} = \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, \mathbf{y}).$$

In Section 6, we apply our generalization bounds to the above construction of a templated MRF, consisting of singleton and pairwise linear potentials (with or without edge features).

2.3 Learning and Generalization

Given a set of m training examples, $\hat{\mathbf{Z}} \triangleq (\mathbf{Z}^{(l)})_{l=1}^{m}$, drawn independently and identically from $\mathbb{D}$, the goal of learning is to produce a hypothesis from a specified class, denoted $\mathcal{H} \subseteq \{h : \mathcal{X}^n \to \mathcal{Y}^n\}$. We do not assume that the data is generated according to some target concept in $\mathcal{H}$, so $\mathcal{H}$ may be misspecified.

Hypotheses are evaluated using a *loss function* of the form $L : \mathcal{H} \times \mathcal{Z}^n \to \mathbb{R}_+$, which may have access to the internal representation of the hypothesis. For a given loss function, L, let $\overline{L}(h) \triangleq \mathbb{E}_{\mathbf{Z} \sim \mathbb{D}}[L(h, \mathbf{Z})]$ denote the expected loss over realizations of an example. This quantity, known as the *risk*, corresponds to the error h will incur on future predictions. Let

$$\hat{L}(h, \hat{\mathbf{Z}}) \triangleq \frac{1}{m} \sum_{l=1}^{m} L(h, \mathbf{Z}^{(l)})$$

denote the average loss on the training set, $\hat{\mathbf{Z}}$. Most learning algorithms minimize (an upper bound on) $\hat{L}(h, \hat{\mathbf{Z}})$, since it is an *empirical* estimate of the risk.

The goal of our analysis is to upper-bound the difference of the expected and empirical risks, $\overline{L}(h) - \hat{L}(h, \hat{\mathbf{Z}})$—which we refer to as the *generalization error*[3]—thereby yielding an upper bound on the risk. As is typically done in generalization analysis, we show that, with high probability over draws of a training set, the generalization error is upper-bounded by a function of certain properties of the domain, hypothesis class and learning algorithm, which decreases as the (effective) size of the training set increases. Note that small generalization error does not necessarily imply small risk, since the empirical risk may be large. Nonetheless, small generalization error implies that the empirical risk will be a good estimate of the risk, thus motivating empirical risk minimization.

2.3.1 PAC-Bayes

PAC-Bayes is a framework for analyzing the risk of a randomized predictor. One begins by fixing a *prior* distribution, $\mathbb{P}$, on the hypothesis space, $\mathcal{H}$. Then, given some training data, one constructs a *posterior* distribution, $\mathbb{Q}$, the parameters of which are typically learned from the training data. For example, when $\mathcal{H}$ is a subset of Euclidean space, a common PAC-Bayesian construction is a standard multivariate Gaussian prior with an isotropic Gaussian posterior, centered at the learned hypothesis. To make a prediction on an input, $\mathbf{x}$, one draws a hypothesis, $h \in \mathcal{H}$, according to $\mathbb{Q}$, then computes $h(\mathbf{x})$.

Since prediction is randomized, the risk quantities are defined over draws of h, which we denote by

$$\hat{L}(\mathbb{Q}, \hat{\mathbf{Z}}) \triangleq \mathop{\mathbb{E}}_{h \sim \mathbb{Q}} \left[\hat{L}(h, \hat{\mathbf{Z}}) \right] \quad \text{and} \quad \overline{L}(\mathbb{Q}) \triangleq \mathop{\mathbb{E}}_{h \sim \mathbb{Q}} \left[\overline{L}(h) \right].$$

The goal of PAC-Bayesian analysis is to upper-bound some measure of discrepancy between these quantities. The discrepancy is sometimes defined as the KL divergence between error rates, or the squared difference. In this work, we upper-bound the difference, $\overline{L}(\mathbb{Q}) - \hat{L}(\mathbb{Q}, \hat{\mathbf{Z}})$, which is the PAC-Bayesian analog of the generalization error.

3. Stability

A key component of our analysis is the *stability* of the loss function. In this section, we introduce some definitions of stability and relate them to other forms found in the literature. Broadly speaking, stability ensures that changes to the input result in proportional changes in the output. In structured prediction, where inference is typically a global optimization over many interdependent variables, changing any single observation may affect many of the inferred values. The structured loss functions we consider *implicitly* require some form of joint inference; therefore, their stability is nontrivial. In this chapter, we introduce some definitions of stability and relate them to other forms found in the literature.

The following definitions will make use of the *Hamming distance*. For vectors $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}^n$, denote their Hamming distance by

$$D_{\text{H}}(\mathbf{z}, \mathbf{z}') \triangleq \sum_{i=1}^{n} \mathbb{1}\{z_i \neq z'_i\}.$$

3. Our definition of generalization error differs from some literature, in which the term is used to refer to the expected loss.

3.1 Uniform and Local Stability

Throughout this section, let $\mathcal{F} \triangleq \{\varphi : \mathcal{Z}^n \to \mathbb{R}\}$ denote an arbitrary class of functionals; e.g., $\mathcal{F}$ could be a structured loss function composed with a class of hypotheses. The definitions in this section describe notions of stability that hold either *uniformly* over the domain of $\mathcal{F}$ (for each $\varphi \in \mathcal{F}$), or *locally* over some subset of the domain (for some subset of $\mathcal{F}$).

Definition 1. We say that a function $\varphi \in \mathcal{F}$ is β-*uniformly stable* if, for any inputs $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}^n$,

$$\left|\varphi(\mathbf{z}) - \varphi(\mathbf{z}')\right| \leq \beta \, D_{\mathrm{H}}(\mathbf{z}, \mathbf{z}'). \tag{1}$$

Similarly, the class $\mathcal{F}$ is β-uniformly stable if every $\varphi \in \mathcal{F}$ is β-uniformly stable.

Equation 1 means that the change in the output should be proportional to the Hamming distance between the inputs. Put differently, a uniformly stable function is Lipschitz under the Hamming norm.

Uniform stability over the entire domain can be a strong requirement. Sometimes, stability only holds for a certain subset of inputs, such as points contained in a Euclidean ball of a certain radius. We refer to the set of inputs for which stability holds as the "good" set; all other inputs are "bad." The precise meaning of good and bad depends on the hypothesis class. Given some delineation of good and bad, we obtain the following localized notion of stability.

Definition 2. For a subset $\mathcal{B}_{\mathcal{Z}} \subseteq \mathcal{Z}^n$, we say that a function $\varphi \in \mathcal{F}$ is $(\beta, \mathcal{B}_{\mathcal{Z}})$-*locally stable* if Equation 1 holds for all $\mathbf{z}, \mathbf{z}' \notin \mathcal{B}_{\mathcal{Z}}$. The class $\mathcal{F}$ is $(\beta, \mathcal{B}_{\mathcal{Z}})$-locally stable if every $\varphi \in \mathcal{F}$ is $(\beta, \mathcal{B}_{\mathcal{Z}})$-locally stable.

Definition 2 has an alternate probabilistic interpretation. If $\mathbb{D}$ is a distribution on $\mathcal{Z}^n$, then Equation 1 holds with some probability over draws of $\mathbf{z}, \mathbf{z}' \sim \mathbb{D}$. If the bad set $\mathcal{B}_{\mathcal{Z}}$ has measure $\mathbb{D}(\mathcal{B}_{\mathcal{Z}}) \leq \nu$, then $(\beta, \mathcal{B}_{\mathcal{Z}})$-local stability is similar to, though slightly weaker than, the *strongly difference-bounded* property proposed by Kutin (2002). If φ is strongly difference-bounded, then Equation 1 must hold for any $\mathbf{z} \notin \mathcal{B}_{\mathcal{Z}}$ and $\mathbf{z}' \in \mathcal{Z}^n$ (which could be in $\mathcal{B}_{\mathcal{Z}}$). All functions that are strongly difference-bounded are locally stable, but the converse is not true.

The notion of probabilistic stability can be extended to distributions on the function class. For any stability parameter β (and bad inputs $\mathcal{B}_{\mathcal{Z}}$), the function class is partitioned into functions that satisfy Equation 1, and those that do not. Therefore, for any distribution $\mathbb{Q}$ on $\mathcal{F}$, uniform (or local) stability holds with some probability over draws of $\varphi \sim \mathbb{Q}$. This idea motivates the following definition.

Definition 3. Fix some $\beta \geq 0$ and $\mathcal{B}_{\mathcal{Z}} \subseteq \mathcal{Z}^n$, and let $\mathcal{B}_{\mathcal{F}} \subseteq \mathcal{F}$ denote the subset of functions that are not $(\beta, \mathcal{B}_{\mathcal{Z}})$-locally stable. We say that a distribution $\mathbb{Q}$ on $\mathcal{F}$ is $(\beta, \mathcal{B}_{\mathcal{Z}}, \eta)$-*locally stable* if $\mathbb{Q}(\mathcal{B}_{\mathcal{F}}) \leq \eta$.

Note the taxonomical relationship between these definitions. Definition 1 is the strongest condition, since it implies Definitions 2 and 3. Clearly, if $\mathcal{F}$ is β-uniformly stable, then it is $(\beta, \emptyset)$-locally and $(\beta, \emptyset, 0)$-locally stable. Definition 2 extends Definition 1 by accommodating broader domains. Definition 3 extends this even further, by accommodating classes in which only some functions satisfy local stability.

Definition 3 is particularly interesting in the PAC-Bayes framework, in which a predictor is selected at random according to a (learned) posterior distribution. With prior knowledge of the hypothesis class (and data distribution), a posterior can be constructed so as to place low mass on predictors that do not satisfy uniform or local stability. As we show in Section 6, this technique lets us relax certain restrictions on the hypothesis class.

Stability measures the change in the output relative to the change in the inputs. A related property is that the change in the output is bounded—i.e., the function has bounded range.

Definition 4. We say that $\varphi \in \mathcal{F}$ is *α-uniformly range-bounded* if, for any $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}^n$,

$$\left|\varphi(\mathbf{z}) - \varphi(\mathbf{z}')\right| \leq \alpha.$$

Range-boundedness is implied by stability, but the range constant, α, may be smaller than the upper bound implied by stability. Our analysis uses range-boundedness as a fall-back property when "good" stability does not hold.

3.2 Connections to Other Notions of Stability

In the learning theory literature, the word "stability" has traditionally been associated with a *learning* algorithm, rather than an inference algorithm. A learning algorithm is said to be stable with respect to a loss function if the loss of a learned hypothesis varies by a bounded amount upon replacing (or deleting) examples from the training set. This property has been used to derive generalization bounds (e.g., Bousquet and Elisseeff, 2002), similar to the way we use stability of inference. The key idea is that stability enables concentration of measure, which is central to generalization. That said, learning stability is distinct from inference stability, and neither property implies the other. Indeed, a learning algorithm might return hypotheses with drastically different losses for slightly different training sets, even if each hypothesis, composed with the loss function, is uniformly stable. Likewise, a stable learning algorithm might produce hypotheses with unstable loss.

Our definition of stability should also be contrasted with *sensitivity analysis*. Since the terms are often used interchangeably, we distinguish the two as follows: stability measures the amount of change induced in the *output* of a function upon perturbing its input within a certain range, and sensitivity analysis measures the amount of perturbation one can apply to the *input* such that its output remains within a certain range. By these definitions, one is the dual of the other. In the context of probabilistic inference, sensitivity analysis has been used to determine the maximum amount one can perturb the model parameters (or evidence) such that the likelihood of a query stays within a given tolerance, or such that the most likely assignment does not change (Chan and Darwiche, 2005, 2006). Stability measures *how much* the likelihood or most likely assignment changes.

Our first generalization bounds for structured prediction (London et al., 2013) crucially relied on a property we referred to as *uniform collective stability*. A class of vector-valued functions, $\mathcal{G} \triangleq \left\{g : \mathcal{Z}^n \to \mathbb{R}^N\right\}$, has *$\beta$-uniform collective stability* if, for any $g \in \mathcal{G}$, and any $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}^n$,

$$\left\|g(\mathbf{z}) - g(\mathbf{z}')\right\|_1 \leq \beta\, D_{\text{H}}(\mathbf{z}, \mathbf{z}').$$

We later relaxed this requirement to various non-uniform definitions of collective stability (London et al., 2014). Because collective stability implicitly involves the maximizing argu-

ment of a high-dimensional global optimization (i.e., a vector of predictions), we restricted our previous analyses to predictors with strongly convex inference objectives. Strong convexity let us bound the collective stability of a predictor, hence, the stability of its *output* composed with an *admissible*[4] loss function. Our new definitions involve the output of a *functional* (i.e., a scalar-valued function of multiple inputs), which essentially means that we are interested in the stability of the loss, instead of the collective stability of the predictions. In our new analysis, the loss function has access to the model and may use it for inference. However, the loss function may not require the same inference used for prediction. (For example, the losses considered in Section 6 use the maximum of the inference objective instead of the maximizing argument.) This framework lets us analyze a broad range of structured losses, without requiring strongly convex inference. Further, it can be shown that any predictor with "good" collective stability (such as one with a strongly convex inference objective), composed with an admissible loss function, satisfies our new definitions of stability. Therefore, our new definitions are strictly more general than collective stability.

4. Statistical Tools

Reasoning about the concentration of functions of dependent random variables requires sophisticated statistical machinery. In this section, we review some supporting definitions and introduce a quantity to summarize the amount of dependence in the data distribution. We use this quantity in a new moment-generating function inequality for locally stable functions of dependent random variables. We then provide some example conditions under which dependence is suitably bounded, thereby supporting improved generalization bounds.

4.1 Quantifying Dependence

For probability measures $\mathbb{P}$ and $\mathbb{Q}$ on a σ-algebra (i.e., event space) Σ, the *total variation distance* is

$$\|\mathbb{P} - \mathbb{Q}\|_{\text{TV}} \triangleq \sup_{A \in \Sigma} |\mathbb{P}(A) - \mathbb{Q}(A)| .$$

As a special case, when the sample space, Ω, is finite,

$$\|\mathbb{P} - \mathbb{Q}\|_{\text{TV}} = \frac{1}{2} \sum_{\omega \in \Omega} |\mathbb{P}(\omega) - \mathbb{Q}(\omega)| .$$

Let π be a permutation of $[n] \triangleq \{1, \ldots, n\}$, where $\pi(i)$ denotes the i^{th} element in the sequence and $\pi(i : j)$ denotes a subsequence of elements i through j. Used to index variables $\mathbf{Z} \triangleq (Z_i)_{i=1}^n$, denote by $Z_{\pi(i)}$ the i^{th} variable in the permutation and $\mathbf{Z}_{\pi(i:j)}$ the subsequence $(Z_{\pi(i)}, \ldots, Z_{\pi(j)})$.

Definition 5. We say that a sequence of permutations, $\boldsymbol{\pi} \triangleq (\pi_i)_{i=1}^n$, is a *filtration* if, for $i = 1, \ldots, n-1$,

$$\pi_i(1 : i) = \pi_{i+1}(1 : i).$$

Let $\Pi(n)$ denote the set of all filtrations for a given n.

4. See (London et al., 2013, 2014) for a precise definition of loss admissibility.

The following data structure quantifies the dependence between subsets of variables defined by a filtration.

Definition 6. Let $\mathbf{Z} \triangleq (Z_i)_{i=1}^n$ denote random variables with joint distribution $\mathbb{D}$ on $\mathcal{Z}^n$. Fix a filtration, $\boldsymbol{\pi} \in \Pi(n)$, and a set of inputs, $\mathcal{B}_{\mathcal{Z}} \subseteq \mathcal{Z}^n$. Let $\overline{\mathcal{B}}$ denote the event $\mathbf{Z} \notin \mathcal{B}_{\mathcal{Z}}$. For $i \in [n]$, let $\mathcal{Z}^i_{\boldsymbol{\pi},\overline{\mathcal{B}}}$ denote the subset of $\mathcal{Z}^i$ such that, for every $\mathbf{z} \in \mathcal{Z}^i_{\boldsymbol{\pi},\overline{\mathcal{B}}}$, $\mathbf{Z}_{\pi_i(1:i)} = \mathbf{z}$ is consistent with $\overline{\mathcal{B}}$. With a slight abuse of notation, for $\mathbf{z} \in \mathcal{Z}^{i-1}_{\boldsymbol{\pi},\overline{\mathcal{B}}}$, let $\mathcal{Z}^i_{\boldsymbol{\pi},\overline{\mathcal{B}}}(\mathbf{z})$ denote the subset of $\mathcal{Z}$ such that, for any $z \in \mathcal{Z}^i_{\boldsymbol{\pi},\overline{\mathcal{B}}}(\mathbf{z})$, $\mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z)$ is consistent with $\overline{\mathcal{B}}$. Then, for $i \in [n]$, $j > i$, $\mathbf{z} \in \mathcal{Z}^{i-1}_{\boldsymbol{\pi},\overline{\mathcal{B}}}$ and $z, z' \in \mathcal{Z}^i_{\boldsymbol{\pi},\overline{\mathcal{B}}}(\mathbf{z})$, we define the *$\vartheta$-mixing coefficients*[5] as

$$\vartheta^{\boldsymbol{\pi}}_{ij}(\mathbf{z}, z, z') \triangleq \left\| \mathbb{D}\left(\mathbf{Z}_{\pi_i(j:n)} \,|\, \overline{\mathcal{B}}, \mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z)\right) - \mathbb{D}\left(\mathbf{Z}_{\pi_i(j:n)} \,|\, \overline{\mathcal{B}}, \mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z')\right) \right\|_{\mathrm{TV}}.$$

We use these to define the upper-triangular *dependency matrix*, $\boldsymbol{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}} \in \mathbb{R}^{n \times n}$, with entries

$$\gamma^{\boldsymbol{\pi}}_{ij} \triangleq \begin{cases} 1 & \text{for } i = j, \\ \sup_{\mathbf{z} \in \mathcal{Z}^{i-1}_{\boldsymbol{\pi},\overline{\mathcal{B}}},\, z, z' \in \mathcal{Z}^i_{\boldsymbol{\pi},\overline{\mathcal{B}}}(\mathbf{z})} \vartheta^{\boldsymbol{\pi}}_{ij}(\mathbf{z}, z, z') & \text{for } i < j, \\ 0 & \text{for } i > j. \end{cases}$$

When $\mathcal{B}_{\mathcal{Z}} = \emptyset$, we simply omit the subscript notation.

Each ϑ-mixing coefficient measures the influence of some variable, $Z_{\pi_i(i)}$, on some subset, $\mathbf{Z}_{\pi_i(j:n)}$, given some assignment to the variables, $\mathbf{Z}_{\pi_i(1:i-1)}$, that preceded $Z_{\pi_i(i)}$ in the filtration; $\gamma^{\boldsymbol{\pi}}_{ij}$ measures the maximal influence of $Z_{\pi_i(i)}$ conditioned on any assignment to $\mathbf{Z}_{\pi_i(1:i-1)}$. Thus, to summarize the amount of dependence in the data distribution, we use the induced matrix infinity norm of $\boldsymbol{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}$, denoted

$$\left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}\right\|_\infty \triangleq \max_{i \in [n]} \sum_{j=1}^n \left|\gamma^{\boldsymbol{\pi}}_{ij}\right|,$$

which effectively measures the maximal aggregate influence of any single variable, given the filtration. Observe that, if $(Z_1, \ldots, Z_n)$ are mutually independent, then $\boldsymbol{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}$ is the identity matrix and $\left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}\right\|_\infty = 1$. At the other extreme, if $(Z_1, \ldots, Z_n)$ are deterministically dependent, then the top row of $\boldsymbol{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}$ is n, so $\left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}\right\|_\infty = n$.

Viewed through the lens of stability, $\left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}\right\|_\infty$ can be interpreted as measuring the stability of the data distribution. From this perspective, distributions with strong, long-range dependencies (when $\left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}\right\|_\infty$ is big) are unstable, whereas distributions with weak, localized dependence (when $\left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}\right\|_\infty$ is small) are stable. Intuitively, the same can be said for inference in MRFs; potentials that emphasize interactions between adjacent variables create long-range dependencies, which causes instability, whereas potentials that emphasize local signal make adjacent variables more independent, which promotes stability. Thus, dependence and stability are two sides of the same coin.

The filtration used to define $\boldsymbol{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}$ can have a strong impact on $\left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}\right\|_\infty$. Since we do not assume that $\mathbf{Z}$ corresponds to a temporal process, there may not be an obvious ordering of

5. The ϑ-mixing coefficients were introduced by Kontorovich and Ramanan (2008) as *η-mixing* and are related to the *maximal coupling coefficients* used by Chazottes et al. (2007).

the variables. However, the types of stochastic processes we are interested in are typically endowed with a topology. If the topology is a graph, the filtration can be determined by traversing the graph. For instance, for a Markov tree process, Kontorovich (2012) ordered the variables via a breadth-first traversal from the root; for an Ising model on a lattice, Chazottes et al. (2007) ordered the variables with a spiraling traversal from the origin. (Both of these examples used a static permutation of the variables, not a filtration.) If the filtration is determined by graph traversal, the ϑ-mixing coefficients can be viewed as measuring the strength of dependence as a function of graph distance. Viewed as such, $\left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}\right\|_\infty$ effectively captures the slowest decay of dependence along any traversal from a given set of traversals.

The aforementioned works (Kontorovich, 2012; Chazottes et al., 2007) showed that, for Markov trees and grids, under suitable contraction or temperature regimes, $\left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}\right\|_\infty$ is bounded independently of n (i.e., $\left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}\right\|_\infty = \mathrm{O}(1)$). By exploiting filtrations, we can show that the same holds for Markov random fields of any bounded-degree structure, provided the distribution exhibits suitable mixing. We discuss these conditions in Section 4.3.

4.2 A Moment-Generating Function Inequality for Local Stability

With the supporting definitions in mind, we now present a new moment-generating function inequality for locally stable functions of dependent random variables. The proof is provided in Appendix A.3.

Proposition 1. *Let $\mathbf{Z} \triangleq (Z_i)_{i=1}^n$ denote random variables with joint distribution $\mathbb{D}$ on $\mathcal{Z}^n$. Fix a set of "bad" inputs, $\mathcal{B}_\mathcal{Z} \subseteq \mathcal{Z}^n$, and let $\overline{\mathcal{B}}$ denote the event $\mathbf{Z} \notin \mathcal{B}_\mathcal{Z}$. Let $\varphi : \mathcal{Z}^n \to \mathbb{R}$ denote a measurable function with $(\beta, \mathcal{B}_\mathcal{Z})$-local stability. Then, for any $\tau \in \mathbb{R}$ and filtration $\boldsymbol{\pi} \in \Pi(n)$,*

$$\mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathbb{D}}\left[e^{\tau(\varphi(\mathbf{Z}) - \mathbb{E}[\varphi(\mathbf{Z}) \,|\, \overline{\mathcal{B}}])} \,|\, \overline{\mathcal{B}}\right] \leq \exp\left(\frac{\tau^2}{8}\, n\beta^2 \left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}\right\|_\infty^2\right).$$

This bound yields a novel concentration inequality for uniformly stable functions of dependent random variables, which we discuss in Appendix A.4. Though we will not use this corollary in our analysis, it may be of independent interest.

