PAC-Bayesian Collective Stability

Ben London
University of Maryland

Bert Huang
University of Maryland

Ben Taskar
University of Washington

Lise Getoor
University of California,
Santa Cruz

Abstract

Recent results have shown that the generalization error of structured predictors decreases with both the number of examples and the size of each example, provided the data distribution has weak dependence and the predictor exhibits a smoothness property called *collective stability.* These results use an especially strong definition of collective stability that must hold *uniformly* over all inputs and all hypotheses in the class. We investigate whether weaker definitions of collective stability suffice. Using the PAC-Bayes framework, which is particularly amenable to our new definitions, we prove that generalization is indeed possible when uniform collective stability happens with high probability over draws of predictors (and inputs). We then derive a generalization bound for a class of structured predictors with variably convex inference, which suggests a novel learning objective that optimizes collective stability.

1 INTRODUCTION

London et al. (2013) recently showed that the generalization error of certain structured predictors is better than was previously known. They provided bounds that decrease with both the number of structured examples and the size of each example, thus enabling generalization from even a single example under certain conditions. In doing so, they introduced the notion of *collective stability*, a measure of the sensitivity of a structured predictor to perturbations of its input. In particular, their analysis relied on a restrictive version of this property that must hold *uniformly* over all inputs and all hypotheses in the class. Though they showed that this condition is met by a class of structured predictors used in practice, such a strict definition may not be necessary for generalization.

In this paper, we show that weaker definitions of collective stability enable generalization from a few large, structured examples. Our results are two-fold. First, we relax the requirement that all hypotheses in the class exhibit collective stability; that is, certain hypotheses are robust to all possible input perturbations, but others may not be. Our analysis is based in the PAC-Bayes framework, in which prediction is randomized over draws from a *posterior* distribution over hypotheses. PAC-Bayesian analysis is particularly amenable to relaxing uniform stability, since one can construct a posterior that places more mass on "good" hypotheses and less weight on "bad" ones. We thus derive PAC-Bayes generalization bounds of order

$$\mathrm{O}\left(\Pr\{h \in \textsc{Bad}\} + \sqrt{\frac{\textsc{Complexity}}{mn}}\right),$$

where m is the number of examples, n is the size of each example, and Complexity is measured by the KL divergence between the posterior and a *prior* distribution over hypotheses. Provided the probability of a bad hypothesis, $\Pr\{h \in \textsc{Bad}\}$, is sufficiently low, the generalization error converges to zero in the limit of either infinite examples or infinitely large examples. Our second generalization bound relaxes the stability condition further by requiring that good hypotheses only exhibit stability with respect to a certain subset of the instance space. If this set has sufficient support under the generating distribution, one obtains generalization guarantees of a similar form.

We apply our generalization bounds to two classes of structured predictors. The first class—which achieves uniform collective stability by assuming parameter-tying, bounded weights and a strongly convex inference function—illustrates how our new PAC-Bayesian analysis can achieve the tightest known generalization bounds for structured prediction, of order $\mathrm{O}\left(\sqrt{\frac{\ln n}{mn}}\right)$. The second class relaxes the assumption of bounded

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33. Copyright 2014 by the authors.

weights, and parameterizes the convexity of the inference function. Despite this class not having uniform collective stability, we are still able to derive a generalization bound with comparable decay. Moreover, since the bound is stated in terms of the parameters of the learned hypothesis (rather than uniform upper bounds), it implies a new learning objective that optimizes collective stability by optimizing the convexity of inference.

Our specific contributions are summarized as follows. We first define new, weaker forms of collective stability, and derive some novel concentration inequalities for functions of interdependent random variables. Using these tools, we then derive improved PAC-Bayes bounds for structured prediction. Unlike previous PAC-Bayes bounds, ours decrease proportionally to both the number of examples and the size of each example. To illustrate the implications of our theory, we give two examples of generalization bounds for structured predictors—the latter of which relaxes some assumptions and suggests a novel learning objective.