Proposition 1 builds on work by Samson (2000), Chazottes et al. (2007) and Kontorovich and Ramanan (2008). Our analysis differs from theirs in that we accommodate functions that are not uniformly stable. In this respect, our analysis is similar to that of Kutin (2002) and Vu (2002), though these works assume independence between variables. Because we allow interdependence—as well as other technical challenges, related to our definitions of local stability—we do not use the same proof techniques as the aforementioned works.

4.3 Bounded Dependence Conditions

The infinity norm of the dependency matrix has a trivial upper bound, $\left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}\right\|_\infty \leq n$. However, we are interested in bounds that are sub-logarithmic in (or, even better, independent of) n. In this section, we describe some general settings in which $\left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}\right\|_\infty$ has a nontrivial upper bound.

For the remainder of this section, let $\mathbf{Z} \triangleq (Z_i)_{i=1}^n$ denote random variables with joint distribution $\mathbb{D}$ on $\mathcal{Z}^n$. Assume that $\mathbb{D}$ is associated with a graph, $G \triangleq (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} \triangleq [n]$ indexes $\mathbf{Z}$.

We use the following notion of distance-based dependence. For simplicity of exposition, we assume that $\mathcal{B}_{\mathcal{Z}} = \emptyset$, so we can omit $\overline{\mathcal{B}}$ from the following notation.

Definition 7. For any two subsets, $\mathcal{S}, \mathcal{T} \subseteq \mathcal{V}$, we define their graph distance, $D_{\mathrm{G}}(\mathcal{S}, \mathcal{T})$, as the length of the shortest path from any node in $\mathcal{S}$ to any node in $\mathcal{T}$. We then define the *distance-based ϑ-mixing coefficients* as

$$\vartheta(k) \triangleq \sup_{\substack{\mathcal{S} \subseteq \mathcal{V},\ i \in \mathcal{S} \\ \mathcal{T} \subseteq \mathcal{V} \setminus \mathcal{S} : D_{\mathrm{G}}(i, \mathcal{T}) \geq k \\ \mathbf{z} \in \mathcal{Z}^{|\mathcal{S}|-1},\ z, z' \in \mathcal{Z}}} \left\| \mathbb{D}\left(\mathbf{Z}_{\mathcal{T}} \,|\, \mathbf{Z}_{\mathcal{S}} = \mathbf{z}, Z_i = z\right) - \mathbb{D}\left(\mathbf{Z}_{\mathcal{T}} \,|\, \mathbf{Z}_{\mathcal{S}} = \mathbf{z}, Z_i = z'\right) \right\|_{\mathrm{TV}},$$

where $\vartheta(0) \triangleq 1$.

The distance-based ϑ-mixing coefficients upper-bound the maximum influence exerted by any subset of the variables on any other subset that is separated by graph distance at least k. The sequence $(\vartheta(0), \vartheta(1), \vartheta(2), \ldots)$ roughly measures how dependence decays with graph distance. Note that $\vartheta(k)$ uniformly upper-bounds $\vartheta_{ij}^{\boldsymbol{\pi}}$ when $D_{\mathrm{G}}(\pi_i(i), \pi_i(j:n)) \geq k$. Therefore, for each upper-triangular entry of $\boldsymbol{\Gamma}^{\boldsymbol{\pi}}$, we have that

$$\gamma_{ij}^{\boldsymbol{\pi}} \leq \vartheta\left(D_{\mathrm{G}}\left(\pi_i(i), \pi_i(j:n)\right)\right).$$

Using the distance-based ϑ-mixing coefficients, we now show that, for certain Markov random fields, it is possible to upper-bound $\left\|\boldsymbol{\Gamma}_{\overline{\mathcal{B}}}^{\boldsymbol{\pi}}\right\|_\infty$ independently of n.

Proposition 2. *Suppose $\mathbb{D}$ is defined by an MRF, such that its graph, G, has maximum degree Δ_G. For any positive constant $\epsilon > 0$, if $\mathbb{D}$ admits a distance-based ϑ-mixing sequence such that, for all $k \geq 1$, $\vartheta(k) \leq (\Delta_G + \epsilon)^{-k}$, then there exists a filtration $\boldsymbol{\pi}$ such that*

$$\left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}\right\|_\infty \leq 1 + \Delta_G/\epsilon.$$

The proof is provided in Appendix A.5.

Uniformly geometric distance-based ϑ-mixing may seem like a restrictive condition. However, our analysis is overly pessimistic, in that it ignores the structure of the MRF beyond simply the maximum degree of the graph. Further, it does not take advantage of the actual conditional independencies present in the distribution. Nevertheless, there is a natural interpretation of the above conditions that follows from considering the mixing coefficients at distance 1: for the immediate neighbors of a node—i.e., its Markov blanket—its ϑ-mixing coefficient must be less than $1/\Delta_G$. This loosely means that the combination of all incoming influence must be less than 1, implying that there is sufficiently strong influence from local observations.

Another important setting is when the graph is a chain. Chain-structured stochastic processes (usually temporal) under various mixing assumptions have been well-studied (see Bradley, 2005 for a comprehensive survey). It can be shown that any *contracting* Markov chain has $\left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}\right\|_\infty = \mathrm{O}(1)$ (Kontorovich, 2012). Here, we provide an alternate condition, using distance-based ϑ-mixing, under which the dependency matrix of a Markov chain has

suitably low norm. The key property of a chain graph is that the number of nodes at distance k from any starting node is constant. We can therefore relax the assumption of geometric decay used in the previous result.

Proposition 3. *Suppose $\mathbb{D}$ is an undirected Markov chain (i.e., chain-structured MRF) of length n. For any constants $\epsilon > 0$ and $p \geq 1$, if $\mathbb{D}$ admits a distance-based ϑ-mixing sequence such that, for all $k \geq 1$, $\vartheta(k) \leq \epsilon k^{-p}$, then there exists a filtration, $\boldsymbol{\pi}$, such that*

$$\left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}\right\|_\infty \leq \begin{cases} 1 + \epsilon\,(1 + \ln(n-1)) & \text{if } p = 1, \\ 1 + \epsilon\,\zeta(p) & \text{if } p > 1, \end{cases}$$

where $\zeta(p) \triangleq \sum_{j=1}^{\infty} j^{-p}$ is the Riemann zeta function.

The proof is provided in Appendix A.6.

For $p > 1$, the Riemann function converges to a constant. For example, $\zeta(2) = \pi^2/6 \approx 1.645$. However, even $p = 1$ yields a sufficiently low growth rate. In the following section, we prove generalization bounds of the form $\mathrm{O}\left(\left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}\right\|_\infty / \sqrt{mn}\right)$, which still converges if $\left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}\right\|_\infty = \mathrm{O}(\ln n)$, albeit at a slower rate.

5. PAC-Bayes Bounds

We now present some new PAC-Bayes generalization bounds using the stability definitions from Section 3. The first theorem is stated for a given stability parameter, β. We then generalize this result to hold for all β simultaneously, meaning β can depend on the posterior. We conclude this section with a general technique for derandomizing the bounds based on stability.

It will help to begin with a high-level sketch of the analysis, which we specialize to various settings in Sections 5.1 and 5.2. It will help to view the composition of the loss function, L, and the hypothesis class, $\mathcal{H}$, as a family of functions, $L \circ \mathcal{H} = \{L(h, \cdot) : h \in \mathcal{H}\}$. If $\mathbb{Q}$ is a distribution on $\mathcal{H}$, it is also a distribution on $L \circ \mathcal{H}$. Each member of $L \circ \mathcal{H}$ is a random function, determined by the draw of $h \sim \mathbb{Q}$. Further, when $L(h, \cdot)$ is composed with a training set $\hat{\mathbf{Z}} \sim \mathbb{D}^m$ in $\hat{L}(h, \cdot)$, the generalization error, $\overline{L}(h) - \hat{L}(h, \hat{\mathbf{Z}})$, becomes a centered random variable. Part of our analysis involves bounding the moment-generating function of this random variable, and to do so requires the notions of stability from Section 3. The stability of $L(h, \cdot)$ is determined by h, so the "bad" members of $L \circ \mathcal{H}$ are in fact the "bad" hypotheses (for the given loss function).

Let $\hat{\mathbf{Z}} \triangleq (\mathbf{Z}^{(l)})_{l=1}^m$ denote a training set of m structured examples, distributed according to $\mathbb{D}^m$. Fix some $\beta \geq 0$ and a set of bad inputs $\mathcal{B}_{\mathcal{Z}}$, with measure $\nu \triangleq \mathbb{D}(\mathcal{B}_{\mathcal{Z}})$. Implicitly, the pair $(\beta, \mathcal{B}_{\mathcal{Z}})$ fixes a set of hypotheses $\mathcal{B}_{\mathcal{H}} \subseteq \mathcal{H}$ for which $L(h, \cdot)$ does not satisfy Equation 1 with $\beta' \triangleq \beta/n$ and $\mathcal{B}_{\mathcal{Z}}$. For the time being, $\mathcal{B}_{\mathcal{H}}$ is independent of $\mathbb{Q}$. Fix a prior $\mathbb{P}$ and posterior $\mathbb{Q}$ on $\mathcal{H}$. (We will later consider all posteriors.) We define a convenience function,

$$\tilde{\phi}(h, \hat{\mathbf{Z}}) \triangleq \begin{cases} \mathbb{E}_{\mathbf{Z} \sim \mathbb{D}}\left[L(h, \mathbf{Z}) \,|\, \overline{\mathcal{B}}\right] - \hat{L}(h, \hat{\mathbf{Z}}) & \text{if } h \notin \mathcal{B}_{\mathcal{H}}, \\ 0 & \text{otherwise,} \end{cases}$$

where $\overline{\mathcal{B}}$ denotes the event $\mathbf{Z} \notin \mathcal{B}_{\mathcal{Z}}$. First, for any uniformly bounded random variable, with $|X| \leq b$, and some event, E,

$$\mathbb{E}[X] = \mathbb{E}[X\,\mathbb{1}\{E\}] + \mathbb{E}[X\,\mathbb{1}\{\neg E\}] \leq b \Pr\{E\} + \mathbb{E}[X\,\mathbb{1}\{\neg E\}].$$

We use this identity to show that, if $L \circ \mathcal{H}$ is α-uniformly range-bounded, and $\mathbb{Q}$ is $(\beta/n, \mathcal{B}_{\mathcal{Z}}, \eta)$-locally stable, then

$$\overline{L}(\mathbb{Q}) - \hat{L}(\mathbb{Q}, \hat{\mathbf{Z}}) \leq \alpha\eta + \alpha\nu + \mathop{\mathbb{E}}_{h\sim\mathbb{Q}}\left[\tilde{\phi}(h, \hat{\mathbf{Z}})\right].$$

To bound the $\mathbb{E}_{h\sim\mathbb{Q}}\big[\tilde{\phi}(h, \hat{\mathbf{Z}})\big]$, we use Donsker and Varadhan's (1975) *change of measure* inequality.

Lemma 1. *For any measurable function $\varphi : \Omega \to \mathbb{R}$, and any two distributions, $\mathbb{P}$ and $\mathbb{Q}$, on Ω,*

$$\mathop{\mathbb{E}}_{\omega\sim\mathbb{P}}[\varphi(\omega)] \leq D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q}) + \ln \mathop{\mathbb{E}}_{\omega\sim\mathbb{Q}}\left[e^{\varphi(\omega)}\right].$$

(McAllester (2003) provides a straightforward proof.) Using Lemma 1, for any free parameter $u \geq 0$, we have that

$$\mathop{\mathbb{E}}_{h\sim\mathbb{Q}}\left[\tilde{\phi}(h, \hat{\mathbf{Z}})\right] \leq \frac{1}{u}\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln \mathop{\mathbb{E}}_{h\sim\mathbb{P}}\left[e^{u\tilde{\phi}(h, \hat{\mathbf{Z}})}\right]\right).$$

Combining the above inequalities yields

$$\overline{L}(\mathbb{Q}) - \hat{L}(\mathbb{Q}, \hat{\mathbf{Z}}) \leq \alpha\eta + \alpha\nu + \frac{1}{u}\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln \mathop{\mathbb{E}}_{h\sim\mathbb{P}}\left[e^{u\tilde{\phi}(h, \hat{\mathbf{Z}})}\right]\right).$$

The remainder of the analysis concerns how to bound $\mathbb{E}_{h\sim\mathbb{P}}\big[e^{u\tilde{\phi}(h, \hat{\mathbf{Z}})}\big]$ and how to optimize u. For the first task, we combine Markov's inequality with the moment-generating function bound from Section 4.2. Optimizing u takes some care, since we would like the bounds to hold simultaneously for all posteriors. We therefore adopt a discretization technique (Seldin et al., 2012) that approximately optimizes the bound for all posteriors. We use a similar technique to obtain bounds that hold for all β.

5.1 Fixed Stability Bounds

In the following theorem, we derive a new PAC-Bayes bound for posteriors with local stability, with β fixed. Fixing β means that the set of "bad" hypotheses is determined by the characteristics of the hypothesis class independently of the posterior.

Theorem 1. *Fix $m \geq 1$, $n \geq 1$, $\boldsymbol{\pi} \in \Pi(n)$, $\delta \in (0, 1)$, $\alpha \geq 0$ and $\beta \geq 0$. Fix a distribution, $\mathbb{D}$, on $\mathcal{Z}^n$. Fix a set of bad inputs, $\mathcal{B}_{\mathcal{Z}}$, with $\nu \triangleq \mathbb{D}(\mathcal{B}_{\mathcal{Z}})$. Let $\boldsymbol{\Gamma}^{\boldsymbol{\pi}}_{\mathcal{B}}$ denote the dependency matrix induced by $\mathbb{D}$, $\boldsymbol{\pi}$ and $\mathcal{B}_{\mathcal{Z}}$. Fix a prior, $\mathbb{P}$, on a hypothesis class, $\mathcal{H}$. Fix a loss function, L, such that $L \circ \mathcal{H}$ is α-uniformly range-bounded. Then, with probability at least $1 - \delta - m\nu$ over realizations of a training set, $\hat{\mathbf{Z}} \triangleq (\mathbf{Z}^{(l)})_{l=1}^{m}$, drawn according to $\mathbb{D}^m$, the following hold: 1) for all $l \in [m]$, $\mathbf{Z}^{(l)} \notin \mathcal{B}_{\mathcal{Z}}$; 2) for all $\eta \in [0, 1]$ and posteriors $\mathbb{Q}$ with $(\beta/n, \mathcal{B}_{\mathcal{Z}}, \eta)$-local stability,*

$$\overline{L}(\mathbb{Q}) \leq \hat{L}(\mathbb{Q}, \hat{\mathbf{Z}}) + \alpha(\eta + \nu) + 2\beta \left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}_{\mathcal{B}}\right\|_{\infty} \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta}}{2mn}}. \tag{2}$$

To interpret the bound, suppose $\alpha = \mathrm{O}(1)$, $\beta = \mathrm{O}(1)$, and that the data distribution is weakly dependent, with $\|\mathbf{\Gamma}^{\boldsymbol{\pi}}\|_\infty = \mathrm{O}(1)$. We would then have that the generalization error decreases at a rate of $\mathrm{O}\left(\eta + \nu + (mn)^{-1/2}\right)$. Since η is a function of the posterior, we can reasonably assume that $\eta = \mathrm{O}\left((mn)^{-1/2}\right)$. (Section 6 provides examples of this.) However, while ν may be proportional to n, it is unreasonable to believe that ν will decrease with m, since $\mathbb{D}$ is almost certainly agnostic to the number of training examples. Thus, Theorem 1 is interesting when either ν is negligible, or when m is a small constant.

It can be shown that any hypothesis class with *collective* stability, composed with a suitable loss function, satisfies the conditions of the bound. Thus, Theorem 1 is strictly more general than our prior PAC-Bayes bounds (London et al., 2014). Moreover, Theorem 1 easily applies to compositions with uniform stability, since $\mathbb{Q}(\mathcal{B}_\mathcal{H}) = 0$ for all posteriors. This insight yields the following corollary.

Corollary 1. *Suppose $L \circ \mathcal{H}$ is (β/n)-uniformly stable. Then, with probability at least $1-\delta$ over realizations of $\hat{\mathbf{Z}}$, for all $\mathbb{Q}$,*

$$\overline{L}(\mathbb{Q}) \leq \hat{L}(\mathbb{Q}, \hat{\mathbf{Z}}) + 2\beta \left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}\right\|_\infty \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln \frac{2}{\delta}}{2mn}}. \tag{3}$$

As we show in Section 6.1.2, Corollary 1 is useful when the hypothesis class and instance space are uniformly bounded. Even when this property does not hold, we obtain an identical bound for all posteriors with $(\beta/n, \emptyset, 0)$-local stability, meaning the support of the posterior is (β/n)-uniformly stable. However, this condition is less useful, since it is assumed that the posterior construction puts nonzero density on a learned hypothesis, which may not satisfy uniform stability for a fixed β.

It is worth noting that, if the hypothesis class does not use joint inference—for example, if a global prediction, $h(\mathbf{X})$, is in fact a set of independent, local predictions, $(h(X_1), \ldots, h(X_n))$—and the loss function decomposes over the labels, then uniform stability is trivially satisfied. In this case, Corollary 1 produces a PAC-Bayes bound for learning traditional predictors from interdependent data. If we further have that $(Z_1, \ldots, Z_n)$ are independent and identically distributed (i.i.d.)—for instance, if they represent "micro examples" drawn independently from some target distribution—then Corollary 1 reduces to standard PAC-Bayes bounds for learning from i.i.d. data (e.g., McAllester, 1999).

We now prove Theorem 1.

Proof (Theorem 1) We begin by defining two convenience functions,

$$\phi(h, \hat{\mathbf{Z}}) \triangleq \overline{L}(h) - \hat{L}(h, \hat{\mathbf{Z}}) \tag{4}$$

$$\text{and} \quad \tilde{\phi}(h, \hat{\mathbf{Z}}) \triangleq \begin{cases} \mathbb{E}_{\mathbf{Z}\sim\mathbb{D}}\left[L(h, \mathbf{Z}) \,|\, \overline{\mathcal{B}}\right] - \hat{L}(h, \hat{\mathbf{Z}}) & \text{if } h \notin \mathcal{B}_\mathcal{H}, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

If $L \circ \mathcal{H}$ is α-uniformly range-bounded (Definition 4), then, for any $h \in \mathcal{H}$,

$$\begin{aligned}\phi(h, \hat{\mathbf{Z}}) &= \frac{1}{m}\sum_{l=1}^{m} \overline{L}(h) - L(h, \mathbf{Z}^{(l)}) \\ &\leq \frac{1}{m}\sum_{l=1}^{m} \sup_{\mathbf{z}\in\mathcal{Z}^n} \left| L(h, \mathbf{z}) - L(h, \mathbf{Z}^{(l)}) \right| \\ &\leq \frac{1}{m}\sum_{l=1}^{m} \alpha = \alpha. \end{aligned} \tag{6}$$

It follows that

$$\begin{aligned}\phi(h, \hat{\mathbf{Z}}) &= \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathbb{D}} \left[L(h, \mathbf{Z}) - \hat{L}(h, \hat{\mathbf{Z}}) \right] \\ &= \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathbb{D}} \left[\left(L(h, \mathbf{Z}) - \hat{L}(h, \hat{\mathbf{Z}}) \right) \mathbb{1}\{\mathbf{Z} \notin \mathcal{B}_{\mathcal{Z}}\} \right] + \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathbb{D}} \left[\left(L(h, \mathbf{Z}) - \hat{L}(h, \hat{\mathbf{Z}}) \right) \mathbb{1}\{\mathbf{Z} \in \mathcal{B}_{\mathcal{Z}}\} \right] \\ &\leq \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathbb{D}} \left[\left(L(h, \mathbf{Z}) - \hat{L}(h, \hat{\mathbf{Z}}) \right) \mathbb{1}\{\mathbf{Z} \notin \mathcal{B}_{\mathcal{Z}}\} \right] + \alpha \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathbb{D}} [\mathbb{1}\{\mathbf{Z} \in \mathcal{B}_{\mathcal{Z}}\}] \\ &\leq \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathbb{D}} \left[\left(L(h, \mathbf{Z}) - \hat{L}(h, \hat{\mathbf{Z}}) \right) \mathbb{1}\{\mathbf{Z} \notin \mathcal{B}_{\mathcal{Z}}\} \right] + \alpha\nu \\ &= \Pr_{\mathbf{Z}\sim\mathbb{D}} \{\mathbf{Z} \notin \mathcal{B}_{\mathcal{Z}}\} \left(\mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathbb{D}} \left[L(h, \mathbf{Z}) \,|\, \overline{\mathcal{B}} \right] - \hat{L}(h, \hat{\mathbf{Z}}) \right) + \alpha\nu \\ &\leq \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathbb{D}} \left[L(h, \mathbf{Z}) \,|\, \overline{\mathcal{B}} \right] - \hat{L}(h, \hat{\mathbf{Z}}) + \alpha\nu. \end{aligned} \tag{7}$$

Moreover, for any posterior $\mathbb{Q}$ with $(\beta/n, \mathcal{B}_{\mathcal{Z}}, \eta)$-local stability,

$$\begin{aligned}\overline{L}(\mathbb{Q}) - \hat{L}(\mathbb{Q}, \hat{\mathbf{Z}}) &= \mathop{\mathbb{E}}_{h\sim\mathbb{Q}} \left[\phi(h, \hat{\mathbf{Z}}) \right] \\ &= \mathop{\mathbb{E}}_{h\sim\mathbb{Q}} \left[\phi(h, \hat{\mathbf{Z}}) \, \mathbb{1}\{h \in \mathcal{B}_{\mathcal{H}}\} \right] + \mathop{\mathbb{E}}_{h\sim\mathbb{Q}} \left[\phi(h, \hat{\mathbf{Z}}) \, \mathbb{1}\{h \notin \mathcal{B}_{\mathcal{H}}\} \right] \\ &\leq \alpha \mathop{\mathbb{E}}_{h\sim\mathbb{Q}} [\mathbb{1}\{h \in \mathcal{B}_{\mathcal{H}}\}] + \mathop{\mathbb{E}}_{h\sim\mathbb{Q}} \left[\phi(h, \hat{\mathbf{Z}}) \, \mathbb{1}\{h \notin \mathcal{B}_{\mathcal{H}}\} \right] \\ &\leq \alpha\eta + \mathop{\mathbb{E}}_{h\sim\mathbb{Q}} \left[\phi(h, \hat{\mathbf{Z}}) \, \mathbb{1}\{h \notin \mathcal{B}_{\mathcal{H}}\} \right] \\ &\leq \alpha\eta + \alpha\nu + \mathop{\mathbb{E}}_{h\sim\mathbb{Q}} \left[\tilde{\phi}(h, \hat{\mathbf{Z}}) \right]. \end{aligned} \tag{8}$$

Then, for any $u \in \mathbb{R}$, using Lemma 1, we have that

$$\begin{aligned}\overline{L}(\mathbb{Q}) - \hat{L}(\mathbb{Q}, \hat{\mathbf{Z}}) &\leq \alpha\eta + \alpha\nu + \frac{1}{u} \mathop{\mathbb{E}}_{h\sim\mathbb{Q}} \left[u\, \tilde{\phi}(h, \hat{\mathbf{Z}}) \right] \\ &\leq \alpha\eta + \alpha\nu + \frac{1}{u} \left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln \mathop{\mathbb{E}}_{h\sim\mathbb{P}} \left[e^{u\tilde{\phi}(h, \hat{\mathbf{Z}})} \right] \right). \end{aligned} \tag{9}$$

Since u cannot depend on $(\eta, \mathbb{Q})$, we define it in terms of fixed quantities. For $j = 0, 1, 2, \ldots$, let $\delta_j \triangleq \delta 2^{-(j+1)}$, let

$$u_j \triangleq 2^j \sqrt{\frac{8mn \ln \frac{2}{\delta}}{\beta^2 \left\| \mathbf{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}} \right\|_\infty^2}}, \tag{10}$$

and define an event,

$$E_j \triangleq \mathbb{1}\left\{ \mathop{\mathbb{E}}_{h\sim\mathbb{P}}\left[e^{u_j\tilde{\phi}(h,\hat{\mathbf{Z}})}\right] \geq \frac{1}{\delta_j}\exp\left(\frac{u_j^2\beta^2\left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}\right\|_\infty^2}{8mn}\right)\right\}. \tag{11}$$

Note that u_j and E_j are independent of $(\eta, \mathbb{Q})$, since β (hence, $\mathcal{B}_\mathcal{H}$) is fixed. Let $E \triangleq \bigcup_{j=0}^\infty E_j$ denote the event that any E_j occurs. We also define an event

$$B \triangleq \bigcup_{l=1}^m \mathbb{1}\left\{\mathbf{Z}^{(l)} \in \mathcal{B}_\mathcal{Z}\right\}, \tag{12}$$

which indicates that at least one of the training examples is "bad." Using the law of total probability and the union bound, we then have that

$$\begin{aligned}
\Pr_{\hat{\mathbf{Z}}\sim\mathbb{D}^m}\{B\cup E\} &= \Pr_{\hat{\mathbf{Z}}\sim\mathbb{D}^m}\{B\} + \Pr_{\hat{\mathbf{Z}}\sim\mathbb{D}^m}\{E\cap\neg B\} \\
&\leq \Pr_{\hat{\mathbf{Z}}\sim\mathbb{D}^m}\{B\} + \Pr_{\hat{\mathbf{Z}}\sim\mathbb{D}^m}\{E\,|\,\neg B\} \\
&\leq \sum_{l=1}^m \Pr_{\mathbf{Z}^{(l)}\sim\mathbb{D}}\{\mathbf{Z}^{(l)}\in\mathcal{B}_\mathcal{Z}\} + \sum_{j=0}^\infty \Pr_{\hat{\mathbf{Z}}\sim\mathbb{D}^m}\{E_j\,|\,\neg B\} \\
&\leq m\nu + \sum_{j=0}^\infty \Pr_{\hat{\mathbf{Z}}\sim\mathbb{D}^m}\{E_j\,|\,\neg B\}.
\end{aligned} \tag{13}$$

The last inequality follows from the definition of ν. Then, using Markov's inequality, and rearranging the expectations, we have that

$$\Pr_{\hat{\mathbf{Z}}\sim\mathbb{D}^m}\{E_j\,|\,\neg B\} \leq \delta_j\exp\left(-\frac{u_j^2\beta^2\left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}\right\|_\infty^2}{8mn}\right)\mathop{\mathbb{E}}_{h\sim\mathbb{P}}\mathop{\mathbb{E}}_{\hat{\mathbf{Z}}\sim\mathbb{D}^m}\left[e^{u_j\tilde{\phi}(h,\hat{\mathbf{Z}})}\,|\,\neg B\right]. \tag{14}$$

Let

$$\varphi(h,\mathbf{Z}) \triangleq \begin{cases} \frac{1}{m}\left(\mathbb{E}_{\mathbf{Z}'\sim\mathbb{D}}\left[L(h,\mathbf{Z}')\,|\,\overline{\mathcal{B}}\right] - L(h,\mathbf{Z})\right) & \text{if } h\notin\mathcal{B}_\mathcal{H}, \\ 0 & \text{otherwise,}\end{cases} \tag{15}$$

and note that $\tilde{\phi}(h,\hat{\mathbf{Z}}) = \sum_{l=1}^m \varphi(h,\mathbf{Z}^{(l)})$. Then, since $\mathbf{Z}^{(1)},\ldots,\mathbf{Z}^{(m)}$ are independent and identically distributed, we can write the inner expectation over $\hat{\mathbf{Z}}$ as

$$\begin{aligned}
\mathop{\mathbb{E}}_{\hat{\mathbf{Z}}\sim\mathbb{D}^m}\left[e^{u_j\tilde{\phi}(h,\hat{\mathbf{Z}})}\,|\,\neg B\right] &= \prod_{l=1}^m \mathop{\mathbb{E}}_{\mathbf{Z}^{(l)}\sim\mathbb{D}}\left[e^{u_j\varphi(h,\mathbf{Z}^{(l)})}\,|\,\neg B\right] \\
&= \prod_{l=1}^m \mathop{\mathbb{E}}_{\mathbf{Z}^{(l)}\sim\mathbb{D}}\left[e^{u_j\varphi(h,\mathbf{Z}^{(l)})}\,|\,\mathbf{Z}^{(l)}\notin\mathcal{B}_\mathcal{Z}\right] \\
&= \prod_{l=1}^m \mathop{\mathbb{E}}_{\mathbf{Z}^{(l)}\sim\mathbb{D}}\left[e^{u_j\varphi(h,\mathbf{Z}^{(l)})}\,|\,\overline{\mathcal{B}}\right].
\end{aligned} \tag{16}$$

By construction, $\varphi(h,\cdot)$ outputs zero whenever $h\in\mathcal{B}_\mathcal{H}$. In these cases, $\varphi(h,\cdot)$ trivially satisfies uniform stability, which implies local stability. Further, if $\mathbb{Q}$ is $(\beta/n, \mathcal{B}_\mathcal{Z}, \eta)$-locally

stable, then every $L(h,\cdot) : h \notin \mathcal{B}_{\mathcal{H}}$ is $(\beta/n, \mathcal{B}_{\mathcal{Z}})$-locally stable, and it is easily verified that $\varphi(h,\cdot) : h \notin \mathcal{B}_{\mathcal{H}}$ is $(\beta/(mn), \mathcal{B}_{\mathcal{Z}})$-locally stable. Thus, $\varphi(h,\cdot) : h \in \mathcal{H}$ is $(\beta/(mn), \mathcal{B}_{\mathcal{Z}})$-locally stable. Since $\mathbb{E}_{\mathbf{Z}\sim\mathbb{D}}[\varphi(h,\mathbf{Z}) \,|\, \overline{\mathcal{B}}] = 0$, we therefore apply Proposition 1 and have, for all $h \in \mathcal{H}$,

$$\mathop{\mathbb{E}}_{\mathbf{Z}^{(l)}\sim\mathbb{D}} \left[e^{u_j \varphi(h,\mathbf{Z}^{(l)})} \,|\, \overline{\mathcal{B}} \right] \leq \exp\left(\frac{u_j^2 \beta^2 \left\| \mathbf{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}} \right\|_\infty^2}{8m^2 n} \right). \tag{17}$$

Combining Equations 14, 16 and 17, we have that

$$\begin{aligned}
\Pr_{\hat{\mathbf{Z}}\sim\mathbb{D}^m} \{E_j \,|\, \neg B\} &\leq \delta_j \exp\left(-\frac{u_j^2 \beta^2 \left\| \mathbf{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}} \right\|_\infty^2}{8mn} \right) \mathop{\mathbb{E}}_{h\sim\mathbb{P}} \left[\prod_{l=1}^m \mathop{\mathbb{E}}_{\mathbf{Z}^{(l)}\sim\mathbb{D}} \left[e^{u_j \varphi(h,\mathbf{Z}^{(l)})} \,|\, \overline{\mathcal{B}} \right] \right] \\
&\leq \delta_j \exp\left(-\frac{u_j^2 \beta^2 \left\| \mathbf{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}} \right\|_\infty^2}{8mn} \right) \mathop{\mathbb{E}}_{h\sim\mathbb{P}} \left[\prod_{l=1}^m \exp\left(\frac{u_j^2 \beta^2 \left\| \mathbf{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}} \right\|_\infty^2}{8m^2 n} \right) \right] \\
&= \delta_j \exp\left(-\frac{u_j^2 \beta^2 \left\| \mathbf{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}} \right\|_\infty^2}{8mn} \right) \exp\left(\frac{u_j^2 \beta^2 \left\| \mathbf{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}} \right\|_\infty^2}{8mn} \right) = \delta_j.
\end{aligned} \tag{18}$$

Then, combining Equations 13 and 18, and using the geometric series identity, we have that

$$\Pr_{\hat{\mathbf{Z}}\sim\mathbb{D}^m} \{B \cup E\} \leq m\nu + \sum_{j=0}^{\infty} \delta_j = m\nu + \delta \sum_{j=0}^{\infty} 2^{-(j+1)} = m\nu + \delta.$$

Thus, with probability at least $1 - \delta - m\nu$ over realizations of $\hat{\mathbf{Z}}$, every $l \in [m]$ satisfies $\mathbf{Z}^{(l)} \notin \mathcal{B}_{\mathcal{Z}}$, and every u_j satisfies

$$\mathop{\mathbb{E}}_{h\sim\mathbb{P}} \left[e^{u_j \tilde{\phi}(h,\hat{\mathbf{Z}})} \right] \leq \frac{1}{\delta_j} \exp\left(\frac{u_j^2 \beta^2 \left\| \mathbf{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}} \right\|_\infty^2}{8mn} \right). \tag{19}$$

We now show how to select j for any particular posterior $\mathbb{Q}$. Let

$$j^\star \triangleq \left\lfloor \frac{1}{2 \ln 2} \ln\left(\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})}{\ln(2/\delta)} + 1 \right) \right\rfloor, \tag{20}$$

and note that $j^\star \geq 0$. For all $v \in \mathbb{R}$, we have that $v - 1 \leq \lfloor v \rfloor \leq v$, and $2^{\ln v} = v^{\ln 2}$. We apply these identities to Equation 20 to show that

$$\frac{1}{2} \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})}{\ln(2/\delta)} + 1} \leq 2^{j^\star} \leq \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})}{\ln(2/\delta)} + 1},$$

implying

$$\sqrt{\frac{2mn\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta}\right)}{\beta^2 \left\| \mathbf{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}} \right\|_\infty^2}} \leq u_{j^\star} \leq \sqrt{\frac{8mn\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta}\right)}{\beta^2 \left\| \mathbf{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}} \right\|_\infty^2}}. \tag{21}$$

Further, by definition of $\delta_{j^\star}$,

$$\begin{aligned} D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{1}{\delta_{j^\star}} &= D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta} + j^\star \ln 2 \\ &\leq D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta} + \frac{\ln 2}{2\ln 2}\ln\left(\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})}{\ln(2/\delta)} + 1\right) \\ &= D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta} + \frac{1}{2}\ln\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta}\right) - \frac{1}{2}\ln\ln\frac{2}{\delta} \\ &\leq D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta} + \frac{1}{2}\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta}\right). \end{aligned} \tag{22}$$

The last inequality uses the identity $v - \ln\ln(2/\delta) \leq v + 1 \leq e^v$, for all $v \in \mathbb{R}$ and $\delta \in (0,1)$. It can be shown that this bound is approximately optimal, in that it is at most twice what it would be for a fixed posterior.

Putting it all together, we now have that, with probability at least $1 - \delta - m\nu$, the approximately optimal $(u_{j^\star}, \delta_{j^\star})$ for any posterior $\mathbb{Q}$ satisfies

$$\begin{aligned} \overline{L}(\mathbb{Q}) - \hat{L}(\mathbb{Q}, \hat{\mathbf{Z}}) &\leq \alpha(\eta + \nu) + \frac{1}{u_{j^\star}}\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln \mathop{\mathbb{E}}_{h\sim\mathbb{P}}\left[e^{u_{j^\star}\tilde{\phi}(h,\hat{\mathbf{Z}})}\right]\right) \\ &\leq \alpha(\eta + \nu) + \frac{1}{u_{j^\star}}\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{1}{\delta_{j^\star}} + \frac{u_{j^\star}^2\beta^2\left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}\right\|_\infty^2}{8mn}\right) \\ &\leq \alpha(\eta + \nu) + \frac{3\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta}\right)}{2u_{j^\star}} + \frac{u_{j^\star}\beta^2\left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}\right\|_\infty^2}{8mn} \\ &\leq \alpha(\eta + \nu) + 2\beta\left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}\right\|_\infty\sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta}}{2mn}}. \end{aligned}$$

The first inequality substitutes $u_{j^\star}$ into Equation 9; the second uses Equation 19; the third is from Equation 22; and the last uses the lower and upper bounds from Equation 21. ■

5.2 Posterior-Dependent Stability

In Theorem 1, we required β to be fixed *a priori*, meaning we required the user to pre-specify a desired stability. In this section, we prove bounds that hold for all $\beta \geq 1$ simultaneously, meaning the value of β can depend on the learned posterior. (The requirement of nonnegativity is not restrictive, since stability with $\beta \leq 1$ implies stability with $\beta = 1$.)

Theorem 2. *Fix $m \geq 1$, $n \geq 1$, $\boldsymbol{\pi} \in \Pi(n)$, $\delta \in (0,1)$ and $\alpha \geq 0$. Fix a distribution, $\mathbb{D}$, on $\mathcal{Z}^n$. Fix a set of bad inputs, $\mathcal{B}_\mathcal{Z}$, with $\nu \triangleq \mathbb{D}(\mathcal{B}_\mathcal{Z})$. Let $\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}$ denote the dependency matrix induced by $\mathbb{D}$, $\boldsymbol{\pi}$ and $\mathcal{B}_\mathcal{Z}$. Fix a prior, $\mathbb{P}$, on a hypothesis class, $\mathcal{H}$. Fix a loss function, L, such that $L\circ\mathcal{H}$ is α-uniformly range-bounded. Then, with probability at least $1-\delta-m\nu$ over realizations of $\hat{\mathbf{Z}} \triangleq (\mathbf{Z}^{(l)})_{l=1}^m$, drawn according to $\mathbb{D}^m$, the following hold: 1) for all $l \in [m]$, $\mathbf{Z}^{(l)} \notin \mathcal{B}_\mathcal{Z}$; 2) for all $\beta \geq 1$, $\eta \in [0,1]$ and posteriors $\mathbb{Q}$ with $(\beta/n, \mathcal{B}_\mathcal{Z}, \eta)$-local stability,*

$$\overline{L}(\mathbb{Q}) \leq \hat{L}(\mathbb{Q}, \hat{\mathbf{Z}}) + \alpha(\eta + \nu) + 4\beta\left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}\right\|_\infty\sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{4}{\delta} + \ln\beta}{2mn}}. \tag{23}$$

The proof is similar to that of Theorem 1, so we defer it to Appendix B.1.

Theorem 2 immediately yields the following corollary by taking $\mathcal{B}_{\mathcal{Z}} \triangleq \emptyset$.

Corollary 2. *With probability at least $1-\delta$ over realizations of $\hat{\mathbf{Z}}$, for all $\beta \geq 1$, $\eta \in [0,1]$ and $\mathbb{Q}$ with $(\beta/n, \emptyset, \eta)$-local stability,*

$$\overline{L}(\mathbb{Q}) \leq \hat{L}(\mathbb{Q}, \hat{\mathbf{Z}}) + \alpha\eta + 4\beta \left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}\right\|_{\infty} \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{4}{\delta} + \ln\beta}{2mn}}. \tag{24}$$

In Section 6, we apply this corollary to unbounded hypothesis classes, with bounded instance spaces. Corollary 2 trivially implies a bound for posteriors with $(\beta/n, \emptyset, 0)$-local stability, such as those with bounded support on an unbounded hypothesis class, where β may depend on a learned model.

5.3 Derandomizing the Loss using Stability

PAC-Bayes bounds are stated in terms of a randomized predictor. Yet, in practice, one is usually interested in the loss of a learned, deterministic predictor. Given a properly constructed posterior distribution, it is possible to convert a PAC-Bayes bound to a generalization bound for the learned hypothesis. There are various ways to go about this for unstructured hypotheses; however, many of these methods fail for structured predictors, since the output is not simply a scalar, but a high-dimensional vector. In this section, we present a generic technique for derandomizing PAC-Bayes bounds for structured prediction based on the idea of stability. An attractive feature of this technique is that it obviates margin-based arguments, which often require a free-parameter for the margin.

We first define a specialized notion of local stability that measures the difference in loss induced by perturbing a given hypothesis. For the following, we view the posterior $\mathbb{Q}$ as a function that, given a hypothesis $h \in \mathcal{H}$, returns a distribution $\mathbb{Q}_h$ on $\mathcal{H}$.

Definition 8. Fix a hypothesis class, $\mathcal{H}$, a set of inputs, $\mathcal{B}_{\mathcal{Z}} \subseteq \mathcal{Z}^n$, a loss function, L, and a posterior, $\mathbb{Q}$. We say that the pair $(L, \mathbb{Q})$ has $(\lambda, \mathcal{B}_{\mathcal{Z}}, \eta)$-local stability if, for any $h \in \mathcal{H}$ and $\mathbf{z} \notin \mathcal{B}_{\mathcal{Z}}$, there exists a set $\mathcal{B}_{\mathcal{H}}(h, \mathbf{z}) \subseteq \mathcal{H}$ such that $\mathbb{Q}_h(\mathcal{B}_{\mathcal{H}}(h, \mathbf{z})) \leq \eta$ and, for all $h' \notin \mathcal{B}_{\mathcal{H}}(h, \mathbf{z})$,

$$\left|L(h, \mathbf{z}) - L(h', \mathbf{z})\right| \leq \lambda. \tag{25}$$

This form of stability is a slightly weaker condition than the previous definitions, in that each input, $(h, \mathbf{z})$, has its own "bad" set, $\mathcal{B}_{\mathcal{H}}(h, \mathbf{z})$. This distinction means that "badness" is relative, whereas, in Definitions 2 and 3, it is absolute.

Proposition 4. *Fix a hypothesis class, $\mathcal{H}$, a set of inputs, $\mathcal{B}_{\mathcal{Z}} \subseteq \mathcal{Z}^n$, with $\nu \triangleq \mathbb{D}(\mathcal{B}_{\mathcal{Z}})$, and a loss function, L, such that, for any $\mathbf{z} \in \mathcal{Z}^n$, $L(\cdot, \mathbf{z})$ is α-uniformly range-bounded. Let $\mathbb{Q}$ denote a posterior function on $\mathcal{H}$. If $(L, \mathbb{Q})$ has $(\lambda, \mathcal{B}_{\mathcal{Z}}, \eta)$-local stability, then, for all $h \in \mathcal{H}$,*

$$\left|\overline{L}(h) - \overline{L}(\mathbb{Q}_h)\right| \leq \alpha(\eta + \nu) + \lambda, \tag{26}$$

and, for all $\hat{\mathbf{z}} \triangleq (\mathbf{z}^{(l)})_{l=1}^m$ such that, $\forall l \in [m]$, $\mathbf{z}^{(l)} \notin \mathcal{B}_{\mathcal{Z}}$,

$$\left|\hat{L}(h, \hat{\mathbf{z}}) - \hat{L}(\mathbb{Q}_h, \hat{\mathbf{z}})\right| \leq \alpha\eta + \lambda. \tag{27}$$

Proof Define a convenience function

$$\varphi(h, h', \mathbf{z}) \triangleq \left|L(h, \mathbf{z}) - L(h', \mathbf{z})\right|.$$

For any $\mathbf{z} \notin \mathcal{B}_{\mathcal{Z}}$, using the range-boundedness and stability assumptions, we have that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{h' \sim \mathbb{Q}_h}\left[\varphi(h, h', \mathbf{z})\right] \\
&\quad = \mathop{\mathbb{E}}_{h' \sim \mathbb{Q}_h}\left[\varphi(h, h', \mathbf{z})\,\mathbb{1}\{h' \in \mathcal{B}_{\mathcal{H}}(h, \mathbf{z})\}\right] + \mathop{\mathbb{E}}_{h' \sim \mathbb{Q}_h}\left[\varphi(h, h', \mathbf{z})\,\mathbb{1}\{h' \notin \mathcal{B}_{\mathcal{H}}(h, \mathbf{z})\}\right] \\
&\quad \leq \alpha\eta + \lambda.
\end{aligned}$$

Therefore, if, $\forall l \in [m]$, $\mathbf{z}^{(l)} \notin \mathcal{B}_{\mathcal{Z}}$, by linearity of expectation and the triangle inequality,

$$\begin{aligned}
\left|\hat{L}(h, \hat{\mathbf{z}}) - \hat{L}(\mathbb{Q}_h, \hat{\mathbf{z}})\right| &= \left|\frac{1}{m}\sum_{l=1}^{m} \mathop{\mathbb{E}}_{h' \sim \mathbb{Q}_h}\left[L\big(h, \mathbf{z}^{(l)}\big) - L\big(h', \mathbf{z}^{(l)}\big)\right]\right| \\
&\leq \frac{1}{m}\sum_{l=1}^{m} \mathop{\mathbb{E}}_{h' \sim \mathbb{Q}_h}\left[\varphi\big(h, h', \mathbf{z}^{(l)}\big)\right] \\
&\leq \alpha\eta + \lambda.
\end{aligned}$$

thus proving Equation 27. Furthermore,

$$\begin{aligned}
\left|\overline{L}(h) - \overline{L}(\mathbb{Q}_h)\right| &= \left|\mathop{\mathbb{E}}_{\mathbf{Z} \sim \mathbb{D}}\mathop{\mathbb{E}}_{h' \sim \mathbb{Q}_h}\left[L(h, \mathbf{Z}) - L(h', \mathbf{Z})\right]\right| \\
&\leq \mathop{\mathbb{E}}_{\mathbf{Z} \sim \mathbb{D}}\mathop{\mathbb{E}}_{h' \sim \mathbb{Q}_h}\left[\varphi(h, h', \mathbf{Z})\right] \\
&= \mathop{\mathbb{E}}_{\mathbf{Z} \sim \mathbb{D}}\mathop{\mathbb{E}}_{h' \sim \mathbb{Q}_h}\left[\varphi(h, h', \mathbf{Z})\,\mathbb{1}\{\mathbf{Z} \in \mathcal{B}_{\mathcal{Z}}\}\right] + \mathop{\mathbb{E}}_{\mathbf{Z} \sim \mathbb{D}}\mathop{\mathbb{E}}_{h' \sim \mathbb{Q}_h}\left[\varphi(h, h', \mathbf{Z})\,\mathbb{1}\{\mathbf{Z} \notin \mathcal{B}_{\mathcal{Z}}\}\right] \\
&\leq \alpha\nu + \alpha\eta + \lambda,
\end{aligned}$$

which proves Equation 26. ■

Proposition 4 can easily be combined with the PAC-Bayes bounds from the previous sections to obtain derandomized generalization bounds. We analyze some examples in Section 6.