1.1 Related Work

Until recently, the generalization error of structured predictors was thought to decay proportionally to the number of examples (Taskar et al., 2004; Bartlett et al., 2005; McAllester, 2007; Keshet et al., 2011). London et al. (2013) then showed that, given suitably weak dependence within each example, certain classes of structured predictors are capable of much faster uniform convergence rates. Their analysis crucially relied on a property they referred to as uniform collective stability, which is akin to a global Lipschitz smoothness condition. Our analysis departs from theirs by relaxing the stability requirement to classes with non-uniform collective stability, thus making our bounds applicable to a wider range of predictors, while maintaining comparable generalization error rates.

There is a large body of theory on learning *local* (i.e., non-structured) predictors from various types of interdependent data. For learning problems that induce a *dependency graph*, Usunier et al. (2006) and Ralaivola et al. (2010) used fractional coloring to analyze the generalization error of local predictors. For *ϕ-mixing* and *β-mixing* temporal data, Mohri and Rostamizadeh (2009, 2010) derived risk bounds using an *independent blocking* technique, due to Yu (1994), though the hypotheses they consider predict each time step independently. McDonald et al. (2011) used a similar technique to bound the risk of autoregressive forecasting models, in which the prediction at time t depends on a moving window of previous observations. We analyze a more general setting in which hypotheses perform joint inference over arbitrarily structured examples. In our setting, techniques such as graph coloring and independent blocking do not apply, since the global prediction does not decompose.

PAC-Bayesian analysis was introduced by McAllester (1999) and later refined by a number of authors (e.g., Langford and Shawe-Taylor, 2002; Seeger, 2002; Ambroladze et al., 2006; Germain et al., 2009). Our PAC-Bayes proofs are based on a martingale technique due to Lever et al. (2010) and Seldin et al. (2012). Our application of PAC-Bayes to interdependent data is related to work by Alquier and Wintenburger (2012), though they consider one-step time series forecasting.

Various notions of stability have been used in machine learning. Bousquet and Elisseeff (2002) used the stability of a learning algorithm to derive generalization bounds in the non-structured setting. Chan and Darwiche (2006), Wainwright (2006) and Honorio (2011) analyzed the sensitivity of probabilistic graphical models to changes in parameters. The notion of stability we use builds off of London et al. (2013), who considered the sensitivity of a predictor to changing inputs.

Our analysis uses concentration inequalities for Lipschitz functions of dependent random variables, similar to those presented by Chazottes et al. (2007) and Kontorovich and Ramanan (2008). To accommodate functions that are not uniformly Lipschitz, we adapt a technique used by Kutin (2002) and Vu (2002), and pair it with a coupling construction due to Fiebig (1993).

2 PRELIMINARIES

In the structured prediction framework we consider, each example contains n interdependent random variables, $\mathbf{Z} \triangleq (Z_i)_{i=1}^n$, with joint distribution $\mathbb{P}$. Each $Z_i \triangleq (X_i, Y_i)$ is an input-output pair, taking values in a sample space $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$.[1] We denote *realizations* of $\mathbf{Z}$ by $\mathbf{z} \in \mathcal{Z}^n$. We use the notation $\mathbb{E}_{\mathbf{Z}\sim\mathbb{P}}$ to specify the expectation over $\mathbf{Z}$, unless it is clear from context.

We are interested in predicting $\mathbf{Y} \triangleq (Y_i)_{i=1}^n$, conditioned on $\mathbf{X} \triangleq (X_i)_{i=1}^n$. Let $\mathcal{H} \subseteq \{h : \mathcal{X}^n \to \hat{\mathcal{Y}}^n\}$ denote a class of hypotheses, where $\hat{\mathcal{Y}} \subseteq \mathbb{R}^k$, for some $k \geq 1$. For example, if $\mathcal{Y}$ contains k states, then $h(\mathbf{X})$ returns a score for each Y_i taking each state. We use $h_i(\mathbf{X})$ to denote the prediction vector for Y_i, and $h_i^j(\mathbf{X})$ to denote its j^{th} entry. Let $\mathbb{H}$ denote a predetermined prior distribution on $\mathcal{H}$, and let $\mathbb{Q}$ denote a posterior distribution, possibly learned from training data. In the PAC-Bayes framework, prediction is stochastic. Given an input $\mathbf{X}$, we first draw a hypothesis $h \in \mathcal{H}$,

[1] 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.

according to $\mathbb{Q}$, then compute the prediction $h(\mathbf{X})$.