5.3.1 Normed Vector Spaces

When the hypothesis class is a normed vector space (as is the case in all of the examples in Section 6), Definition 8 can be decomposed into properties of the loss function and posterior separately.

Definition 9. Fix a hypothesis class, $\mathcal{H}$, equipped with a norm, $\|\cdot\|$. Fix a set of inputs, $\mathcal{B}_{\mathcal{Z}} \subseteq \mathcal{Z}^n$. We say that a loss function, L, has $(\lambda, \mathcal{B}_{\mathcal{Z}})$-*local hypothesis stability* if, for all $h, h' \in \mathcal{H}$ and $\mathbf{z} \notin \mathcal{B}_{\mathcal{Z}}$,

$$\left|L(h, \mathbf{z}) - L(h', \mathbf{z})\right| \leq \lambda\left\|h - h'\right\|.$$

Definition 10. Fix a hypothesis class, $\mathcal{H}$, equipped with a norm, $\|\cdot\|$. We say that a posterior, $\mathbb{Q}$, has (β, η)-*local hypothesis stability* if, for any $h \in \mathcal{H}$, there exists a set $\mathcal{B}_{\mathcal{H}}(h) \subseteq \mathcal{H}$ such that $\mathbb{Q}_h(\mathcal{B}_{\mathcal{H}}(h)) \leq \eta$ and, for all $h' \notin \mathcal{B}_{\mathcal{H}}(h)$, $\|h - h'\| \leq \beta$.

When both of these properties hold, we have the following.

Proposition 5. *Fix a hypothesis class, $\mathcal{H}$, equipped with a norm, $\|\cdot\|$. Fix a set of inputs, $\mathcal{B}_{\mathcal{Z}} \subseteq \mathcal{Z}^n$. If a loss function, L, has $(\lambda, \mathcal{B}_{\mathcal{Z}})$-local hypothesis stability, and a posterior, $\mathbb{Q}$, has (β, η)-local hypothesis stability, then $(L, \mathbb{Q})$ has $(\lambda\beta, \mathcal{B}_{\mathcal{Z}}, \eta)$-local stability.*

The proof is provided in Appendix B.2.

6. Example Applications

To illustrate how various learning algorithms and modeling decisions affect the generalization error, we now apply our PAC-Bayes bounds to the class of pairwise MRFs with templated, linear potentials (described in Section 2.2.3). We derive generalization bounds for two popular training regimes, *max-margin* and *soft-max* learning, under various assumptions about the instance space and feature functions. The bounds in this section are stated in terms of a deterministic predictor, meaning we use the PAC-Bayes framework as an analytic tool only. That said, one could easily adapt our analysis to obtain bounds for a randomized predictor by skipping the derandomization step.

6.1 Max-Margin Learning

For classification tasks, the goal is to output the labeling that is closest to the true labeling, by some measure of closeness. This is usually measured by the *Hamming loss*,

$$L_{\text{H}}(h, \mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n} D_{\text{H}}\left(\mathbf{y}, h(\mathbf{x})\right).$$

The Hamming loss can be considered the structured equivalent of the *0-1 loss*. Unfortunately, the Hamming loss is not convex, making it difficult to minimize directly. Thus, many learning algorithms minimize a convex upper bound.

One such method is *max-margin* learning. Max-margin learning aims to find the "simplest" model that scores the correct outputs higher than all incorrect outputs by a specified margin. Though typically formulated as a quadratic program, the learning objective can also be stated as minimizing a *hinge loss*, with model regularization.

Structured predictors learned with a max-margin objective are alternatively referred to as *max-margin Markov networks* (Taskar et al., 2004) or *StructSVM* (Tsochantaridis et al., 2005), depending on the form of the hinge loss. In this section, we consider the former formulation, defining the structured hinge loss as

$$L_{\text{h}}(h, \mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n}\left(\max_{\mathbf{y}' \in \mathcal{Y}^n} D_{\text{H}}(\mathbf{y}, \mathbf{y}') + h(\mathbf{x}, \mathbf{y}') - h(\mathbf{x}, \mathbf{y})\right), \tag{28}$$

where

$$h(\mathbf{x}, \mathbf{y}) \triangleq \boldsymbol{\theta}(\mathbf{x}; \mathbf{w}) \cdot \hat{\mathbf{y}} \tag{29}$$

is the unnormalized log-likelihood. The Hamming distance, $D_{\mathrm{H}}(\mathbf{y}, \mathbf{y}')$, implies that the margin, $h(\mathbf{x}, \mathbf{y}) - h(\mathbf{x}, \mathbf{y}')$, should scale linearly with the distance between $\mathbf{y}$ and $\mathbf{y}'$.

In theory, the structured hinge loss can be defined with any distance function; though, in practice, the Hamming distance is commonly used. One attractive property of the Hamming distance is that, when

$$h(\mathbf{x}) \triangleq \operatorname*{arg\,max}_{\mathbf{y} \in \mathcal{Y}^n} h(\mathbf{x}, \mathbf{y}) = \operatorname*{arg\,max}_{\mathbf{y} \in \mathcal{Y}^n} p\left(\mathbf{Y} = \mathbf{y} \,|\, \mathbf{X} = \mathbf{x}; \mathbf{w}\right) \tag{30}$$

(i.e., MAP inference), the hinge loss upper-bounds the Hamming loss. Another benefit is that it decomposes along the unary cliques. Indeed, with $\delta(\mathbf{y}) \triangleq \left[\begin{smallmatrix} \mathbf{1} - \mathbf{y} \\ \mathbf{0} \end{smallmatrix}\right]$ (i.e., one minus the unary clique states, then zero-padded to be the same length as $\hat{\mathbf{y}}$), observe that $D_{\mathrm{H}}(\mathbf{y}, \mathbf{y}') = \delta(\mathbf{y}) \cdot \hat{\mathbf{y}}'$. This identity yields a convenient equivalence:

$$L_{\mathrm{h}}(h, \mathbf{x}, \mathbf{y}) = \frac{1}{n} \left(\max_{\mathbf{y}' \in \mathcal{Y}^n} \left(\boldsymbol{\theta}(\mathbf{x}; \mathbf{w}) + \delta(\mathbf{y})\right) \cdot \hat{\mathbf{y}}' - \boldsymbol{\theta}(\mathbf{x}; \mathbf{w}) \cdot \hat{\mathbf{y}} \right).$$

The term $\boldsymbol{\theta}(\mathbf{x}; \mathbf{w}) \cdot \hat{\mathbf{y}}$ is constant with respect to $\mathbf{y}'$, and is thus irrelevant to the maximization. Therefore, letting

$$\tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}) \triangleq \boldsymbol{\theta}(\mathbf{x}; \mathbf{w}) + \delta(\mathbf{y}), \tag{31}$$

computing the hinge loss is equivalent to performing *loss-augmented* MAP inference with $\tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w})$. Provided inference can be computed efficiently with the given class of models, so too can the hinge loss.[6]

6.1.1 Structured Ramp Loss

Applying our generalization bounds requires a uniformly range-bounded loss function. Since the hinge loss is not uniformly range-bounded for certain hypothesis classes, we therefore introduce the structured *ramp loss*:

$$L_{\mathrm{r}}(h, \mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n} \left(\max_{\mathbf{y}' \in \mathcal{Y}^n} D_{\mathrm{H}}(\mathbf{y}, \mathbf{y}') + h(\mathbf{x}, \mathbf{y}') - \max_{\mathbf{y}'' \in \mathcal{Y}^n} h(\mathbf{x}, \mathbf{y}'') \right),$$

where $h(\mathbf{x}, \mathbf{y})$ is defined in Equation 29. The ramp loss is 1-uniformly range-bounded. Further, when $h(\mathbf{x})$ performs MAP inference (Equation 30),

$$L_{\mathrm{H}}(h, \mathbf{x}, \mathbf{y}) \leq L_{\mathrm{r}}(h, \mathbf{x}, \mathbf{y}) \leq L_{\mathrm{h}}(h, \mathbf{x}, \mathbf{y}). \tag{32}$$

Thus, we can analyze the generalization properties of the ramp loss to obtain bounds for the difference of the expected Hamming loss and empirical hinge loss. To distinguish quantities of different loss functions, we will use a subscript notation; e.g., $\overline{L}_{\mathrm{H}}$ is the expected Hamming loss, and $\hat{L}_{\mathrm{h}}$ is the empirical hinge loss.

Using the templated, linear potentials defined in Section 2.2.3, we obtain two technical lemmas for the structured ramp loss. Proofs are provided in Appendices C.1 and C.2.

6. The results in this section are easily extended to approximate MAP inference algorithms, such as linear programming relaxations. The bounds are the same, but the semantics of the loss functions change, since approximate MAP solutions might be fractional.

Lemma 2. *Fix any $p, q \geq 1$ such that $1/p + 1/q = 1$. Fix a graph, $G \triangleq (\mathcal{V}, \mathcal{E})$, with maximum degree Δ_G. Assume that $\sup_{x \in \mathcal{X}} \|x\|_p \leq R$. Then, for any MRF h with weights $\mathbf{w}$, and any $\mathbf{z}, \mathbf{z} \in \mathcal{Z}^n$, where $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ and $\mathbf{z}' = (\mathbf{x}', \mathbf{y}')$,*

$$\left|L_r(h, \mathbf{z}) - L_r(h, \mathbf{z}')\right| \leq \frac{1}{n}\left((2\Delta_G + 4)R\|\mathbf{w}\|_q + 1\right) D_{\mathrm{H}}(\mathbf{z}, \mathbf{z}'). \tag{33}$$

Further, if the model does not use edge observations (i.e., $f_{ij}(\mathbf{x}, \mathbf{y}) \triangleq y_i \otimes y_j$), then

$$\left|L_r(h, \mathbf{z}) - L_r(h, \mathbf{z}')\right| \leq \frac{1}{n}\left(4R\|\mathbf{w}\|_q + 1\right) D_{\mathrm{H}}(\mathbf{z}, \mathbf{z}'). \tag{34}$$

Lemma 3. *Fix any $p, q \geq 1$ such that $1/p + 1/q = 1$. Fix a graph, $G \triangleq (\mathcal{V}, \mathcal{E})$. Assume that $\sup_{x \in \mathcal{X}} \|x\|_p \leq R$. Then, for any example $\mathbf{z} \in \mathcal{Z}^n$, and any two MRFs, h, h' with weights $\mathbf{w}, \mathbf{w}'$,*

$$\left|L_r(h, \mathbf{z}) - L_r(h', \mathbf{z})\right| \leq \frac{2|G|R}{n}\left\|\mathbf{w} - \mathbf{w}'\right\|_q.$$

Lemma 3 implies that L_{r} has $(2|G|R/n, \emptyset)$-local hypothesis stability.

6.1.2 Generalization Bounds for Max-Margin Learning

We now apply our PAC-Bayes bounds to the class of max-margin Markov networks that perform MAP inference, with the templated, linear potentials defined in Section 2.2.3. We denote this class by $\mathcal{H}_{\mathrm{M3N}}$. As a warm-up, we first assume that both the observations and weights are uniformly bounded by the 2-norm unit ball. By Lemma 2, this means that the ramp loss satisfies uniform stability, meaning we can apply Corollary 1.

Example 1. Fix any $m \geq 1$, $n \geq 1$, $\boldsymbol{\pi} \in \Pi(n)$ and $\delta \in (0, 1)$. Fix a graph, $G \triangleq (\mathcal{V}, \mathcal{E})$, with maximum degree Δ_G. Assume that $\sup_{x \in \mathcal{X}} \|x\|_2 \leq 1$. Then, with probability at least $1 - \delta$ over realizations of $\hat{\mathbf{Z}} \triangleq (\mathbf{Z}^{(l)})_{l=1}^m$, for all $h \in \mathcal{H}_{\mathrm{M3N}}$ with $\|\mathbf{w}\|_2 \leq 1$,

$$\overline{L}_{\mathrm{H}}(h) \leq \hat{L}_{\mathrm{h}}(h, \hat{\mathbf{Z}}) + \frac{4}{mn} + (4\Delta_G + 10)\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}\|_\infty \sqrt{\frac{d\ln(2m|G|) + \ln\frac{2}{\delta}}{2mn}}.$$

The proof is given in Appendix C.3. Note that, with the bounded degree assumption, $|G| \leq n\Delta_G = \mathrm{O}(n)$.

We now relax the assumption that the hypothesis class is bounded. One approach is to apply a covering argument directly to Example 1. However, it is interesting to see how other prior/posterior constructions behave. Of particular interest are Gaussian constructions, which correspond to 2-norm regularization. Since the support of a Gaussian is unbounded, this construction requires a non-uniform notion of stability. The following example illustrates how to use posterior-dependent, local stability.

Example 2. Fix any $m \geq 1$, $n \geq 1$, $\boldsymbol{\pi} \in \Pi(n)$ and $\delta \in (0, 1)$. Fix a graph, $G \triangleq (\mathcal{V}, \mathcal{E})$, with maximum degree Δ_G. Assume that $\sup_{x \in \mathcal{X}} \|x\|_2 \leq 1$. Then, with probability at least $1 - \delta$ over realizations of $\hat{\mathbf{Z}} \triangleq (\mathbf{Z}^{(l)})_{l=1}^m$, for all $h \in \mathcal{H}_{\mathrm{M3N}}$,

$$\overline{L}_{\mathrm{H}}(h) \leq \hat{L}_{\mathrm{h}}(h, \hat{\mathbf{Z}}) + \frac{7}{mn} + 4\beta_h\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}\|_\infty \sqrt{\frac{\frac{1}{2}\|\mathbf{w}\|_2^2 + \frac{d}{2}\ln\left(2d(m|G|)^2\ln(2dmn)\right) + \ln\frac{4\beta_h}{\delta}}{2mn}},$$

where

$$\beta_h \triangleq (2\Delta_G + 4)\left(\|\mathbf{w}\|_2 + \frac{1}{m\,|G|}\right) + 1.$$

Example 2 is only slightly worse than Example 1, incurring a $\mathrm{O}\left(\ln\ln(mn)\right)$ term for the Gaussian construction. Both bounds guarantee generalization when either m or n is large.

The proof of Example 2 uses a concentration inequality for vectors of Gaussian random variables, the proof of which is given Appendix C.4.

Lemma 4. *Let $\mathbf{X} \triangleq (X_i)_{i=1}^d$ be independent Gaussian random variables, with mean vector $\boldsymbol{\mu} \triangleq (\mu_1, \ldots, \mu_d)$ and variance σ^2. Then, for any $p \geq 1$ and $\epsilon > 0$,*

$$\Pr\left\{\|\mathbf{X} - \boldsymbol{\mu}\|_p \geq \epsilon\right\} \leq 2d \exp\left(-\frac{\epsilon^2}{2\sigma^2 d^{2/p}}\right).$$

For $p = 2$ and small σ^2, this bound can be significantly sharper than Chebyshev's inequality.

Proof (Example 2) Define the prior, $\mathbb{P}$, as a standard multivariate Gaussian, with zero mean and unit variance. More precisely, let

$$p(h) \triangleq (2\pi)^{-d/2} e^{-\frac{1}{2}\|\mathbf{w}\|_2^2}$$

denote the density of $\mathbb{P}$. Given a (learned) hypothesis, h, we construct the posterior, $\mathbb{Q}_h$, as an isotropic Gaussian, centered at $\mathbf{w}$, with variance

$$\sigma^2 \triangleq \left(2d(m\,|G|)^2 \ln(2dmn)\right)^{-1}$$

in all dimensions. Its density is

$$q_h(h') \triangleq (2\pi\sigma^2)^{-d/2} e^{-\frac{1}{2\sigma^2}\|\mathbf{w}' - \mathbf{w}\|_2^2}.$$

Note that the support of both distributions is $\mathbb{R}^d$, which is unbounded.

Our proof technique involves four steps. First, we upper-bound the KL divergence between $\mathbb{Q}_h$ and $\mathbb{P}$. Then, we identify a β_h and η such that $\mathbb{Q}_h$ is $(\beta_h/n, \emptyset, \eta)$-locally stable. Combining the first two steps with Corollary 2 yields a PAC-Bayes bound for the randomized predictor. The final step is to derandomize this bound using Proposition 4.

The KL divergence between Gaussians is well known. Thus, it is easily verified that

$$\begin{aligned}
D_{\mathrm{KL}}(\mathbb{Q}_h \| \mathbb{P}) &= \frac{1}{2}\left[d\left(\sigma^2 - 1\right) + \|\mathbf{w}\|_2^2 - d\ln\sigma^2\right] \\
&= \frac{1}{2}\left[d\left(\frac{1}{2d(m\,|G|)^2\ln(2dmn)} - 1\right) + \|\mathbf{w}\|_2^2 + d\ln\left(2d(m\,|G|)^2\ln(2dmn)\right)\right] \\
&\leq \frac{1}{2}\left[\|\mathbf{w}\|_2^2 + d\ln\left(2d(m\,|G|)^2\ln(2dmn)\right)\right].
\end{aligned}$$

The inequality follows from the fact that $\sigma^2 \leq 1$ for all $d \geq 1$, $m \geq 1$ and $n \geq 1$ (implying $|G| \geq 1$).

To determine the local stability of $\mathbb{Q}_h$, for any $h \in \mathcal{H}_{\mathrm{M3N}}$, we define a "bad" set of hypotheses,

$$\mathcal{B}_{\mathcal{H}_{\mathrm{M3N}}}(h) \triangleq \left\{h' \in \mathcal{H}_{\mathrm{M3N}} : \left\|\mathbf{w}' - \mathbf{w}\right\|_2 \geq \frac{1}{m\,|G|}\right\}.$$

Using Lemma 4,

$$\begin{aligned}\mathbb{Q}_h\left(\mathcal{B}_{\mathcal{H}_{\text{M3N}}}(h)\right) &= \Pr_{h'\sim\mathbb{Q}_h}\left\{\left\|\mathbf{w}'-\mathbf{w}\right\|_2 \geq \frac{1}{m\left|G\right|}\right\} \\ &\leq 2d\exp\left(-\frac{2d(m\left|G\right|)^2\ln(2dmn)}{2d(m\left|G\right|)^2}\right) \\ &= \frac{1}{mn}. \end{aligned} \tag{35}$$

Further, for every $h' \notin \mathcal{B}_{\mathcal{H}_{\text{M3N}}}(h)$,

$$\left\|\mathbf{w}'\right\|_2 - \left\|\mathbf{w}\right\|_2 \leq \left\|\mathbf{w}'-\mathbf{w}\right\|_2 \leq \frac{1}{m\left|G\right|}.$$

When combined with Lemma 2, with $R=1$, we have that

$$\begin{aligned}\left|L_{\text{r}}(h,\mathbf{z}) - L_{\text{r}}(h,\mathbf{z}')\right| &\leq \frac{1}{n}\left(\left(2\Delta_G+4\right)\left\|\mathbf{w}'\right\|_2+1\right)D_{\text{H}}(\mathbf{z},\mathbf{z}') \\ &\leq \frac{1}{n}\left(\left(2\Delta_G+4\right)\left(\left\|\mathbf{w}\right\|_2+\frac{1}{m\left|G\right|}\right)+1\right)D_{\text{H}}(\mathbf{z},\mathbf{z}') \\ &= \frac{\beta_h}{n}D_{\text{H}}(\mathbf{z},\mathbf{z}').\end{aligned}$$

Thus, every $\mathbb{Q}_h$ is $(\beta_h/n, \emptyset, 1/(mn))$-locally stable.

Having established an upper bound on the KL divergence and local stability of all posteriors, we can now apply one of our PAC-Bayes bounds. Since the definition of β_h depends on the posterior via $\mathbf{w}$, we must use a bound from Section 5.2. In this case, there are no "bad" inputs, since the observations are bounded in the unit ball, so we can invoke Corollary 2. Recalling that the ramp loss is 1-uniformly difference bounded, we then have that, with probability at least $1-\delta$, every $\mathbb{Q}_h : h \in \mathcal{H}_{\text{M3N}}$ satisfies

$$\begin{aligned}\overline{L}_{\text{r}}(\mathbb{Q}_h) \leq \hat{L}_{\text{r}}(\mathbb{Q}_h,\hat{\mathbf{Z}}) + \frac{1}{mn} \\ + 4\beta_h\left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}\right\|_\infty\sqrt{\frac{\frac{1}{2}\left\|\mathbf{w}\right\|_2^2+\frac{d}{2}\ln\left(2d(m\left|G\right|)^2\ln(2dmn)\right)+\ln\frac{4\beta_h}{\delta}}{2mn}}.\end{aligned} \tag{36}$$

We now derandomize the loss terms in Equation 36. Observe that $\mathcal{H}_{\text{M3N}}$ is a normed vector space, since it consists of weight vectors in $\mathbb{R}^d$. In this case, we will use the 2-norm. By Equation 35, it is clear that $\mathbb{Q}$ has $(1/(m\left|G\right|), 1/(mn))$-local hypothesis stability (Definition 10), since every $h \in \mathcal{H}_{\text{M3N}}$ results in the same probability bound. Further, by Lemma 3, with $R=1$,

$$\left|L_{\text{r}}(h,\mathbf{z}) - L_{\text{r}}(h',\mathbf{z})\right| \leq \frac{2\left|G\right|}{n}\left\|\mathbf{w}-\mathbf{w}'\right\|_2, \tag{37}$$

meaning L_{r} has $(2\left|G\right|/n, \emptyset)$-local hypothesis stability (Definition 9). Therefore, by Proposition 5, $(L_{\text{r}}, \mathbb{Q})$ has $(2/(mn), \emptyset, 1/(mn))$-local stability. It then follows, via Proposition 4 and Equation 32, that

$$\overline{L}_{\text{H}}(h) \leq \overline{L}_{\text{r}}(h) \leq \overline{L}_{\text{r}}(\mathbb{Q}_h) + \frac{3}{mn}, \tag{38}$$

and

$$\hat{L}_{\mathrm{r}}(\mathbb{Q}_h, \hat{\mathbf{Z}}) \leq \hat{L}_{\mathrm{r}}(h, \hat{\mathbf{Z}}) + \frac{3}{mn} \leq \hat{L}_{\mathrm{h}}(h, \hat{\mathbf{Z}}) + \frac{3}{mn}. \tag{39}$$

Combining Equations 36, 38 and 39 completes the proof. ■

6.2 Soft-Max Learning

A drawback of max-margin learning is that the learning objective is not differentiable everywhere, due to the hinge loss. Thus, researchers (Gimpel and Smith, 2010; Hazan and Urtasun, 2010) have proposed a smooth alternative, based on the *soft-max* function. This form of learning has been popularized for learning conditional random fields (CRFs).

The soft-max loss, for a given temperature parameter, $\epsilon \in [0, 1]$, is defined as

$$L_{\mathrm{sm}}(h, \mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n} \left(\Phi_\epsilon(\mathbf{x}, \mathbf{y}; \mathbf{w}) - h(\mathbf{x}, \mathbf{y}) \right), \tag{40}$$

where $h(\mathbf{x}, \mathbf{y})$ is the unnormalized log-likelihood (Equation 29) and

$$\begin{aligned} \Phi_\epsilon(\mathbf{x}, \mathbf{y}; \mathbf{w}) &\triangleq \epsilon \ln \sum_{\mathbf{y}' \in \mathcal{Y}^n} \exp\left(\frac{1}{\epsilon} \left(D_{\mathrm{H}}(\mathbf{y}, \mathbf{y}') + h(\mathbf{x}, \mathbf{y}') \right) \right) \\ &= \epsilon \ln \sum_{\mathbf{y}' \in \mathcal{Y}^n} \exp\left(\frac{1}{\epsilon} \tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}) \cdot \hat{\mathbf{y}}' \right). \end{aligned} \tag{41}$$

is the soft-max function. We purposefully overload the notation of the log-partition function due to its relationship to the soft-max. Observe that, for $\epsilon = 1$, the soft-max becomes the log-partition of the distribution induced by the loss-augmented potentials, and Equation 40 is the corresponding negative log-likelihood, scaled by $1/n$. Further, as $\epsilon \to 0$, the soft-max approaches the max operator and Equation 40 becomes the hinge loss (Equation 28).

The latter equivalence can be illustrated by convex conjugacy. This requires some additional notation. Let $\boldsymbol{\mu} \in [0, 1]^{|\mathcal{V}||\mathcal{Y}| + |\mathcal{E}||\mathcal{Y}|^2}$ denote a vector of marginal probabilities for all cliques and clique states. Let $\mathcal{M}$ denote the set of all consistent marginal vectors, often called the *marginal polytope*. For every $\boldsymbol{\mu} \in \mathcal{M}$, there is a corresponding distribution, $p_{\boldsymbol{\mu}}$, such that $\boldsymbol{\mu}_c \cdot \mathbf{y}_c = p_{\boldsymbol{\mu}}(\mathbf{Y}_c = \mathbf{y}_c)$ for every clique, $c \in \mathcal{C}$, and clique state, $\mathbf{y}_c$. Let $\Phi^*(\boldsymbol{\mu})$ denote the *convex conjugate* of the log-partition, which, for $\boldsymbol{\mu} \in \mathcal{M}$, is equal to the negative entropy of $p_{\boldsymbol{\mu}}$.[7] With these definitions, the soft-max, like the log-partition, has the following variational form:

$$\begin{aligned} \Phi_\epsilon(\mathbf{x}, \mathbf{y}; \mathbf{w}) &= \max_{\boldsymbol{\mu} \in \mathcal{M}} \tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}) \cdot \boldsymbol{\mu} - \epsilon \, \Phi^*(\boldsymbol{\mu}) \\ &= \max_{\boldsymbol{\mu} \in \mathcal{M}} \left(\boldsymbol{\theta}(\mathbf{x}, \mathbf{y}; \mathbf{w}) + \delta(\mathbf{y}) \right) \cdot \boldsymbol{\mu} - \epsilon \, \Phi^*(\boldsymbol{\mu}). \end{aligned} \tag{42}$$

7. We omit some details of the conjugate function for simplicity of exposition. See Wainwright and Jordan (2008) for a precise definition.

This maximization is equivalent to marginal inference with loss-augmented potentials.[8] Let $\boldsymbol{\mu}_{\mathrm{u}}$ denote the marginals of the unary cliques, and observe that

$$\delta(\mathbf{y}) \cdot \boldsymbol{\mu} = \frac{1}{2} \left\| \mathbf{y} - \boldsymbol{\mu}_{\mathrm{u}} \right\|_1 \triangleq D_1(\mathbf{y}, \boldsymbol{\mu}). \tag{43}$$

With a slight abuse of notation, we define an alternate scoring function for marginals:

$$h_\epsilon(\mathbf{x}, \boldsymbol{\mu}) \triangleq \boldsymbol{\theta}(\mathbf{x}; \mathbf{w}) \cdot \boldsymbol{\mu} - \epsilon \, \Phi^*(\boldsymbol{\mu}). \tag{44}$$

Note that each full labeling, $\hat{\mathbf{y}}$, corresponds to a vertex of the marginal polytope, so $\hat{\mathbf{y}} \in \mathcal{M}$. Further, $h_\epsilon(\mathbf{x}, \hat{\mathbf{y}}) = h(\mathbf{x}, \mathbf{y})$, since $\Phi^*(\hat{\mathbf{y}}) = 0$. Thus, combining Equations 42 to 44, we have that the soft-max loss (Equation 40) is equivalent to

$$L_{\mathrm{sm}}(h, \mathbf{x}, \mathbf{y}) = \frac{1}{n} \left(\max_{\boldsymbol{\mu} \in \mathcal{M}} D_1(\mathbf{y}, \boldsymbol{\mu}) + h_\epsilon(\mathbf{x}, \boldsymbol{\mu}) - h_\epsilon(\mathbf{x}, \hat{\mathbf{y}}) \right),$$

which resembles a smoothed hinge loss for $\epsilon \in (0, 1)$.

Like the regular hinge loss, $L_{\mathrm{sm}}(h, \mathbf{x}, \mathbf{y})$ is not uniformly range-bounded for certain hypothesis classes, so it cannot be used with our PAC-Bayes bounds. However, we can use the ramp loss, with a slight modification:

$$L_{\mathrm{sr}}(h, \mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n} \left(\max_{\boldsymbol{\mu} \in \mathcal{M}} D_1(\mathbf{y}, \boldsymbol{\mu}) + h_\epsilon(\mathbf{x}, \boldsymbol{\mu}) - \max_{\boldsymbol{\mu}' \in \mathcal{M}} h_\epsilon(\mathbf{x}, \boldsymbol{\mu}') \right).$$

We have essentially just replaced the maxes over $\mathcal{Z}^n$ with maxes over $\mathcal{M}$ and used Equation 44 instead of Equation 29. We refer to this loss as the *soft ramp loss*. The stability properties of the regular ramp loss carry over to the soft ramp loss; it is straightforward to show that Lemmas 2 and 3 hold when $L_{\mathrm{r}}(h, \mathbf{x}, \mathbf{y})$ is replaced with $L_{\mathrm{sr}}(h, \mathbf{x}, \mathbf{y})$.[9]

The distance function, $D_1(\mathbf{y}, \boldsymbol{\mu})$, has a probabilistic interpretation:

$$D_1(\mathbf{y}, \boldsymbol{\mu}) = \sum_{i=1}^{n} 1 - p_{\boldsymbol{\mu}} \left(Y_i = y_i \,|\, \mathbf{X} = \mathbf{x} \right).$$

This identity motivates another loss function; with

$$h_\epsilon(\mathbf{x}) \triangleq \arg\max_{\boldsymbol{\mu} \in \mathcal{M}} h_\epsilon(\mathbf{x}, \boldsymbol{\mu}),$$

let

$$L_1(h, \mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n} D_1 \left(\mathbf{y}, h_\epsilon(\mathbf{x}) \right) = \frac{1}{n} \sum_{i=1}^{n} 1 - p \left(Y_i = y_i \,|\, \mathbf{X} = \mathbf{x}; \mathbf{w} \right).$$

Note that

$$L_1(h, \mathbf{x}, \mathbf{y}) \leq L_{\mathrm{sr}}(h, \mathbf{x}, \mathbf{y}) \leq L_{\mathrm{sm}}(h, \mathbf{x}, \mathbf{y}).$$

8. Since marginal inference is often intractable, exact inference could be replaced with a tractable surrogate, such as the Bethe approximation.
9. The additional $\epsilon \, \Phi^*(\cdot)$ term in Equation 44 is canceled out in Equations 55, 56, 58 and 59.

Marginal inference, $h_\epsilon(\mathbf{x})$, can be decoded by selecting the labels with the highest marginal probabilities. This technique is sometimes referred to as *posterior decoding*. Conveniently, because the marginals sum to one, it can be shown that the Hamming loss of the posterior decoding is at most twice L_1.

In the following example, we consider the class of soft-max CRFs, $\mathcal{H}_{\text{CRF}}$. For historical reasons, these models typically do not use edge observations, which is a common modeling decision in, e.g., sequence models. We therefore assume that the edge features are simply $f_{ij}(\mathbf{x}, \mathbf{y}) \triangleq y_i \otimes y_j$.

Example 3. Fix any $m \geq 1$, $n \geq 1$, $\boldsymbol{\pi} \in \Pi(n)$, $\delta \in (0, 1)$ and $G \triangleq (\mathcal{V}, \mathcal{E})$. Assume that $\sup_{x \in \mathcal{X}} \|x\|_2 \leq 1$. Then, with probability at least $1 - \delta$ over realizations of $\hat{\mathbf{Z}} \triangleq (\mathbf{Z}^{(l)})_{l=1}^m$, for all $h \in \mathcal{H}_{\text{CRF}}$,

$$\overline{L}_1(h) \leq \hat{L}_{\text{sm}}(h, \hat{\mathbf{Z}}) + \frac{7}{mn} + 4\beta_h \left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}\right\|_\infty \sqrt{\frac{\frac{1}{2}\|\mathbf{w}\|_2^2 + \frac{d}{2}\ln\left(2d(m\,|G|)^2 \ln(2dmn)\right) + \ln\frac{4\beta_h}{\delta}}{2mn}},$$

where

$$\beta_h \triangleq 4\left(\|\mathbf{w}\|_2 + \frac{1}{m\,|G|}\right) + 1.$$

We omit the proof, since it is almost identical to Example 2. The key difference worth noting is that, since the model does not use edge observations, the graph's maximum degree does not appear in β_h.

6.3 Possibly Unbounded Domains

Until now, we have assumed that the observations are uniformly bounded in the unit ball. This assumption is common in the literature, but it does not quite match what happens in practice. Typically, one will rescale each dimension of the input space using the minimum and maximum values found in the training data. While this procedure guarantees a bound on the observations at training time, the bound may not hold at test time when one rescales by the limits estimated from the training set. This outcome would violate the preconditions of the stability guarantees used to prove the previous examples.

Now, suppose we knew that the observations were bounded with high probability. In the following example, we construct a hypothetical data distribution under which this assumption holds. We combine this with Theorem 2 to derive a variant of Example 2.

Example 4. Fix any $m \geq 1$, $n \geq 1$, $\boldsymbol{\pi} \in \Pi(n)$, $\delta \in (0, 1)$ and $G \triangleq (\mathcal{V}, \mathcal{E})$. Suppose the data generating process, $\mathbb{D}$, is defined as follows. For each $y \in \mathcal{Y}$, assume there is an associated isotropic Gaussian over $\mathcal{X} \subseteq \mathbb{R}^k$, with mean $\mu_y \in \mathcal{X} : \|\mu_y\|_2 \leq 1$ and variance $\sigma_y^2 \leq \left(2k\ln(2kn^2)\right)^{-1}$. First, $\mathbf{Y}$ is sampled according to some arbitrary distribution, conditioned on G. Then, for each $i \in [n]$, conditioned on $Y_i = y_i$, a vector of observations, $x_i \in \mathcal{X}$, is sampled according to $(\mu_{y_i}, \sigma_{y_i}^2)$.

Note that, conditioned on the labels, $(y_1, \ldots, y_n)$, the observations, $(x_1, \ldots, x_n)$, are mutually independent. It therefore does not make sense to model edge observations, so we use $f_{ij}(\mathbf{x}, \mathbf{y}) \triangleq y_i \otimes y_j$. For the following, we abuse our previous notation and let $\mathcal{H}_{\text{M3N}}$ denote the class of max-margin Markov networks that use these edge features.

Let $\mathcal{B}_{\mathcal{Z}} \triangleq \{\exists i : \|X_i\|_2 \geq 2\}$ denote a set of "bad" inputs, and let $\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\mathbb{B}}$ denote the dependency matrix induced by $\mathbb{D}$, $\boldsymbol{\pi}$ and $\mathcal{B}_{\mathcal{Z}}$. Then, with probability at least $1 - \delta - m/n$ over realizations of $\hat{\mathbf{Z}} \triangleq (\mathbf{Z}^{(l)})_{l=1}^{m}$, for all $h \in \mathcal{H}_{\text{M3N}}$,

$$\overline{L}_{\text{H}}(h) \leq \hat{L}_{\text{h}}(h, \hat{\mathbf{Z}}) + \frac{11}{mn} + \frac{2}{n} + 4\beta_h \left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\mathbb{B}}\right\|_{\infty} \sqrt{\frac{\frac{1}{2}\|\mathbf{w}\|_2^2 + \frac{d}{2}\ln\left(2d(m\,|G|)^2 \ln(2dmn)\right) + \ln\frac{4\beta_h}{\delta}}{2mn}},$$

where

$$\beta_h \triangleq 8\left(\|\mathbf{w}\|_2 + \frac{1}{m\,|G|}\right) + 1.$$

The proof is provided in Appendix C.5.

Note that the dominating term is $2/n$, meaning the bound is meaningful for large n and small m. This rate follows intuition, since one should not expect η to depend on the number of training examples; moreover, the probability of drawing a "bad" example should increase proportionally to the number of independent draws.

7. Discussion

We have proposed new PAC-Bayes bounds for structured prediction that can decrease with both the number of examples, m, and the size of each example, n, thus proving that generalization is indeed possible from a few large examples. Under suitable conditions, our bounds can be as tight as $\tilde{\text{O}}\left(1/\sqrt{mn}\right)$. The bounds reveal the connection between generalization and the stability of a structured loss function, as well as the role of dependence in the generating distribution. The stability conditions used in this work generalize our previous work, thereby accommodating a broader range of structured loss functions, including max-margin and soft-max learning. We also provide bounds on the norm of the dependency matrix, which is a result that may be useful outside of this context.

The examples in Section 6 identify several take-aways for practitioners. Primarily, they indicate the importance of templating (or, parameter-tying). Observe that all of the bounds depend on d, the number of parameters[10], via a term that is $\tilde{\text{O}}(d/n)$. Clearly, if d scales linearly with n, the number of nodes, then this term is bounded away from zero as $n \to \infty$. Consequently, one cannot hope to generalize from one example. Though we do not prove this formally, the intuition is fairly simple: if there is a different $\mathbf{w}_i$ for each node i, and $\mathbf{w}_{ij}$ for each edge $\{i, j\}$, then one example provides exactly one "micro example" from which one can estimate $\{\mathbf{w}_i\}_{i\in\mathcal{V}}$ and $\{\mathbf{w}_{ij}\}_{\{i,j\}\in\mathcal{E}}$. In this setting, our bounds become $\tilde{\text{O}}(1/\sqrt{m})$, which is no better (and no worse) than previous bounds. Thus, templating is crucial to achieving the fast generalization rate.[11]

Another observation is that Examples 2 to 4 depend on the norm of the weight vector, $\mathbf{w}$. Specifically, we used the 2-norm, for its relationship to Gaussian priors; though, one could substitute any norm, due to the equivalence of norms in finite dimension. Dependence

10. We believe that this dependence is unavoidable when derandomizing PAC-Bayes bounds for structured prediction. Evidence to support this conjecture is given by McAllester's (2007) bound, which depends on the number templates, and the number of parameters is roughly linear in the number of templates.
11. It may be possible to achieve a fast rate without templating if one imposes a sparsity assumption on the optimal weight vector, but it seems likely that the sparsity would depend on n.

on the norm of the weights is a standard feature of most generalization bounds. This term is commonly interpreted as a measure of hypothesis complexity. Weight regularization during training controls the norm of the weights, thereby effectively limiting the complexity of the learned model.

We also find that the structure of the the model influences the bounds via Δ_G, the maximum degree of the graph, and $|G|$, the total number of nodes and edges. (Since the bounds are sub-logarithmic in G, and $\frac{1}{n}\ln|G| \leq \frac{2}{n}\ln n$, one could reasonably argue that Δ_G is the only important structural term.) It is important to note that the edges in the model need not necessarily correspond to concrete relationships in the data. For example, there are many ways to define the "influential" neighbors of a user in a social network, though the user may be connected to nearly everyone in the network; the adjacencies one models may be a subset of the true adjacencies. Therefore, Δ_G and $|G|$ are quantities that one can control; they become part of the trade-off between representational power and overfitting. In light of this trade-off, recall that the stability term, β_h, partially depends on whether one conditions on the observations in the edge features; as shown in Examples 3 and 4, β_h can be reduced to $\mathrm{O}(\|\mathbf{w}\|_2)$ if one does not. On the other hand, if observations are modeled in the edge features, and $\Delta_G = \mathrm{O}(\sqrt{n})$, then the bounds become $\tilde{\mathrm{O}}(1/\sqrt{m})$. Thus, under this modeling assumption, controlling the maximum degree is critical.

Our improved generalization rate critically relies on the dependency matrix, $\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\mathcal{B}}$, having low infinity norm. If this condition does not hold—for instance, suppose every variable has some non-negligible dependence on every other variable, and $\left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\mathcal{B}}\right\|_\infty = \mathrm{O}(n)$—then our bounds are no more optimistic than previous results and may in fact be slightly looser than some. However, if the dependence is sub-logarithmic, i.e., $\left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\mathcal{B}}\right\|_\infty = \mathrm{O}(\ln n)$, then our bounds are much more optimistic. In Section 4.3, we examined two settings in which this assumption holds; these settings can be characterized by the following conditions: strong local signal, bounded interactions (i.e., degree), and dependence that decays with graph distance. Since the data distribution is determined by nature, it is not a variable one can control. There may be situations in which the mixing coefficients can be estimated from data, as done by McDonald et al. (2011) for β-mixing time series. We leave this as a question for future research. Identifying weaker sufficient dependence conditions is also of interest.

There are several ways in which our analysis can be refined and extended. In Lemma 2, which we use to establish the stability of the ramp loss, we used a rather course application of Hölder's inequality to isolate the influence of the weights. This technique ignores the relative magnitudes of the node and edge weights. Indeed, it may be the case that the edge weights are significantly lower than the node weights. A finer analysis of the weights could improve Equation 33 and might yield new insights for weight regularization. One could also abstract the desirable properties of the potential functions to accommodate a broader class than the linear potentials used in our examples. Finally, we conjecture that our bounds could be tightened by adapting Germain et al.'s (2009) analysis to bound $\phi^2(h, \hat{\mathbf{Z}}) \triangleq \left(\overline{L}(h) - \hat{L}(h, \hat{\mathbf{Z}})\right)^2$ instead of $\phi(h, \hat{\mathbf{Z}}) \triangleq \overline{L}(\mathbb{Q}) - \hat{L}(\mathbb{Q}, \hat{\mathbf{Z}})$. The primary challenge would be bounding the moment-generating function, $\mathbb{E}_{\hat{\mathbf{Z}}\sim\mathbb{D}^m}\left[e^{u\,\phi^2(h,\hat{\mathbf{Z}})}\right]$, since our martingale-based method would not work. If successful, this analysis could yield bounds that tighten when the empirical loss is small.

Acknowledgments

This paper is dedicated to the memory of our friend and collaborator, Ben Taskar. This work was supported by the National Science Foundation (NSF), under grant number IIS1218488, and by the Intelligence Advanced Research Projects Activity (IARPA), via Department of Interior National Business Center (DoI/NBC) contract number D12PC00337. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF, IARPA, DoI/NBC, or the U.S. Government.

Appendix A. Proofs from Section 4

This appendix contains the deferred proofs from Section 6. We begin with some supplemental background in measure concentration. We then prove Proposition 1, and derive a concentration inequality implied by the result. We conclude with the proofs of Propositions 2 and 3.

A.1 The Method of Bounded Differences

Our proof of Proposition 1 follows McDiarmid's *method of bounded differences* (McDiarmid, 1989), which uses a construction known as a *Doob martingale difference sequence*. Let $\varphi : \mathcal{Z}^n \to \mathbb{R}$ denote a measurable function. Let $\mathbf{Z} \triangleq (Z_i)_{i=1}^n$ denote a set of random variables with joint distribution $\mathbb{D}$, and let $\mu \triangleq \mathbb{E}[\varphi(\mathbf{Z})]$ denote the mean of φ. For $i \in [n]$, let

$$V_i \triangleq \mathbb{E}[\varphi(\mathbf{Z}) \,|\, \mathbf{Z}_{1:i}] - \mathbb{E}[\varphi(\mathbf{Z}) \,|\, \mathbf{Z}_{1:i-1}],$$

where $V_1 \triangleq \mathbb{E}[\varphi(\mathbf{Z}) \,|\, \mathbf{Z}_1] - \mu$. The sequence $(V_1, \ldots, V_n)$ has the convenient property that

$$\sum_{i=1}^{n} V_i = \varphi(\mathbf{Z}) - \mu.$$

Therefore, using the law of total expectation, we have that, for any $\tau \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}\left[e^{\tau(\varphi(\mathbf{Z})-\mu)}\right] &= \mathbb{E}\left[\prod_{i=1}^{n} e^{\tau V_i}\right] \\
&= \mathbb{E}\left[\left(\prod_{i=1}^{n-1} e^{\tau V_i}\right) \mathbb{E}\left[e^{\tau V_n} \,|\, \mathbf{Z}_{1:n-1}\right]\right] \\
&\leq \mathbb{E}\left[\prod_{i=1}^{n-1} e^{\tau V_i}\right] \sup_{\mathbf{z} \in \mathcal{Z}^{n-1}} \mathbb{E}\left[e^{\tau V_n} \,|\, \mathbf{Z}_{1:n-1} = \mathbf{z}\right] \\
&\vdots \\
&\leq \prod_{i=1}^{n} \sup_{\mathbf{z} \in \mathcal{Z}^{i-1}} \mathbb{E}\left[e^{\tau V_i} \,|\, \mathbf{Z}_{1:i-1} = \mathbf{z}\right]. \qquad (45)
\end{aligned}$$

Note that the order in which we condition on variables is arbitrary, and does not necessarily need to correspond to any spatio-temporal process. The important property is that the

sequence of σ-algebras generated by the conditioned variables are *nested* (McDiarmid (1998) called this a *filter*), which is guaranteed by the construction of $(V_1, \ldots, V_n)$.

One can then use Hoeffding's lemma (Hoeffding, 1963) to bound each term in the above product.