For a loss function $\ell : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R}_+$ and hypothesis h, denote the average loss on a set of m structured examples, $\hat{\mathbf{Z}} \triangleq (\mathbf{Z}^{(l)})_{l=1}^m = ((Z_i^{(l)})_{i=1}^n)_{l=1}^m$, by

$$L(h, \hat{\mathbf{Z}}) \triangleq \frac{1}{mn} \sum_{l=1}^{m} \sum_{i=1}^{n} \ell \left(Y_i^{(l)}, h_i(\mathbf{X}^{(l)}) \right). \qquad (1)$$

(Decomposable losses, such as this, are common in the tasks we consider.) Let $\overline{L}(h) \triangleq \mathbb{E}_{\mathbf{Z} \sim \mathbb{P}}[L(h, \mathbf{Z})]$ denote the expected loss (also known as the *risk*) over realizations of a single example $\mathbf{Z}$, which corresponds to the error h will incur on future predictions. Since prediction in the PAC-Bayes framework is randomized, we use the expectations of these measures over draws of h, which we denote by

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

We are interested in the difference of $\overline{L}(\mathbb{Q}) - L(\mathbb{Q}, \hat{\mathbf{Z}})$.

3 STRUCTURED PREDICTORS

We are interested in hypotheses that perform joint reasoning over all variables simultaneously, according to some prior knowledge about the structure of the problem. One such model is a *Markov random field* (MRF). An MRF consists of a graph $G \triangleq (\mathcal{V}, \mathcal{E})$ with cliques $\mathcal{C}$, random variables $\mathbf{Z} \triangleq (Z_i)_{i \in \mathcal{V}}$, *feature functions* $\mathbf{f}(\mathbf{Z}) \triangleq (f_c(\mathbf{Z}))_{c \in \mathcal{C}}$, and weights $\mathbf{w} \triangleq (w_c)_{c \in \mathcal{C}}$, which define a distribution $p_{\mathbf{w}}(\mathbf{Z}) \propto \exp(\langle \mathbf{w}, \mathbf{f}(\mathbf{Z}) \rangle)$. The edge set $\mathcal{E}$ captures the dependencies in $\mathbf{Z}$, and is typically determined by the problem structure. For now, assume that the sample space is discrete, and that each feature outputs a basis vector representation wherein $f_c^j(\mathbf{Z}) = 1$ if $\mathbf{Z}_c$ is in its j^{th} state and 0 otherwise.

The canonical inference problems for MRFs are *maximum a posteriori* (MAP) inference, which computes the mode of the distribution, and *marginal* inference, which computes the marginal distribution of a subset of the variables. We represent the marginals of the cliques by a vector $\boldsymbol{\mu} \in \mathbb{R}^N$, where $N \triangleq \sum_{c \in \mathcal{C}} |c|^{|\mathcal{Z}|}$ and μ_c^j indicates the probability that $\mathbf{Z}_c$ is in its j^{th} state. The set of all consistent marginal vectors is called the *marginal polytope*, which we denote by $\mathcal{M}$. When $\mathcal{Z}$ is discrete and the features output the above representation, the marginals are the solution to

$$\mathop{\arg\max}_{\boldsymbol{\mu} \in \mathcal{M}} \langle \mathbf{w}, \boldsymbol{\mu} \rangle + H(\boldsymbol{\mu}), \qquad (2)$$