Lemma 5. *If ξ is a random variable, such that $\mathbb{E}[\xi] = 0$ and $a \leq \xi \leq b$ almost surely, then for any $\tau \in \mathbb{R}$,*

$$\mathbb{E}\left[e^{\tau\xi}\right] \leq \exp\left(\frac{\tau^2(b-a)^2}{8}\right).$$

Clearly, $\mathbb{E}[V_i \,|\, \mathbf{Z}_{1:i-1}] = 0$. Thus, if, for all $i \in [n]$, there exists a value $c_i \geq 0$ such that

$$\sup_{\mathbf{z} \in \mathcal{Z}^{i-1}} \sup_{z \in \mathcal{Z}} (V_i) - \inf_{z' \in \mathcal{Z}} (V_i) = \sup_{\substack{\mathbf{z} \in \mathcal{Z}^{i-1} \\ z, z' \in \mathcal{Z}}} \mathbb{E}\left[\varphi(\mathbf{Z}) \,|\, \mathbf{Z}_{1:i} = (\mathbf{z}, z)\right] - \mathbb{E}\left[\varphi(\mathbf{Z}) \,|\, \mathbf{Z}_{1:i} = (\mathbf{z}, z')\right] \leq c_i,$$

then

$$\mathbb{E}\left[e^{\tau(\varphi(\mathbf{Z}) - \mu)}\right] \leq \prod_{i=1}^{n} \exp\left(\frac{\tau^2 c_i^2}{8}\right) = \exp\left(\frac{\tau^2}{8} \sum_{i=1}^{n} c_i^2\right).$$

When $Z_1, \ldots, Z_n$ are mutually independent, and φ has β-uniformly stability, upper-bounding c_i is straightforward; it becomes complicated when we relax the independence assumption, or when φ is not uniformly stable. The following section addresses the former challenge.

A.2 Coupling

To analyze interdependent random variables, we use a theoretical construction known as *coupling*. For random variables Z_1 and Z_2, with respective distributions $\mathbb{D}_1$ and $\mathbb{D}_2$ over a common sample space $\mathcal{Z}$, a coupling is any joint distribution $\hat{\mathbb{D}}$ over $\mathcal{Z} \times \mathcal{Z}$ such that the marginal distributions, $\hat{\mathbb{D}}(Z_1)$ and $\hat{\mathbb{D}}(Z_2)$, are equal to $\mathbb{D}_1(Z_1)$ and $\mathbb{D}_2(Z_2)$ respectively.

Using a construction due to Fiebig (1993), one can create a coupling of two sequences of random variables, such that the probability that any two corresponding variables are different is upper-bounded by the ϑ-mixing coefficients in Definition 6. The following is an adaptation of this result (due to Samson, 2000) for continuous domains.

Lemma 6. *Let $\mathbf{Z}^{(1)} \triangleq (Z_i^{(1)})_{i=1}^n$ and $\mathbf{Z}^{(2)} \triangleq (Z_i^{(2)})_{i=1}^n$ be random variables with respective distributions $\mathbb{D}_1$ and $\mathbb{D}_2$ over a sample space $\mathcal{Z}^n$. Then there exists a coupling $\hat{\mathbb{D}}$, with marginal distributions $\hat{\mathbb{D}}(\mathbf{Z}^{(1)}) = \mathbb{D}_1(\mathbf{Z}^{(1)})$ and $\hat{\mathbb{D}}(\mathbf{Z}^{(2)}) = \mathbb{D}_2(\mathbf{Z}^{(2)})$, such that, for any $i \in [n]$,*

$$\Pr_{(\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}) \sim \hat{\mathbb{D}}} \left\{ Z_i^{(1)} \neq Z_i^{(2)} \right\} \leq \left\| \mathbb{D}_1\big(\mathbf{Z}_{i:n}^{(1)}\big) - \mathbb{D}_2\big(\mathbf{Z}_{i:n}^{(2)}\big) \right\|_{\mathrm{TV}},$$

where $\Pr_{(\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}) \sim \hat{\mathbb{D}}} \{ Z_i^{(1)} \neq Z_i^{(2)} \}$ denotes the marginal probability that $Z_i^{(1)} \neq Z_i^{(2)}$ under $\hat{\mathbb{D}}$.

Note that the requirement of strictly positive densities is not restrictive, since one can always construct a positive density from a simply nonnegative one. We defer to Samson (2000) for details.

We are now equipped with the tools to prove Proposition 1.

A.3 Proof of Proposition 1

Conditioned on $\overline{\mathcal{B}}$, every realization of $\mathbf{Z}$ is in the "good" set. We define a Doob martingale difference sequence, using the filtration $\boldsymbol{\pi}$:

$$V_i^{\boldsymbol{\pi}} \triangleq \mathbb{E}\left[\varphi(\mathbf{Z}) \,|\, \overline{\mathcal{B}}, \mathbf{Z}_{\pi_i(1:i)}\right] - \mathbb{E}\left[\varphi(\mathbf{Z}) \,|\, \overline{\mathcal{B}}, \mathbf{Z}_{\pi_i(1:i-1)}\right],$$

where $V_1^{\boldsymbol{\pi}} \triangleq \mathbb{E}[\varphi(\mathbf{Z}) \,|\, \overline{\mathcal{B}}, \mathbf{Z}_{\pi_i(1)}] - \mathbb{E}[\varphi(\mathbf{Z}) \,|\, \overline{\mathcal{B}}]$. Note that $\mathbb{E}[V_i^{\boldsymbol{\pi}} \,|\, \overline{\mathcal{B}}] = 0$ and, for $\mathbf{Z} \notin \mathcal{B}_{\mathcal{Z}}$,

$$\sum_{i=1}^{n} V_i^{\boldsymbol{\pi}} = \varphi(\mathbf{Z}) - \mathbb{E}[\varphi(\mathbf{Z}) \,|\, \overline{\mathcal{B}}].$$

We therefore have, via Equation 45, that

$$\mathbb{E}\left[e^{\tau(\varphi(\mathbf{Z}) - \mathbb{E}[\varphi(\mathbf{Z}) \,|\, \overline{\mathcal{B}}])} \,|\, \overline{\mathcal{B}}\right] \leq \prod_{i=1}^{n} \sup_{\mathbf{z} \in \mathcal{Z}^{i-1}_{\boldsymbol{\pi}, \overline{\mathcal{B}}}} \mathbb{E}\left[e^{\tau V_i^{\boldsymbol{\pi}}} \,|\, \overline{\mathcal{B}}, \mathbf{Z}_{\pi_i(1:i-1)} = \mathbf{z}\right],$$

where the supremum over $\mathcal{Z}^{i-1}_{\boldsymbol{\pi}, \overline{\mathcal{B}}}$ comes from the fact that the expectations are conditioned on $\overline{\mathcal{B}}$. Recall that each permutation in $\boldsymbol{\pi}$ has the same prefix, thus preserving the order of conditioned variables, and ensuring that the sequence of σ-algebras is nested.

What remains is to show that, for all $i \in [n]$,

$$\begin{aligned}
&\sup_{\mathbf{z} \in \mathcal{Z}^{i-1}_{\boldsymbol{\pi}, \overline{\mathcal{B}}}} \sup_{z \in \mathcal{Z}^{i}_{\boldsymbol{\pi}, \overline{\mathcal{B}}}(\mathbf{z})} (V_i^{\boldsymbol{\pi}}) - \inf_{z' \in \mathcal{Z}^{i}_{\boldsymbol{\pi}, \overline{\mathcal{B}}}(\mathbf{z})} (V_i^{\boldsymbol{\pi}}) \\
&\quad = \sup_{\substack{\mathbf{z} \in \mathcal{Z}^{i-1}_{\boldsymbol{\pi}, \overline{\mathcal{B}}} \\ z, z' \in \mathcal{Z}^{i}_{\boldsymbol{\pi}, \overline{\mathcal{B}}}(\mathbf{z})}} \mathbb{E}\left[\varphi(\mathbf{Z}) \,|\, \overline{\mathcal{B}}, \mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z)\right] - \mathbb{E}\left[\varphi(\mathbf{Z}) \,|\, \overline{\mathcal{B}}, \mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z')\right] \qquad (46)
\end{aligned}$$

is bounded, so as to apply Lemma 5. (Again, the suprema over $\mathcal{Z}^{i}_{\boldsymbol{\pi}, \overline{\mathcal{B}}}(\mathbf{z})$ stem from conditioning on $\overline{\mathcal{B}}$.) To do so, we will use the coupling construction from Lemma 6. Fix any $\mathbf{z} \in \mathcal{Z}^{i-1}_{\boldsymbol{\pi}, \overline{\mathcal{B}}}$ and $z, z' \in \mathcal{Z}^{i}_{\boldsymbol{\pi}, \overline{\mathcal{B}}}(\mathbf{z})$, and let $N \triangleq n - i$. Define random variables $\boldsymbol{\xi}^{(1)} \triangleq (\xi_j^{(1)})_{j=1}^{N}$ and $\boldsymbol{\xi}^{(2)} \triangleq (\xi_j^{(2)})_{j=1}^{N}$, with coupling distribution $\hat{\mathbb{D}}$ such that

$$\begin{aligned}
&\hat{\mathbb{D}}\big(\boldsymbol{\xi}^{(1)}\big) \triangleq \mathbb{D}\left(\mathbf{Z}_{\pi_i(i+1:n)} \,|\, \overline{\mathcal{B}}, \mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z)\right) \\
\text{and} \quad &\hat{\mathbb{D}}\big(\boldsymbol{\xi}^{(2)}\big) \triangleq \mathbb{D}\left(\mathbf{Z}_{\pi_i(i+1:n)} \,|\, \overline{\mathcal{B}}, \mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z')\right). \qquad (47)
\end{aligned}$$

In other words, the marginal distributions of $\boldsymbol{\xi}^{(1)}$ and $\boldsymbol{\xi}^{(2)}$ are equal to the conditional distributions of $\mathbf{Z}_{\pi_i(i+1:n)}$ given $\overline{\mathcal{B}}$ and, respectively, $\mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z)$ or $\mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z')$. Note that we have renumbered the coupled variables according to π_i. This does not affect the distribution, but it does affect how we later apply Lemma 6. Denote by π_i^{-1} the inverse of π_i (i.e., $\pi_i^{-1}(\pi_i(1:n)) = [n]$), and let

$$\psi(\mathbf{z}) = \varphi\big(\mathbf{z}_{\pi_i^{-1}(1:n)}\big).$$

Put simply, ψ inverts the permutation applied to its input, so as to ensure $\psi(\mathbf{z}_{\pi_i(1:n)}) = \varphi(\mathbf{z})$. For convenience, let

$$\Delta\psi \triangleq \psi\big(\mathbf{z}, z, \boldsymbol{\xi}^{(1)}\big) - \psi\big(\mathbf{z}, z', \boldsymbol{\xi}^{(2)}\big)$$

denote the difference. Using these definitions, we have the following equivalence:

$$\mathbb{E}\left[\varphi(\mathbf{Z}) \,|\, \overline{\mathcal{B}}, \mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z)\right] - \mathbb{E}\left[\varphi(\mathbf{Z}) \,|\, \overline{\mathcal{B}}, \mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z')\right] = \mathbb{E}\left[\psi\big(\mathbf{z}, z, \boldsymbol{\xi}^{(1)}\big) - \psi\big(\mathbf{z}, z', \boldsymbol{\xi}^{(2)}\big)\right].$$

Because the expectations are conditioned on $\overline{\mathcal{B}}$, both realizations, $(\mathbf{z}, z, \boldsymbol{\xi}^{(1)})$ and $(\mathbf{z}, z', \boldsymbol{\xi}^{(2)})$, are "good," in the sense that Equation 1 holds. We therefore have that

$$\begin{aligned}
\mathbb{E}\left[\psi\big(\mathbf{z}, z, \boldsymbol{\xi}^{(1)}\big) - \psi\big(\mathbf{z}, z', \boldsymbol{\xi}^{(2)}\big)\right] &\leq \beta \;\; \mathbb{E}\left[D_{\mathrm{H}}((\mathbf{z}, z, \boldsymbol{\xi}^{(1)}), (\mathbf{z}, z', \boldsymbol{\xi}^{(2)}))\right] \\
&\leq \beta \left(1 + \mathbb{E}\left[\sum_{j=1}^{N} \mathbb{1}\{\xi_j^{(1)} \neq \xi_j^{(2)}\}\right]\right) \\
&= \beta \left(1 + \sum_{j=1}^{N} \Pr_{(\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}) \sim \hat{\mathbb{D}}} \left\{\xi_j^{(1)} \neq \xi_j^{(2)}\right\}\right).
\end{aligned}$$

In the second inequality, we assumed that $z \neq z'$. Recall from Lemma 6 and Definition 6 that

$$\begin{aligned}
&1 + \sum_{j=1}^{N} \Pr_{(\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}) \sim \hat{\mathbb{D}}} \left\{\xi_j^{(1)} \neq \xi_j^{(2)}\right\} \\
&\quad \leq 1 + \sum_{j=i+1}^{n} \left\| \mathbb{D}\left(\mathbf{Z}_{\pi_i(j:n)} \,|\, \overline{\mathcal{B}}, \mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z)\right) - \mathbb{D}\left(\mathbf{Z}_{\pi_i(j:n)} \,|\, \overline{\mathcal{B}}, \mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z')\right) \right\|_{\mathrm{TV}} \\
&\quad = 1 + \sum_{j=i+1}^{n} \vartheta_{ij}^{\boldsymbol{\pi}}(\mathbf{z}, z, z') \\
&\quad \leq 1 + \sum_{j=i+1}^{n} \gamma_{ij}^{\boldsymbol{\pi}} = \sum_{j=i}^{n} \gamma_{ij}^{\boldsymbol{\pi}}.
\end{aligned}$$

The above inequalities hold uniformly for all $\mathbf{z} \in \mathcal{Z}_{\boldsymbol{\pi}, \overline{\mathcal{B}}}^{i-1}$ and $z, z' \in \mathcal{Z}_{\boldsymbol{\pi}, \overline{\mathcal{B}}}^{i}(\mathbf{z})$; thus,

$$\sup_{\mathbf{z} \in \mathcal{Z}_{\boldsymbol{\pi}, \overline{\mathcal{B}}}^{i-1}} \; \sup_{z \in \mathcal{Z}_{\boldsymbol{\pi}, \overline{\mathcal{B}}}^{i}(\mathbf{z})} (V_i^{\boldsymbol{\pi}}) - \inf_{z' \in \mathcal{Z}_{\boldsymbol{\pi}, \overline{\mathcal{B}}}^{i}(\mathbf{z})} (V_i^{\boldsymbol{\pi}}) \;\leq\; \beta \sum_{j=i}^{n} \gamma_{ij}^{\boldsymbol{\pi}}.$$

Then, since we have identified a uniform upper bound for Equation 46, we apply Lemma 5 and obtain

$$\begin{aligned}
\mathbb{E}\left[e^{\tau(\varphi(\mathbf{Z}) - \mathbb{E}[\varphi(\mathbf{Z}) \,|\, \overline{\mathcal{B}}])} \,|\, \overline{\mathcal{B}}\right] &\leq \exp\left(\frac{\tau^2}{8} \sum_{i=1}^{n} \left(\beta \sum_{j=i}^{n} \gamma_{ij}^{\boldsymbol{\pi}}\right)^2\right) \\
&\leq \exp\left(\frac{\tau^2}{8} n\beta^2 \max_{i \in [n]} \left(\sum_{j=i}^{n} \gamma_{ij}^{\boldsymbol{\pi}}\right)^2\right) \\
&= \exp\left(\frac{\tau^2}{8} n\beta^2 \left\|\boldsymbol{\Gamma}_{\overline{\mathcal{B}}}^{\boldsymbol{\pi}}\right\|_{\infty}^2\right),
\end{aligned}$$

which completes the proof.

A.4 A New Concentration Inequality

Proposition 1, implies the following concentration inequality, which may be of independent interest.

Corollary 3. *Let* $\mathbf{Z} \triangleq (Z_i)_{i=1}^n$ *denote random variables with joint distribution* $\mathbb{D}$ *on* $\mathcal{Z}^n$*, and let* $\varphi : \mathcal{Z}^n \to \mathbb{R}$ *denote a measurable function. If* φ *is* β*-uniformly stable, then, for any* $\epsilon > 0$ *and* $\boldsymbol{\pi} \in \Pi(n)$*,*

$$\Pr\left\{\varphi(\mathbf{Z}) - \mathbb{E}[\varphi(\mathbf{Z})] \geq \epsilon\right\} \leq \exp\left(\frac{-2\epsilon^2}{n\beta^2 \left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}\right\|_\infty^2}\right).$$

Proof First, note that, for any $\tau \in \mathbb{R}$,

$$\Pr\left\{\varphi(\mathbf{Z}) - \mathbb{E}[\varphi(\mathbf{Z})] \geq \epsilon\right\} = \Pr\left\{e^{\tau(\varphi(\mathbf{Z}) - \mathbb{E}[\varphi(\mathbf{Z})])} \geq e^{\tau\epsilon}\right\},$$

due to the monotonicity of exponentiation. We then apply Markov's inequality and obtain

$$\Pr\left\{e^{\tau(\varphi(\mathbf{Z}) - \mathbb{E}[\varphi(\mathbf{Z})])} \geq e^{\tau\epsilon}\right\} \leq \frac{1}{e^{\tau\epsilon}} \mathbb{E}\left[e^{\tau(\varphi(\mathbf{Z}) - \mathbb{E}[\varphi(\mathbf{Z})])}\right].$$

Since φ has β-uniform stability, we can apply Proposition 1 by taking $\mathcal{B}_{\mathcal{Z}} \triangleq \emptyset$. Thus,

$$\Pr\left\{e^{\tau(\varphi(\mathbf{Z}) - \mathbb{E}[\varphi(\mathbf{Z})])} \geq e^{\tau\epsilon}\right\} \leq \frac{1}{e^{\tau\epsilon}} \exp\left(\frac{\tau^2}{8}\, n\, \beta^2 \left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}\right\|_\infty^2\right).$$

Optimizing with respect to τ, we take $\tau \triangleq \frac{4\epsilon}{n\beta^2 \left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}\right\|_\infty^2}$ to complete the proof. ■

Corollary 3 extends some current state-of-the-art results (e.g., Kontorovich and Ramanan, 2008, Theorem 1.1) by supporting filtrations of the mixing coefficients. Further, when $Z_1, \ldots, Z_n$ are mutually independent (i.e., $\left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}\right\|_\infty = 1$), we recover McDiarmid's inequality.

A.5 Proof of Proposition 2

We construct the filtration $\boldsymbol{\pi}$ recursively. We initialize π_1 using a breadth-first traversal of the graph, starting from any node. Then, for $i = 2, \ldots, n$, we set $\pi_i(1 : i-1) \triangleq \pi_{i-1}(1 : i-1)$, and determine $\pi_i(i : n)$ using a breadth-first traversal over the induced subgraph of $\pi_{i-1}(i : n)$, starting from $\pi_{i-1}(i-1)$. This ensures that nodes closer to $\pi_i(i)$ appear earlier in the permutation, so that the higher mixing coefficients are not incurred for all $j = i+1, \ldots, n$.

The degree of any node in this induced subgraph is at most the maximum degree of the whole graph, Δ_G, so the number of nodes at distance k from node $\pi_i(i)$ is at most Δ_G^k. Hence, the number of subsets, $\pi_i(j : n) : j > i$, at distance k from $\pi_i(i)$ is at most Δ_G^k. Therefore,

$$\sum_{j=i}^{n} \gamma_{ij}^{\boldsymbol{\pi}} \leq \sum_{k=0}^{\infty} \Delta_G^k\, \vartheta(k) \leq \sum_{k=0}^{\infty} \left(\frac{\Delta_G}{\Delta_G + \epsilon}\right)^k.$$

Since $\Delta_G/(\Delta_G+\epsilon) < 1$ for $\epsilon > 0$, this geometric series converges to

$$\frac{1}{1-\Delta_G/(\Delta_G+\epsilon)} = 1 + \Delta_G/\epsilon,$$

which completes the proof.

A.6 Proof of Proposition 3

For a chain graph, we define each permutation uniformly as $\pi_i \triangleq [n]$. Each upper-triangular entry of $\mathbf{\Gamma}^{\boldsymbol{\pi}}$ then satisfies $\gamma^{\boldsymbol{\pi}}_{ij} \leq \vartheta(j-i)$. The number of unconditioned variables at distance $k = j - i$ is exactly one. Thus, for any row i,

$$\sum_{j=i}^{n} \gamma^{\boldsymbol{\pi}}_{ij} \leq 1 + \sum_{k=1}^{n-i} \vartheta(k) \leq 1 + \epsilon \sum_{k=1}^{n-i} k^{-p}.$$

For $p = 1$, $(k^{-p})_{k=1}^{\infty}$ is a Harmonic series. Thus, the partial sum, $\sum_{k=1}^{n-i} k^{-p}$, is the $(n-i)^{\text{th}}$ Harmonic number, which is upper-bounded by $\ln(n-i)+1$, and maximized at row $i = 1$. For $p > 1$,

$$1 + \epsilon \sum_{k=1}^{n-i} k^{-p} \leq 1 + \epsilon \sum_{k=1}^{\infty} k^{-p} = 1 + \zeta(p),$$

by definition.

Appendix B. Proofs from Section 5

This appendix contains the deferred proofs from Section 5.

B.1 Proof of Theorem 2

For $i = 0, 1, 2, \ldots$, let $\beta_i \triangleq 2^{i+1}$. Since Equation 2 fails with probability $\delta + m\nu$, we could simply invoke Theorem 1 for each β_i with $\delta_i \triangleq \beta_i^{-1}(\delta + m\nu)$. This approach would introduce an additional $\mathrm{O}\left(\ln(m\nu)^{-1}\right)$ term in the numerator of Equation 23. We therefore choose instead to cover β and u simultaneously. Accordingly, for $j = 0, 1, 2, \ldots$, let

$$u_{ij} \triangleq 2^j \sqrt{\frac{8mn \ln \frac{2\beta_i}{\delta}}{\beta_i^2 \left\| \mathbf{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}} \right\|_\infty^2}}.$$

Each β_i defines a set of "bad" hypotheses, $\mathcal{B}^i_{\mathcal{H}}$, which we use in Equation 5 to define a function $\tilde{\phi}_i$. Let $\delta_{ij} \triangleq \delta \beta_i^{-1} 2^{-(j+1)}$, and define an event

$$E_{ij} \triangleq \mathbb{1}\left\{ \mathop{\mathbb{E}}_{h \sim \mathbb{P}} \left[e^{u_{ij} \tilde{\phi}_i(h, \hat{\mathbf{Z}})} \right] \geq \frac{1}{\delta_{ij}} \exp\left(\frac{u_{ij}^2 \beta_i^2 \left\| \mathbf{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}} \right\|_\infty^2}{8mn} \right) \right\}.$$

Note that none of the above depend on $(\beta, \eta, \mathbb{Q})$. Using the event B defined in Equation 12, we have, via Proposition 1, that

$$\Pr_{\hat{\mathbf{Z}} \sim \mathbb{D}^m} \{ E_{ij} \,|\, \neg B \} \leq \delta_{ij} \exp\left(-\frac{u_{ij}^2 \beta_i^2 \left\| \mathbf{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}} \right\|_\infty^2}{8mn} \right) \mathop{\mathbb{E}}_{h \sim \mathbb{P}} \mathop{\mathbb{E}}_{\hat{\mathbf{Z}} \sim \mathbb{D}^m} \left[e^{u_{ij} \tilde{\phi}_i(h, \hat{\mathbf{Z}})} \,|\, \neg B \right] \leq \delta_{ij}.$$

Then, using the same reasoning as Equation 13, with $E \triangleq \bigcup_{i=0}^{\infty} \bigcup_{j=0}^{\infty} E_{ij}$,

$$\begin{aligned}
\Pr_{\hat{\mathbf{Z}} \sim \mathbb{D}^m} \{B \cup E\} &\leq m\nu + \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \Pr_{\hat{\mathbf{Z}} \sim \mathbb{D}^m} \{E_{ij} \,|\, \neg B\} \\
&\leq m\nu + \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \delta_{ij} \\
&= m\nu + \delta \sum_{i=0}^{\infty} \beta_i^{-1} \sum_{j=0}^{\infty} 2^{-(j+1)} \\
&= m\nu + \delta \sum_{i=0}^{\infty} 2^{-(i+1)} \sum_{j=0}^{\infty} 2^{-(j+1)} \\
&= m\nu + \delta.
\end{aligned}$$

Therefore, with probability at least $1 - \delta - m\nu$, every $l \in [m]$ satisfies $\mathbf{Z}^{(l)} \notin \mathcal{B}_{\mathcal{Z}}$, and every (i, j) satisfies

$$\mathop{\mathbb{E}}_{h \sim \mathbb{P}} \left[e^{u_{ij} \tilde{\phi}_i(h, \hat{\mathbf{Z}})} \right] \leq \frac{1}{\delta_{ij}} \exp\left(\frac{u_{ij}^2 \beta_i^2 \left\| \mathbf{\Gamma}_{\overline{\mathcal{B}}}^{\boldsymbol{\pi}} \right\|_{\infty}^2}{8mn} \right). \tag{48}$$

Observe that $(\beta/n, \mathcal{B}_{\mathcal{Z}}, \eta)$-local stability implies $(\beta_j/n, \mathcal{B}_{\mathcal{Z}}, \eta)$-local stability for all $\beta_j \geq \beta$. Therefore, for any particular $(\beta, \eta, \mathbb{Q})$ such that $\mathbb{Q}$ is $(\beta/n, \mathcal{B}_{\mathcal{Z}}, \eta)$-locally stable, we select $i^\star \triangleq \left\lfloor (\ln 2)^{-1} \ln \beta \right\rfloor$. This ensures that $\beta \leq \beta_{i^\star}$, so $\mathbb{Q}$ also satisfies $(\beta_{i^\star}/n, \mathcal{B}_{\mathcal{Z}}, \eta)$-local stability. Then, letting

$$j^\star \triangleq \left\lfloor \frac{1}{2 \ln 2} \ln \left(\frac{D_{\mathrm{KL}}(\mathbb{Q} \| \mathbb{P})}{\ln(2\beta_{i^\star}/\delta)} + 1 \right) \right\rfloor,$$

we have that

$$\frac{1}{2} \sqrt{\frac{8mn \left(D_{\mathrm{KL}}(\mathbb{Q} \| \mathbb{P}) + \ln \frac{2\beta_{i^\star}}{\delta} \right)}{\beta_{i^\star}^2 \left\| \mathbf{\Gamma}_{\overline{\mathcal{B}}}^{\boldsymbol{\pi}} \right\|_{\infty}^2}} \leq u_{i^\star j^\star} \leq \sqrt{\frac{8mn \left(D_{\mathrm{KL}}(\mathbb{Q} \| \mathbb{P}) + \ln \frac{2\beta_{i^\star}}{\delta} \right)}{\beta_{i^\star}^2 \left\| \mathbf{\Gamma}_{\overline{\mathcal{B}}}^{\boldsymbol{\pi}} \right\|_{\infty}^2}}. \tag{49}$$

Moreover,

$$\begin{aligned}
D_{\mathrm{KL}}(\mathbb{Q} \| \mathbb{P}) + \ln \frac{1}{\delta_{i^\star j^\star}} &\leq D_{\mathrm{KL}}(\mathbb{Q} \| \mathbb{P}) + \ln \frac{2\beta_{i^\star}}{\delta} + \frac{1}{2} \ln \left(\frac{D_{\mathrm{KL}}(\mathbb{Q} \| \mathbb{P})}{\ln(2\beta_{i^\star}/\delta)} + 1 \right) \\
&\leq D_{\mathrm{KL}}(\mathbb{Q} \| \mathbb{P}) + \ln \frac{2\beta_{i^\star}}{\delta} + \frac{1}{2} \left(D_{\mathrm{KL}}(\mathbb{Q} \| \mathbb{P}) + \ln \frac{2\beta_{i^\star}}{\delta} \right).
\end{aligned} \tag{50}$$

Thus, with probability at least $1 - \delta - m\nu$,

$$\begin{aligned}
\overline{L}(\mathbb{Q}) - \hat{L}(\mathbb{Q}, \hat{\mathbf{Z}}) &\leq \alpha(\eta + \nu) + \frac{1}{u_{i^\star j^\star}} \left(D_{\mathrm{KL}}(\mathbb{Q} \| \mathbb{P}) + \ln \mathop{\mathbb{E}}_{h \sim \mathbb{P}} \left[e^{u_{i^\star j^\star} \tilde{\phi}_{i^\star}(h, \hat{\mathbf{Z}})} \right] \right) \\
&\leq \alpha(\eta + \nu) + \frac{1}{u_{i^\star j^\star}} \left(D_{\mathrm{KL}}(\mathbb{Q} \| \mathbb{P}) + \ln \frac{1}{\delta_{i^\star j^\star}} + \frac{u_{i^\star j^\star}^2 \beta_{i^\star}^2 \left\| \mathbf{\Gamma}^{\boldsymbol{\pi}}_{\mathcal{B}} \right\|_\infty^2}{8mn} \right) \\
&\leq \alpha(\eta + \nu) + \frac{3 \left(D_{\mathrm{KL}}(\mathbb{Q} \| \mathbb{P}) + \ln \frac{2\beta_{i^\star}}{\delta} \right)}{2u_{i^\star j^\star}} + \frac{u_{i^\star j^\star} \beta_{i^\star}^2 \left\| \mathbf{\Gamma}^{\boldsymbol{\pi}}_{\mathcal{B}} \right\|_\infty^2}{8mn} \\
&\leq \alpha(\eta + \nu) + 2\beta_{i^\star} \left\| \mathbf{\Gamma}^{\boldsymbol{\pi}}_{\mathcal{B}} \right\|_\infty \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q} \| \mathbb{P}) + \ln \frac{2\beta_{i^\star}}{\delta}}{2mn}}.
\end{aligned}$$

The first inequality uses Equation 9; the second uses Equation 48; the third and fourth use Equations 49 and 50. Noting that $\beta_{i^\star} \leq 2\beta$ completes the proof.

B.2 Proof of Proposition 5

Fix any $h \in \mathcal{H}$ and $\mathbf{z} \notin \mathcal{B}_{\mathcal{Z}}$. By Definition 10, there exists a set $\mathcal{B}_{\mathcal{H}}(h)$ with measure $\mathbb{Q}_h(\mathcal{B}_{\mathcal{H}}(h)) \leq \eta$. For any $\mathbf{z} \notin \mathcal{B}_{\mathcal{Z}}$, let $\mathcal{B}_{\mathcal{H}}(h, \mathbf{z}) \triangleq \mathcal{B}_{\mathcal{H}}(h)$, and note that $\mathbb{Q}_h(\mathcal{B}_{\mathcal{H}}(h, \mathbf{z})) \leq \eta$ as well. Further, for any $h' \notin \mathcal{B}_{\mathcal{H}}(h, \mathbf{z})$, $\|h - h'\| \leq \beta$. Thus, by Definition 9,

$$\left| L(h, \mathbf{z}) - L(h', \mathbf{z}) \right| \leq \lambda \left\| h - h' \right\| \leq \lambda\beta,$$

which completes the proof.

Appendix C. Proofs from Section 6

This appendix contains the deferred proofs from Section 6. Certain proofs require the following technical lemmas, which apply to the linear feature functions defined in Section 2.2.3.

Lemma 7. *Fix a graph, $G \triangleq (\mathcal{V}, \mathcal{E})$, with maximum degree Δ_G. Suppose $\mathcal{X}$ is uniformly bounded by the p-norm ball with radius R; i.e., $\sup_{x \in \mathcal{X}} \|x\|_p \leq R$. Then, for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}^n$ and $\mathbf{y} \in \mathcal{Y}^n$,*

$$\left\| \mathbf{f}(\mathbf{x}, \mathbf{y}) - \mathbf{f}(\mathbf{x}', \mathbf{y}) \right\|_p \leq (\Delta_G + 2) R \, D_{\mathrm{H}}(\mathbf{x}, \mathbf{x}'). \tag{51}$$

Further, if the model does not use edge observations (i.e., $f_{ij}(\mathbf{x}, \mathbf{y}) \triangleq y_i \otimes y_j$), then

$$\left\| \mathbf{f}(\mathbf{x}, \mathbf{y}) - \mathbf{f}(\mathbf{x}', \mathbf{y}) \right\|_p \leq 2R \, D_{\mathrm{H}}(\mathbf{x}, \mathbf{x}'). \tag{52}$$

Proof We start by considering a pair, $\mathbf{x}, \mathbf{x}' \in \mathcal{X}^n : D_{\mathrm{H}}(\mathbf{x}, \mathbf{x}') = 1$, that differ at a single coordinate, corresponding to a node i. This means that the aggregate features differ at one local feature, and any edge involving i. Thus, using the triangle inequality, we have that

$$\begin{aligned}
\left\| \mathbf{f}(\mathbf{x}, \mathbf{y}) - \mathbf{f}(\mathbf{x}', \mathbf{y}) \right\|_p &= \left\| \begin{bmatrix} f_i(\mathbf{x}, \mathbf{y}) - f_i(\mathbf{x}', \mathbf{y}) \\ \sum_{j : \{i,j\} \in \mathcal{E}} f_{ij}(\mathbf{x}, \mathbf{y}) - f_{ij}(\mathbf{x}', \mathbf{y}) \end{bmatrix} \right\|_p \\
&\leq \left\| f_i(\mathbf{x}, \mathbf{y}) - f_i(\mathbf{x}', \mathbf{y}) \right\|_p + \sum_{j : \{i,j\} \in \mathcal{E}} \left\| f_{ij}(\mathbf{x}, \mathbf{y}) - f_{ij}(\mathbf{x}', \mathbf{y}) \right\|_p .
\end{aligned} \tag{53}$$

Note that the second term disappears when the model does not use edge observations.

Recall that the features are defined using a Kronecker product. For any vectors $\mathbf{u}, \mathbf{v}$, $\|\mathbf{u} \otimes \mathbf{v}\|_p = \|\mathbf{u}\|_p \|\mathbf{v}\|_p$. Using this identity, and the fact that each $y \in \mathcal{Y}$ has $\|y\|_1 = 1$, we have that

$$\begin{aligned}
\left\|f_i(\mathbf{x}, \mathbf{y}) - f_i(\mathbf{x}', \mathbf{y})\right\|_p &= \left\|(x_i - x_i') \otimes y_i\right\|_p \\
&= \left\|x_i - x_i'\right\|_p \|y_i\|_p \\
&\leq \left(\|x_i\|_p + \left\|x_i'\right\|_p\right) \times 1 \\
&\leq 2R,
\end{aligned}$$

and

$$\begin{aligned}
\left\|f_{ij}(\mathbf{x}, \mathbf{y}) - f_{ij}(\mathbf{x}', \mathbf{y})\right\|_p &= \left\|\frac{1}{2}\left(\begin{bmatrix} x_i \\ x_j \end{bmatrix} - \begin{bmatrix} x_i' \\ x_j \end{bmatrix}\right) \otimes (y_i \otimes y_j)\right\|_p \\
&= \frac{1}{2}\left\|x_i - x_i'\right\|_p \|y_i\|_p \|y_j\|_p \\
&\leq \frac{1}{2}\left(\|x_i\|_p + \left\|x_i'\right\|_p\right) \times 1 \times 1 \\
&\leq R.
\end{aligned}$$

Combining these inequalities with Equation 53, and using the fact that i participates in at most Δ_G edges, we have that

$$\left\|\mathbf{f}(\mathbf{x}, \mathbf{y}) - \mathbf{f}(\mathbf{x}', \mathbf{y})\right\|_p \leq 2R + \sum_{j:\{i,j\} \in \mathcal{E}} R \leq (2 + \Delta_G)R.$$

For no edge observations, the righthand side is simply $2R$. Thus, since the bounds hold for any single coordinate perturbation, Equations 51 and 52 follow from the triangle inequality. ■

Lemma 8. *Fix a graph, $G \triangleq (\mathcal{V}, \mathcal{E})$, and recall that $|G| \triangleq |\mathcal{V}| + |\mathcal{E}|$. Suppose $\mathcal{X}$ is uniformly bounded by the p-norm ball with radius R; i.e., $\sup_{x \in \mathcal{X}} \|x\|_p \leq R$. Then, for all $\mathbf{x} \in \mathcal{X}^n$ and $\mathbf{y} \in \mathcal{Y}^n$,*

$$\|\mathbf{f}(\mathbf{x}, \mathbf{y})\|_p \leq |G|\, R.$$

Proof Invoking the triangle inequality, we have that

$$\begin{aligned}
\left\|\mathbf{f}(\mathbf{x},\mathbf{y})\right\|_p &= \left\| \begin{bmatrix} \sum_{i\in\mathcal{V}} f_i(\mathbf{x},\mathbf{y}) \\ \sum_{\{i,j\}\in\mathcal{E}} f_{ij}(\mathbf{x},\mathbf{y}) \end{bmatrix} \right\|_p \\
&\leq \sum_{i\in\mathcal{V}} \left\|f_i(\mathbf{x},\mathbf{y})\right\|_p + \sum_{\{i,j\}\in\mathcal{E}} \left\|f_{ij}(\mathbf{x},\mathbf{y})\right\|_p \\
&= \sum_{i\in\mathcal{V}} \left\|x_i \otimes y_i\right\|_p + \sum_{\{i,j\}\in\mathcal{E}} \left\| \frac{1}{2} \begin{bmatrix} x_i \\ x_j \end{bmatrix} \otimes (y_i \otimes y_j) \right\|_p \\
&= \sum_{i\in\mathcal{V}} \left\|x_i\right\|_p \left\|y_i\right\|_p + \sum_{\{i,j\}\in\mathcal{E}} \frac{1}{2} \left\| \begin{bmatrix} x_i \\ x_j \end{bmatrix} \right\|_p \left\|y_i\right\|_p \left\|y_j\right\|_p \\
&\leq \sum_{i\in\mathcal{V}} \left\|x_i\right\|_p \left\|y_i\right\|_p + \sum_{\{i,j\}\in\mathcal{E}} \frac{1}{2} \left(\left\|x_i\right\|_p + \left\|x_j\right\|_p \right) \left\|y_i\right\|_p \left\|y_j\right\|_p \\
&\leq \sum_{i\in\mathcal{V}} R \times 1 + \sum_{\{i,j\}\in\mathcal{E}} \frac{1}{2} \left(R + R\right) \times 1 \times 1 \\
&= (|\mathcal{V}| + |\mathcal{E}|)R = |G|\, R,
\end{aligned}$$

which completes the proof. ■

Note that Lemmas 7 and 8 hold when discrete labels are replaced with marginals, since each clique's marginals sum to one. This adaptation enables the proof of Example 3.

C.1 Proof of Lemma 2

To simplify notation, let:

$$\begin{aligned}
\mathbf{y}_1 &\triangleq \arg\max_{\mathbf{u}\in\mathcal{Y}^n} D_{\mathrm{H}}(\mathbf{y},\mathbf{u}) + h(\mathbf{x},\mathbf{u}); & \mathbf{y}_2 &\triangleq \arg\max_{\mathbf{u}\in\mathcal{Y}^n} h(\mathbf{x},\mathbf{u}); \\
\mathbf{y}'_1 &\triangleq \arg\max_{\mathbf{u}\in\mathcal{Y}^n} D_{\mathrm{H}}(\mathbf{y}',\mathbf{u}) + h(\mathbf{x}',\mathbf{u}); & \mathbf{y}'_2 &\triangleq \arg\max_{\mathbf{u}\in\mathcal{Y}^n} h(\mathbf{x}',\mathbf{u}).
\end{aligned}$$

Using this notation, we have that

$$\begin{aligned}
&n \left| L_{\mathrm{r}}(h,\mathbf{z}) - L_{\mathrm{r}}(h,\mathbf{z}') \right| \\
&\quad = \left| (D_{\mathrm{H}}(\mathbf{y},\mathbf{y}_1) + h(\mathbf{x},\mathbf{y}_1) - h(\mathbf{x},\mathbf{y}_2)) - \left(D_{\mathrm{H}}(\mathbf{y}',\mathbf{y}'_1) + h(\mathbf{x}',\mathbf{y}'_1) - h(\mathbf{x}',\mathbf{y}'_2)\right) \right| \\
&\quad \leq \left| (D_{\mathrm{H}}(\mathbf{y},\mathbf{y}_1) + h(\mathbf{x},\mathbf{y}_1)) - \left(D_{\mathrm{H}}(\mathbf{y}',\mathbf{y}'_1) + h(\mathbf{x}',\mathbf{y}'_1)\right) \right| + \left| h(\mathbf{x},\mathbf{y}_2) - h(\mathbf{x}',\mathbf{y}'_2) \right|, \quad (54)
\end{aligned}$$

using the triangle inequality.

Focusing on the second absolute difference, we can assume, without loss of generality, that $h(\mathbf{x},\mathbf{y}_2) \geq h(\mathbf{x}',\mathbf{y}'_2)$, meaning

$$\begin{aligned}
\left| h(\mathbf{x},\mathbf{y}_2) - h(\mathbf{x}',\mathbf{y}'_2) \right| &= h(\mathbf{x},\mathbf{y}_2) - h(\mathbf{x}',\mathbf{y}'_2) \\
&\leq h(\mathbf{x},\mathbf{y}_2) - h(\mathbf{x}',\mathbf{y}_2) \\
&= \mathbf{w} \cdot \left(\mathbf{f}(\mathbf{x},\mathbf{y}_2) - \mathbf{f}(\mathbf{x}',\mathbf{y}_2)\right) \\
&\leq \left\|\mathbf{w}\right\|_q \left\|\mathbf{f}(\mathbf{x},\mathbf{y}_2) - \mathbf{f}(\mathbf{x}',\mathbf{y}_2)\right\|_p \\
&\leq \left\|\mathbf{w}\right\|_q (\Delta_G + 2) R\, D_{\mathrm{H}}(\mathbf{x},\mathbf{x}'). \quad (55)
\end{aligned}$$

The first inequality uses the optimality of $\mathbf{y}_2'$, implying $-h(\mathbf{x}', \mathbf{y}_2') \leq -h(\mathbf{x}', \mathbf{y}_2)$; the second inequality uses Hölder's inequality; the third inequality uses Lemma 7 (Equation 51). Note that we obtain the same upper bound if we assume that $h(\mathbf{x}, \mathbf{y}_2) \leq h(\mathbf{x}', \mathbf{y}_2')$, since we can reverse the terms inside the absolute value and proceed with $\mathbf{y}_2'$ instead of $\mathbf{y}_2$.

We now return to the first absolute difference. To reduce clutter, it will help to use the loss-augmented potentials, $\tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w})$, from Equation 31. Recall that $\delta(\mathbf{y})$ denotes the loss augmentation vector for $\mathbf{y}$. We then have that

$$\left|(D_{\text{H}}(\mathbf{y}, \mathbf{y}_1) + h(\mathbf{x}, \mathbf{y}_1)) - \left(D_{\text{H}}(\mathbf{y}', \mathbf{y}_1') + h(\mathbf{x}', \mathbf{y}_1')\right)\right| = \left|\tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}) \cdot \hat{\mathbf{y}}_1 - \tilde{\boldsymbol{\theta}}(\mathbf{x}', \mathbf{y}'; \mathbf{w}) \cdot \hat{\mathbf{y}}_1'\right|.$$

If we assume (without loss of generality) that $\tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}) \cdot \hat{\mathbf{y}}_1 \geq \tilde{\boldsymbol{\theta}}(\mathbf{x}', \mathbf{y}'; \mathbf{w}) \cdot \hat{\mathbf{y}}_1'$, then

$$\begin{aligned}
\left|\tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}) \cdot \hat{\mathbf{y}}_1 - \tilde{\boldsymbol{\theta}}(\mathbf{x}', \mathbf{y}'; \mathbf{w}) \cdot \hat{\mathbf{y}}_1'\right| &= \tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}) \cdot \hat{\mathbf{y}}_1 - \tilde{\boldsymbol{\theta}}(\mathbf{x}', \mathbf{y}'; \mathbf{w}) \cdot \hat{\mathbf{y}}_1' \\
&\leq \tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}) \cdot \hat{\mathbf{y}}_1 - \tilde{\boldsymbol{\theta}}(\mathbf{x}', \mathbf{y}'; \mathbf{w}) \cdot \hat{\mathbf{y}}_1 \\
&= \left(\boldsymbol{\theta}(\mathbf{x}; \mathbf{w}) + \delta(\mathbf{y}) - \boldsymbol{\theta}(\mathbf{x}'; \mathbf{w}) - \delta(\mathbf{y}')\right) \cdot \hat{\mathbf{y}}_1 \\
&= \mathbf{w} \cdot \left(\mathbf{f}(\mathbf{x}, \mathbf{y}_1) - \mathbf{f}(\mathbf{x}', \mathbf{y}_1)\right) + \left(\delta(\mathbf{y}) - \delta(\mathbf{y}')\right) \cdot \hat{\mathbf{y}}_1 \\
&\leq \|\mathbf{w}\|_q \, (\Delta_G + 2) R \, D_{\text{H}}(\mathbf{x}, \mathbf{x}') + \left(\delta(\mathbf{y}) - \delta(\mathbf{y}')\right) \cdot \hat{\mathbf{y}}_1 \\
&\leq \|\mathbf{w}\|_q \, (\Delta_G + 2) R \, D_{\text{H}}(\mathbf{x}, \mathbf{x}') + D_{\text{H}}(\mathbf{y}, \mathbf{y}'). \qquad (56)
\end{aligned}$$

The first inequality uses the optimality of $\mathbf{y}_1'$; the second inequality uses Hölder's inequality and Lemma 7 again; the last inequality uses the fact that

$$\left(\delta(\mathbf{y}) - \delta(\mathbf{y}')\right) \cdot \hat{\mathbf{y}}_1 = D_{\text{H}}(\mathbf{y}, \mathbf{y}_1) - D_{\text{H}}(\mathbf{y}', \hat{\mathbf{y}}_1) \leq D_{\text{H}}(\mathbf{y}, \mathbf{y}').$$

The upper bound in Equation 56 also holds when $\tilde{\boldsymbol{\theta}}(\mathbf{x}', \mathbf{y}'; \mathbf{w}) \cdot \hat{\mathbf{y}}_1' \geq \tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}) \cdot \hat{\mathbf{y}}_1$.