where $H(\boldsymbol{\mu})$ is the entropy of the distribution consistent with $\boldsymbol{\mu}$ (Wainwright and Jordan, 2008). This identity can be adapted for approximate marginal inference by relaxing $\mathcal{M}$ and replacing H with a tractable surrogate, such as the *Bethe approximation*. Further, Equation 2 has an interesting relationship with MAP inference, in that the mode is given by

$$\mathop{\arg\max}_{\mathbf{z} \in \mathcal{Z}^n} p_{\mathbf{w}}(\mathbf{z}) = \Gamma \left(\mathop{\arg\max}_{\boldsymbol{\mu} \in \mathcal{M}} \langle \mathbf{w}, \boldsymbol{\mu} \rangle \right),$$

where $\Gamma : \mathcal{M} \to \mathcal{Z}^n$ is a linear projection that selects and decodes the unary clique marginals. The key insight is that MAP inference is equivalent to marginal inference without entropy maximization.

For discriminative tasks, in which each Z_i is actually a tuple (X_i, Y_i) of input-output pairs, an MRF can be used to model the conditional distribution $p_{\mathbf{w}}(\mathbf{Y} \,|\, \mathbf{X})$. For an observation $\mathbf{x} \in \mathcal{X}^n$ (regardless of whether $\mathcal{X}$ is discrete), the conditional marginals are the solution to

$$\mathop{\arg\max}_{\boldsymbol{\mu} \in \mathcal{M}_{\mathbf{Y}}} \langle \mathbf{w}, \mathbf{f}(\mathbf{x}, \boldsymbol{\mu}) \rangle + H(\boldsymbol{\mu}),$$

where $\mathcal{M}_{\mathbf{Y}}$ is the marginal polytope of $\mathbf{Y}$, and $\mathbf{f}$ conditions $\boldsymbol{\mu}$ on $\mathbf{x}$ via a linear map. The relationship with MAP inference holds in this case as well.

A common technique for defining MRFs is *templating* (also known as *parameter-tying*). A *clique template* is a complete subgraph pattern, such as a singleton, pair or triangle. Given a graph, a set of templates partitions the cliques into subgraphs with common structure. Thus, a templated MRF replaces the per-clique features and weights with per-template ones, which are then applied to each *grounding* (i.e., matching clique). Since the features are no longer tied to specific groundings, 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), Markov logic networks (Richardson and Domingos, 2006) and hinge-loss MRFs (Bach et al., 2013).

3.1 Templated Structured Models

We now present a general class of models that includes variations of the above graphical models.

Definition 1. A *templated structured model* (TSM) is defined by:

- a search space $\mathcal{S}$;
- a set of clique templates $\mathcal{T}$;
- a set of feature functions $\{f_t\}_{t \in \mathcal{T}}$, with output length $d_t \geq 1$;
- a set of weights $\{w_t \in \mathbb{R}^{d_t}\}_{t \in \mathcal{T}}$;
- a regularizer $\Psi : \mathcal{S} \to \mathbb{R}$;
- a linear projection $\Gamma : \mathcal{S} \to \hat{\mathcal{Y}}^n$.

Given a graph G and input $\mathbf{x} \in \mathcal{X}^n$, let $t(G)$ denote the groundings of G, and let

$$\mathbf{f}(\mathbf{x}, \mathbf{s}) \triangleq \left(\sum_{c \in t(G)} f_t(\mathbf{x}_c, \mathbf{s}_c) \right)_{t \in \mathcal{T}}$$

and $\mathbf{w} \triangleq (w_t)_{t \in \mathcal{T}}$, both of which have (output) length $d \triangleq \sum_{t \in \mathcal{T}} d_t$. Define the *energy function* E as

$$E_{\mathbf{w}}(\mathbf{x}, \mathbf{s}) \triangleq \langle \mathbf{w}, \mathbf{f}(\mathbf{x}, \mathbf{s}) \rangle - \Psi(\mathbf{s}).$$

A TSM hypothesis h outputs

$$h(\mathbf{x}) \triangleq \Gamma \left(\arg\max_{\mathbf{s} \in \mathcal{S}} E_{\mathbf{w}}(\mathbf{x}, \mathbf{s}) \right).$$

The search space, clique templates, feature functions, regularizer and projection are typically chosen *a priori*. The graph for a given input is determined implicitly by the data or prior knowledge of the problem structure. Thus, learning a TSM usually amounts to learning the weights.

Though the TSM representation is abstract, one can show that inference in TSMs is equivalent to inference in some of the previous models. To recreate (approximate) marginal inference in a templated MRF, we define $\mathcal{S}$ as the (local) marginal polytope of $\mathbf{Y}$, and each f_t as a linear map that conditions on $\mathbf{x}_c$; Ψ is (a surrogate for) the negative entropy, and the projection Γ selects the unary clique (pseudo)marginals. We can also recover (approximate) MAP inference by letting $\Psi(\mathbf{s}) \triangleq 0$ and decoding the unary terms.

4 COLLECTIVE STABILITY

A key component of our analysis is the *algorithmic stability* of joint inference. Broadly speaking, stability ensures that small changes to the input result in bounded variation in the output. In learning theory, it has traditionally been used to quantify the variation in the output of a learning algorithm upon adding or removing training examples (Bousquet and Elisseeff, 2002). We apply this concept to an arbitrary class of vector-valued functions, $\mathcal{F} \triangleq \{\varphi : \mathcal{Z}^n \to \mathbb{R}^N\}$, where N does not necessarily equal n. For vectors $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}^n$, denote their Hamming distance by

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

Definition 2. We say that a function $\varphi \in \mathcal{F}$ has β-*uniform collective stability* if, for any inputs $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}^n$,

$$\left\| \varphi(\mathbf{z}) - \varphi(\mathbf{z}') \right\|_1 \leq \beta \, D_{\mathrm{H}}(\mathbf{z}, \mathbf{z}'). \tag{3}$$

Similarly, we say that the class $\mathcal{F}$ has β-uniform collective stability if every $\varphi \in \mathcal{F}$ has β-uniform collective stability.

Put differently, a function with uniform collective stability is Lipschitz under the Hamming norm of its domain and 1-norm of its range.

Though uniform stability seems like a strong requirement, it is met by a broad class of models used in practice (London et al., 2013). Nonetheless, part of the scope of this paper is to explore weaker definitions of collective stability. For example, suppose uniform stability holds for *most* functions in the class, but not all. This is of particular interest in the PAC-Bayes framework, in which a predictor is selected according to a distribution over hypotheses.

Definition 3. Let $\mathbb{Q}$ be a distribution on $\mathcal{F}$. We say that $\mathcal{F}$ has $(\mathbb{Q}, \eta, \beta)$ *collective stability* if there exists a "bad" set $\mathcal{B}_{\mathcal{F}} \subseteq \mathcal{F}$ such that $\mathbb{Q}\{\varphi \in \mathcal{B}_{\mathcal{F}}\} \leq \eta$ and every $\varphi \notin \mathcal{B}_{\mathcal{F}}$ has β-uniform collective stability.

We might also allow that uniform stability holds for *most* inputs, but not all.

Definition 4. Let $\mathbf{Z} \triangleq (Z_i)_{i=1}^n$ be random variables with joint distribution $\mathbb{P}$. We say that $\varphi \in \mathcal{F}$ has $(\mathbb{P}, \nu, \beta)$ *collective stability* if there exists a "bad" set $\mathcal{B} \subseteq \mathcal{Z}^n$ such that $\mathbb{P}\{\mathbf{Z} \in \mathcal{B}\} \leq \nu$ and Equation 3 holds for any $\mathbf{z}, \mathbf{z}' \notin \mathcal{B}$.

A still weaker definition combines Definitions 3 and 4.