Combining Equations 55 to 57, we then have that

$$\begin{aligned}
n \left|L_{\text{r}}(h, \mathbf{z}) - L_{\text{r}}(h, \mathbf{z}')\right| &\leq 2(\Delta_G + 2) R \, \|\mathbf{w}\|_q \; D_{\text{H}}(\mathbf{x}, \mathbf{x}') + D_{\text{H}}(\mathbf{y}, \mathbf{y}') \\
&\leq 2(\Delta_G + 2) R \, \|\mathbf{w}\|_q \; D_{\text{H}}(\mathbf{z}, \mathbf{z}') + D_{\text{H}}(\mathbf{z}, \mathbf{z}').
\end{aligned}$$

Dividing both sides by n yields Equation 33. To obtain Equation 34, we use Lemma 7's Equation 52 in Equations 55 and 56, which reduces the term $(\Delta_G + 2)$ to just 2.

C.2 Proof of Lemma 3

The proof proceeds similarly to that of Lemma 2. Let

$$\begin{aligned}
\mathbf{y}_1 &\triangleq \operatorname*{arg\,max}_{\mathbf{u} \in \mathcal{Y}^n} D_{\text{H}}(\mathbf{y}, \mathbf{u}) + h(\mathbf{x}, \mathbf{u}); & \mathbf{y}_2 &\triangleq \operatorname*{arg\,max}_{\mathbf{u} \in \mathcal{Y}^n} h(\mathbf{x}, \mathbf{u}); \\
\mathbf{y}_1' &\triangleq \operatorname*{arg\,max}_{\mathbf{u} \in \mathcal{Y}^n} D_{\text{H}}(\mathbf{y}, \mathbf{u}) + h'(\mathbf{x}, \mathbf{u}); & \mathbf{y}_2' &\triangleq \operatorname*{arg\,max}_{\mathbf{u} \in \mathcal{Y}^n} h'(\mathbf{x}, \mathbf{u}).
\end{aligned}$$

Using this notation, we have that

$$\begin{aligned}
&n \left|L_{\text{r}}(h, \mathbf{z}) - L_{\text{r}}(h', \mathbf{z})\right| \\
&\quad \leq \left|(D_{\text{H}}(\mathbf{y}, \mathbf{y}_1) + h(\mathbf{x}, \mathbf{y}_1)) - \left(D_{\text{H}}(\mathbf{y}, \mathbf{y}_1') + h'(\mathbf{x}, \mathbf{y}_1')\right)\right| + \left|h(\mathbf{x}, \mathbf{y}_2) - h'(\mathbf{x}, \mathbf{y}_2')\right|, \quad (57)
\end{aligned}$$

via the triangle inequality. Assuming $h(\mathbf{x}, \mathbf{y}_2) \geq h'(\mathbf{x}, \mathbf{y}'_2)$, we have that

$$\begin{aligned}
\left|h(\mathbf{x}, \mathbf{y}_2) - h'(\mathbf{x}, \mathbf{y}'_2)\right| &= h(\mathbf{x}, \mathbf{y}_2) - h'(\mathbf{x}, \mathbf{y}'_2) \\
&\leq h(\mathbf{x}, \mathbf{y}_2) - h'(\mathbf{x}, \mathbf{y}_2) \\
&= (\mathbf{w} - \mathbf{w}') \cdot \mathbf{f}(\mathbf{x}, \mathbf{y}_2) \\
&\leq \left\|\mathbf{w} - \mathbf{w}'\right\|_q \left\|\mathbf{f}(\mathbf{x}, \mathbf{y}_2)\right\|_p \\
&\leq \left\|\mathbf{w} - \mathbf{w}'\right\|_q |G|\, R, \qquad (58)
\end{aligned}$$

via Lemma 8. Further, using the loss-augmented potentials, and assuming $\tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}) \cdot \hat{\mathbf{y}}_1 \geq \tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}') \cdot \hat{\mathbf{y}}'_1$, we have that

$$\begin{aligned}
\left|\left(D_{\text{H}}(\mathbf{y}, \mathbf{y}_1) + h(\mathbf{x}, \mathbf{y}_1)\right) - \left(D_{\text{H}}(\mathbf{y}, \mathbf{y}'_1) + h'(\mathbf{x}, \mathbf{y}'_1)\right)\right| &= \tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}) \cdot \hat{\mathbf{y}}_1 - \tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}') \cdot \hat{\mathbf{y}}'_1 \\
&\leq \tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}) \cdot \hat{\mathbf{y}}_1 - \tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}') \cdot \hat{\mathbf{y}}_1 \\
&= \left(\boldsymbol{\theta}(\mathbf{x}; \mathbf{w}) + \delta(\mathbf{y}) - \boldsymbol{\theta}(\mathbf{x}; \mathbf{w}') - \delta(\mathbf{y})\right) \cdot \hat{\mathbf{y}}_1 \\
&= (\mathbf{w} - \mathbf{w}') \cdot \mathbf{f}(\mathbf{x}, \mathbf{y}_1) \\
&\leq \left\|\mathbf{w} - \mathbf{w}'\right\|_q \left\|\mathbf{f}(\mathbf{x}, \mathbf{y}_1)\right\|_p \\
&\leq \left\|\mathbf{w} - \mathbf{w}'\right\|_q |G|\, R. \qquad (59)
\end{aligned}$$

Combining the inequalities and dividing by n completes the proof.

C.3 Proof of Example 1

Since the weights are uniformly bounded, we define the prior, $\mathbb{P}$, as a uniform distribution on the d-dimensional unit ball. Given a (learned) hypothesis, h, with weights $\mathbf{w}$, we construct a posterior, $\mathbb{Q}_h$, as a uniform distribution on a d-dimensional ball with radius ϵ, centered at $\mathbf{w}$, and clipped at the boundary of the unit ball; i.e., its support is $\{\mathbf{w}' \in \mathbb{R}^d : \|\mathbf{w}' - \mathbf{w}\|_2 \leq \epsilon,\ \|\mathbf{w}'\|_2 \leq 1\}$. We let $\epsilon \triangleq (m\,|G|)^{-1}$, meaning the radius of the ball should decrease as the size of the training set increases.

For a uniform distribution, $\mathbb{U}$, with *support* $\text{supp}(\mathbb{U}) \subseteq \mathcal{H}$, we denote its *volume* by

$$\text{vol}(\mathbb{U}) \triangleq \int_{\mathcal{H}} \mathbb{1}\{h \in \text{supp}(\mathbb{U})\}\ \text{d}h.$$

The probability density function of $\mathbb{U}$ is the inverse of its volume. The volume of $\mathbb{P}$ is the volume of a unit ball, which is proportional to 1. Similarly, the volume of $\mathbb{Q}_h$ is at least the volume of a d-dimensional ball with radius $\epsilon/2$ (due to the intersection with the unit ball), which is proportional to $(\epsilon/2)^d$.[12] Therefore, using p and q_h to denote their respective

12. We withhold the precise definitions for simplicity of exposition. It will suffice to recognize their relative proportions, since the withheld constant depends only on d, and is thereby canceled out in the KL divergence.

densities, we have that

$$\begin{aligned} D_{\mathrm{KL}}(\mathbb{Q}_h \| \mathbb{P}) &= \int_{\mathcal{H}} q_h(h') \ln \frac{q_h(h')}{p(h')} \, \mathrm{d}h' \\ &= \int_{\mathcal{H}} q_h(h') \ln \frac{\mathrm{vol}(\mathbb{P})}{\mathrm{vol}(\mathbb{Q}_h)} \, \mathrm{d}h' \\ &\leq \int_{\mathcal{H}} q_h(h') \ln (2/\epsilon)^d \, \mathrm{d}h' \\ &= d \ln(2m \left|G\right|). \end{aligned}$$

By assumption, every allowable hypothesis has a weight vector $\mathbf{w}$ with $\left\|\mathbf{w}\right\|_2 \leq 1$. We also assume that $\sup_{x \in \mathcal{X}} \left\|x\right\|_2 \leq 1$. Therefore, with $R = 1$ and $\beta \triangleq (2\Delta_G + 4) + 1$, Lemma 2 immediately proves that $L_{\mathrm{r}} \circ \{h \in \mathcal{H}_{\mathrm{M3N}} : \left\|\mathbf{w}\right\|_2 \leq 1\}$ is (β/n)-uniformly stable. Invoking Corollary 1, we then have that, with probability at least $1 - \delta$, every $\mathbb{Q}_h : \left\|\mathbf{w}\right\|_2 \leq 1$ satisfies

$$\overline{L}_{\mathrm{r}}(\mathbb{Q}_h) \leq \hat{L}_{\mathrm{r}}(\mathbb{Q}_h, \hat{\mathbf{Z}}) + 2\left((2\Delta_G + 4) + 1\right) \left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}\right\|_{\infty} \sqrt{\frac{d \ln(2m \left|G\right|) + \ln \frac{2}{\delta}}{2mn}}. \tag{60}$$

By construction, every $h' \sim \mathbb{Q}_h$ satisfies $\left\|\mathbf{w}' - \mathbf{w}\right\|_2 \leq (m \left|G\right|)^{-1}$, so $\mathbb{Q}$ has $(1/(m \left|G\right|), 0)$-local hypothesis stability. As demonstrated in Equation 37, L_{r} has $(2 \left|G\right| / n, \emptyset)$-local hypothesis stability. Thus, via Proposition 5, $(L_{\mathrm{r}}, \mathbb{Q})$ has $(2/(mn), \emptyset, 0)$-local stability. Then, via Proposition 4 and Equation 32, we have that

$$\overline{L}_{\mathrm{H}}(h) \leq \overline{L}_{\mathrm{r}}(h) \leq \overline{L}_{\mathrm{r}}(\mathbb{Q}_h) + \frac{2}{mn}, \tag{61}$$

and

$$\hat{L}_{\mathrm{r}}(\mathbb{Q}_h, \hat{\mathbf{Z}}) \leq \hat{L}_{\mathrm{r}}(h, \hat{\mathbf{Z}}) + \frac{2}{mn} \leq \hat{L}_{\mathrm{h}}(h, \hat{\mathbf{Z}}) + \frac{2}{mn}. \tag{62}$$

Combining Equations 60 to 62 completes the proof.

C.4 Proof of Lemma 4

We begin with a fundamental property of the normal distribution, which is used to prove the concentration inequality.

Fact 1. *If X is a Gaussian random variable, with mean μ and variance σ^2, then, for any $\epsilon > 0$,*

$$\Pr\left\{\left|X - \mu\right| \geq \epsilon\right\} \leq 2 \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right). \tag{63}$$

Observe that, if $\left\|\mathbf{X} - \boldsymbol{\mu}\right\|_p \geq \epsilon$, then there must exist at least one coordinate $i \in [d]$ such that $\left|X_i - \mu_i\right| \geq \epsilon / d^{1/p}$; otherwise, we would have

$$\left\|\mathbf{X} - \boldsymbol{\mu}\right\|_p = \left(\sum_{i=1}^{d} \left|X_i - \mu_i\right|^p\right)^{1/p} < \left(d \left(\frac{\epsilon}{d^{1/p}}\right)^p\right)^{1/p} = \epsilon.$$

We therefore have that

$$\begin{aligned}\Pr\left\{\|\mathbf{X}-\boldsymbol{\mu}\|_p \geq \epsilon\right\} &\leq \Pr\left\{\exists i : |X_i - \mu_i| \geq \frac{\epsilon}{d^{1/p}}\right\} \\ &\leq \sum_{i=1}^{d} \Pr\left\{|X_i - \mu_i| \geq \frac{\epsilon}{d^{1/p}}\right\} \\ &\leq \sum_{i=1}^{d} 2\exp\left(-\frac{\epsilon^2}{2\sigma^2 d^{2/p}}\right).\end{aligned}$$

The second inequality uses the union bound; the last uses Fact 1. Summing over $i = 1, \ldots, d$ completes the proof.

C.5 Proof of Example 4

We first show that $\mathbb{D}(\mathcal{B}_{\mathcal{Z}}) \leq 1/n$. Then, the rest of the proof is a simple modification of the previous analyses.

Observe that, for any x and μ_y,

$$\|x\|_2 - 1 \leq \|x\|_2 - \|\mu_y\|_2 \leq \|x - \mu_y\|_2 .$$

So, if $\|x\|_2 \geq 2$, then $\|x - \mu_y\|_2 \geq 1$. Therefore, using the union bound, and Lemma 4, we can upper-bound the measure of $\mathcal{B}_{\mathcal{Z}}$ as follows:

$$\begin{aligned}\mathbb{D}(\mathcal{B}_{\mathcal{Z}}) &= \Pr_{\mathbf{Z}\sim\mathbb{D}}\{\exists i : \|X_i\|_2 \geq 2\} \\ &\leq \sup_{\mathbf{y}\in\mathcal{Y}^n} \Pr_{\mathbf{X}\sim\mathbb{D}}\{\exists i : \|X_i\|_2 \geq 2 \,|\, \mathbf{Y} = \mathbf{y}\} \\ &= \sup_{\mathbf{y}\in\mathcal{Y}^n} \sum_{i=1}^{n} \Pr_{X_i\sim\mathbb{D}}\{\|X_i\|_2 \geq 2 \,|\, Y_i = y_i\} \\ &\leq \sup_{\mathbf{y}\in\mathcal{Y}^n} \sum_{i=1}^{n} \Pr_{X_i\sim\mathbb{D}}\left\{\|X_i - \mu_{y_i}\|_2 \geq 1 \,|\, Y_i = y_i\right\} \\ &\leq \sup_{\mathbf{y}\in\mathcal{Y}^n} \sum_{i=1}^{n} 2k\exp\left(-\frac{1}{2k\sigma_{y_i}^2}\right) \\ &\leq \sum_{i=1}^{n} 2k\exp\left(-\frac{2k\ln(2kn^2)}{2k}\right) = \frac{1}{n}.\end{aligned}$$

Conditioned on $\mathbf{Z} \notin \mathcal{B}_{\mathcal{Z}}$, we have that Lemmas 7 and 8 hold for $R = 2$; hence, so do Lemmas 2 and 3. With $\mathbb{P}$, $\mathbb{Q}_h$ and $\mathcal{B}_{\mathcal{H}_{\text{M3N}}}(h)$ constructed identically to Example 2, this means that $\mathbb{Q}_h$ is $(\beta_h/n, \mathcal{B}_{\mathcal{Z}}, 1/(mn))$-locally stable. Further, L_{r} has $(4\,|G|\,/n, \mathcal{B}_{\mathcal{Z}})$-local hypothesis stability, and $\mathbb{Q}$ has $(1/(m\,|G|), 1/(mn))$-local hypothesis stability; by Proposition 5, this means that $(L_{\text{r}}, \mathbb{Q})$ has $(4/(mn), \mathcal{B}_{\mathcal{Z}}, 1/(mn))$-local stability. Thus, invoking Theorem 2 and Proposition 4, with $\nu = 1/n$, we have that, with probability at least $1 - \delta - m/n$, all

$l \in [m]$ satisfy $\mathbf{Z}^{(l)} \notin \mathcal{B}_{\mathcal{Z}}$, and all $h \in \mathcal{H}_{\text{M3N}}$ satisfy

$$\begin{aligned}\overline{L}_{\text{H}}(h) &\leq \overline{L}_{\text{r}}(\mathbb{Q}_h) + \frac{5}{mn} + \frac{1}{n} \\ &\leq \hat{L}_{\text{r}}(\mathbb{Q}_h, \hat{\mathbf{Z}}) + \frac{6}{mn} + \frac{2}{n} \\ &\qquad + 4\beta_h \left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{S}}}\right\|_\infty \sqrt{\frac{\frac{1}{2}\left\|\mathbf{w}\right\|_2^2 + \frac{d}{2}\ln\left(2d(m\left|G\right|)^2\ln(2dmn)\right) + \ln\frac{4\beta_h}{\delta}}{2mn}}.\end{aligned}$$

Further, since none of the training examples in the sample are "bad," we also have that

$$\hat{L}_{\text{r}}(\mathbb{Q}_h, \hat{\mathbf{Z}}) \leq \hat{L}_{\text{r}}(h, \hat{\mathbf{Z}}) + \frac{5}{mn} \leq \hat{L}_{\text{h}}(h, \hat{\mathbf{Z}}) + \frac{5}{mn}.$$

Combining these inequalities completes the proof.

References

P. Alquier and O. Wintenburger. Model selection for weakly dependent time series forecasting. *Bernoulli*, 18(3):883–913, 2012.

A. Ambroladze, E. Parrado-Hernández, and J. Shawe-Taylor. Tighter PAC-Bayes bounds. In *Neural Information Processing Systems*, 2006.

P. Bartlett, M. Collins, D. McAllester, and B. Taskar. Large margin methods for structured classification: Exponentiated gradient algorithms and PAC-Bayesian generalization bounds. Extended version of paper appearing in Advances in Neural Information Processing Systems 17, 2005.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

J. Bradley and C. Guestrin. Sample complexity of composite likelihood. In *Artificial Intelligence and Statistics*, 2012.

R. Bradley. Basic properties of strong mixing conditions: A survey and some open questions. *Probability Surveys*, 2(2):107–144, 2005.

O. Catoni. *Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *Institute of Mathematical Statistics Lecture Notes – Monograph Series*. Institute of Mathematical Statistics, 2007.

H. Chan and A. Darwiche. Sensitivity analysis in Markov networks. In *International Joint Conference on Artificial Intelligence*, 2005.

H. Chan and A. Darwiche. On the robustness of most probable explanations. In *Uncertainty in Artificial Intelligence*, 2006.

J. Chazottes, P. Collet, C. Külske, and F. Redig. Concentration inequalities for random fields via coupling. *Probability Theory and Related Fields*, 137:201–225, 2007.

M. Collins. Parameter estimation for statistical parsing models: Theory and practice of distribution-free methods. In *International Conference on Parsing Technologies*, 2001.

M. Donsker and S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

D. Fiebig. Mixing properties of a class of Bernoulli processes. *Transactions of the American Mathematical Society*, 338:479–492, 1993.

P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. PAC-Bayesian learning of linear classifiers. In *International Conference on Machine Learning*, 2009.

S. Giguère, F. Laviolette, M. Marchand, and K. Sylla. Risk bounds and learning algorithms for the regression approach to structured output prediction. In *International Conference on Machine Learning*, 2013.

K. Gimpel and N. Smith. Softmax-margin CRFs: Training log-linear models with cost functions. In *Conference of the North American Chapter of the Association of Computational Linguistics*, 2010.

T. Hazan and R. Urtasun. A primal-dual message-passing algorithm for approximated large scale structured prediction. In *Neural Information Processing Systems*, 2010.

T. Hazan, S. Maji, J. Keshet, and T. Jaakkola. Learning efficient random maximum a posteriori predictors with non-decomposable loss functions. In *Neural Information Processing Systems*, 2013.

R. Herbrich and T. Graepel. A PAC-Bayesian margin bound for linear classifiers: Why SVMs work. In *Neural Information Processing Systems*, 2001.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

J. Honorio. Lipschitz parametrization of probabilistic graphical models. In *Uncertainty in Artificial Intelligence*, 2011.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

A. Kontorovich. Obtaining measure concentration from Markov contraction. *Markov Processes and Related Fields*, 18:613–638, 2012.

A. Kontorovich and K. Ramanan. Concentration inequalities for dependent random variables via the martingale method. *Annals of Probability*, 36(6):2126–2158, 2008.

S. Kutin. Extensions to McDiarmid's inequality when differences are bounded with high probability. Technical report, University of Chicago, 2002.

J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Intl. Conference on Machine Learning*, 2001.

J. Langford and J. Shawe-Taylor. PAC-Bayes and margins. In *Neural Information Processing Systems*, 2002.

G. Lever, F. Laviolette, and J. Shawe-Taylor. Distribution-dependent PAC-Bayes priors. In *Conference on Algorithmic Learning Theory*, 2010.

B. London, B. Huang, B. Taskar, and L. Getoor. Collective stability in structured prediction: Generalization from one example. In *International Conference on Machine Learning*, 2013.

B. London, B. Huang, B. Taskar, and L. Getoor. PAC-Bayesian collective stability. In *Artificial Intelligence and Statistics*, 2014.

D. McAllester. Some PAC-Bayesian theorems. In *Conference on Computational Learning Theory*, 1998.

D. McAllester. PAC-Bayesian model averaging. In *Conference on Computational Learning Theory*, 1999.

D. McAllester. Simplified PAC-Bayesian margin bounds. In *Conference on Computational Learning Theory*, 2003.

D. McAllester. Generalization bounds and consistency for structured labeling. In G. Bakir, T. Hofmann, B. Schölkopf, A. Smola, B. Taskar, and S. Vishwanathan, editors, *Predicting Structured Data*. MIT Press, 2007.

D. McAllester and J. Keshet. Generalization bounds and consistency for latent structural probit and ramp loss. In *Neural Information Processing Systems*, 2011.

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

C. McDiarmid. Concentration. *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 195–248, 1998.

D. McDonald, C. Shalizi, and M. Schervish. Estimating β-mixing coefficients. In *Artificial Intelligence and Statistics*, 2011.

D. McDonald, C. Shalizi, and M. Schervish. Time series forecasting: model evaluation and selection using nonparametric risk bounds. *CoRR*, abs/1212.0463, 2012.

M. Mohri and A. Rostamizadeh. Rademacher complexity bounds for non-i.i.d. processes. In *Neural Information Processing Systems*, 2009.

M. Mohri and A. Rostamizadeh. Stability bounds for stationary ϕ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 11:789–814, 2010.

J. Neville and D. Jensen. Dependency networks for relational data. In *International Conference on Data Mining*, 2004.

L. Ralaivola, M. Szafranski, and G. Stempfel. Chromatic PAC-Bayes bounds for non-i.i.d. data: Applications to ranking and stationary β-mixing processes. *Journal of Machine Learning Research*, 11:1927–1956, 2010.

M. Richardson and P. Domingos. Markov logic networks. *Machine Learning*, 62(1-2):107–136, 2006.

Dan Roth. On the hardness of approximate reasoning. *Artificial Intelligence*, 82(1-2): 273–302, 1996.

P. Samson. Concentration of measure inequalities for Markov chains and ϕ-mixing processes. *Annals of Probability*, 28(1):416–461, 2000.

M. Seeger. PAC-Bayesian generalization error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 3:233–269, 2002.

Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093, 2012.

B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. In *Uncertainty in Artificial Intelligence*, 2002.

B. Taskar, C. Guestrin, and D. Koller. Max-margin Markov networks. In *Neural Information Processing Systems*, 2004.

I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005.

N. Usunier, M. Amini, and P. Gallinari. Generalization error bounds for classifiers trained with interdependent data. In *Neural Information Processing Systems*, 2006.

V. Vu. Concentration of non-Lipschitz functions and applications. *Random Structures and Algorithms*, 20(3):262–316, 2002.

M. Wainwright. Estimating the "wrong" graphical model: Benefits in the computation-limited setting. *Journal of Machine Learning Research*, 7:1829–1859, 2006.

M. Wainwright and M. Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., 2008.

R. Xiang and J. Neville. Relational learning with one network: An asymptotic analysis. In *Artificial Intelligence and Statistics*, 2011.

B. Yu. Rates of convergence for empirical processes of stationary mixing sequences. *Annals of Probability*, 22(1):94–116, 1994.