Definition 5. Let $\mathbb{P}$ be the distribution of $\mathbf{Z}$, and $\mathbb{Q}$ a distribution on $\mathcal{F}$. We say that $\mathcal{F}$ has $(\mathbb{P}, \nu, \mathbb{Q}, \eta, \beta)$ *collective stability* if there exist "bad" sets $\mathcal{B} \subseteq \mathcal{Z}^n$ and $\mathcal{B}_{\mathcal{H}} \subseteq \mathcal{F}$ such that:

1. $\mathbb{P}\{\mathbf{Z} \in \mathcal{B} \,|\, \varphi \notin \mathcal{B}_{\mathcal{F}}\} \leq \nu$;
2. $\mathbb{Q}\{\varphi \in \mathcal{B}_{\mathcal{F}}\} \leq \eta$;
3. Equation 3 holds for any $\varphi \notin \mathcal{B}_{\mathcal{F}}$ and $\mathbf{z}, \mathbf{z}' \notin \mathcal{B}$.

There is a taxonomical relationship between these definitions, with Definition 2 being the strongest. Clearly, if $\mathcal{F}$ has β-uniform collective stability, then it has $(\mathbb{Q}, 0, \beta)$ collective stability and $(\mathbb{P}, 0, \mathbb{Q}, 0, \beta)$ collective stability with respect to any distributions $\mathbb{P}$ and $\mathbb{Q}$. Definitions 3 and 4 both extend Definition 2, but in different ways; Definition 3 accommodates broader function classes, and Definition 4 accommodates broader instance spaces. Definition 5 is the weakest in the hierarchy, accommodating classes in which only some functions satisfy Definition 4.

As shown in Appendix B, the collective stability of a hypothesis extends to any admissible loss function, meaning a stable predictor will have stable loss. For functionals (i.e., when $N = 1$), such as the average loss, L, we use the term *difference-bounded* (following Kutin, 2002) instead of collective stability. Further, we say that a functional φ is α-*uniformly range-bounded* if, for any $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}^n$, $|\varphi(\mathbf{z}) - \varphi(\mathbf{z}')| \leq \alpha$.

5 STATISTICAL TOOLS

Before presenting our generalization bounds, we review some supporting definitions and introduce a novel moment-generating function inequality for functions of interdependent random variables. We use this later to obtain high-probability bounds on the difference of the expected and empirical risks.

We first introduce a data structure to measure dependence. Let π be a permutation of $[n] \triangleq \{1, 2, \ldots, n\}$, 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 6. 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.

For probability measures $\mathbb{P}$ and $\mathbb{Q}$ on a σ-algebra Σ, recall the standard definition of *total variation distance*,

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

Definition 7. Fix a filtration $\boldsymbol{\pi} \in \Pi(n)$. For $i \in [n]$, $j > i$, $\mathbf{z} \in \mathcal{Z}^{i-1}$ and $z, z' \in \mathcal{Z}$, define the *η-mixing coefficients*[2],

$$\vartheta^{\boldsymbol{\pi}}_{i,j}(\mathbf{z}, z, z') \triangleq \left\| \begin{array}{c} \mathbb{P}\left(\mathbf{Z}_{\pi_i(j:n)} \,|\, \mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z)\right) \\ -\mathbb{P}\left(\mathbf{Z}_{\pi_i(j:n)} \,|\, \mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z')\right) \end{array} \right\|_{\text{TV}} .$$

We use these to define the upper-triangular *dependency matrix* $\boldsymbol{\Theta}^{\boldsymbol{\pi}}_n \in \mathbb{R}^{n \times n}$, with entries

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

Finally, recall the definition of the induced matrix ∞-norm, $\|\boldsymbol{\Theta}^{\boldsymbol{\pi}}_n\|_\infty \triangleq \max_{i \in [n]} \sum_{j=1}^n |\theta^{\boldsymbol{\pi}}_{i,j}|$. Observe that, if $Z_1, \ldots, Z_n$ are mutually independent, then $\boldsymbol{\Theta}^{\boldsymbol{\pi}}_n$ is the identity and $\|\boldsymbol{\Theta}^{\boldsymbol{\pi}}_n\|_\infty = 1$.

We do not assume that $\mathbf{Z}$ corresponds to a temporal process, which is why permuting the order can have such a strong impact on $\|\boldsymbol{\Theta}^{\boldsymbol{\pi}}_n\|_\infty$. In general, given an arbitrary graph topology, $\|\boldsymbol{\Theta}^{\boldsymbol{\pi}}_n\|_\infty$ measures the decay of dependence over graph distance. For example, for a Markov tree process, Kontorovich (2012) orders the variables via a breadth-first traversal from the root; for an Ising model on a lattice, Chazottes et al. (2007) order the variables with a spiraling traversal from the origin. Both these instances use a static permutation, not a filtration. Nonetheless, under suitable contraction or temperature regimes, the authors show that $\|\boldsymbol{\Theta}^{\boldsymbol{\pi}}_n\|_\infty$ is bounded independent of n (i.e., $\|\boldsymbol{\Theta}^{\boldsymbol{\pi}}_n\|_\infty = \mathrm{O}(1)$). 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 Appendix A.4.

With the supporting definitions in mind, we are ready to present our moment-generating function inequality. The proof is provided in Appendix A.2.

Theorem 1. *Let $\mathbf{Z} \triangleq (Z_i)_{i=1}^n$ be random variables with joint distribution $\mathbb{P}$. Let $\varphi : \mathcal{Z}^n \to \mathbb{R}$ be a measurable function that is $(\mathbb{P}, \nu, \beta)$ difference-bounded, and α-uniformly range-bounded. Then, for any $\lambda \in [0, 1]$, there exists a set $\mathcal{B}_\lambda \subseteq \mathcal{Z}^n$ such that $\mathbb{P}\{\mathbf{Z} \in \mathcal{B}_\lambda\} \leq n\nu/\lambda$ and, for any $\tau \in \mathbb{R}$ and $\boldsymbol{\pi} \in \Pi(n)$,*

$$\mathbb{E}\left[e^{\tau(\varphi(\mathbf{Z}) - \mathbb{E}[\varphi(\mathbf{Z})])} \,|\, \mathbf{Z} \notin \mathcal{B}_\lambda\right] \leq \exp\left(\frac{n\tau^2(2\lambda\alpha + \beta)^2 \|\boldsymbol{\Theta}^{\boldsymbol{\pi}}_n\|_\infty^2}{8}\right).$$

Some implications of this result, including novel concentration inequalities (which may be of interest outside of this context), are discussed in Appendix A.3.

6 PAC-BAYES BOUNDS

We now present two new PAC-Bayes generalization bounds using the non-uniform definitions of collective stability from Section 4. The so-called "explicit" bounds we present, while not as tight as some "implicit" bounds, are arguably more interpretable, and are easily obtained using our martingale-based concentration inequalities. Proofs are provided in Appendix B, so here we provide only a high-level sketch.

Let $\hat{\mathbf{Z}} \triangleq ((Z_i^{(l)})_{i=1}^n)_{l=1}^m$ denote a training set of m structured examples, distributed according to $\mathbb{P}^m$. We define a function $\Phi(h, \hat{\mathbf{Z}}) \triangleq \overline{L}(h) - L(h, \hat{\mathbf{Z}})$. Then, for some set $\mathcal{B}_\mathcal{H} \subseteq \mathcal{H}$ of "bad" hypotheses, we let

$$\Phi'(h, \hat{\mathbf{Z}}) \triangleq \begin{cases} \Phi(h, \hat{\mathbf{Z}}) & \text{if } h \notin \mathcal{B}_\mathcal{H} \\ 0 & \text{otherwise} \end{cases} . \tag{4}$$

Observe that

$$\begin{aligned} \overline{L}(\mathbb{Q}) - L(\mathbb{Q}, \hat{\mathbf{Z}}) &= \mathop{\mathbb{E}}_{h \sim \mathbb{Q}}\left[\Phi(h, \hat{\mathbf{Z}})\right] \\ &= \mathbb{Q}\{h \in \mathcal{B}_\mathcal{H}\} \mathop{\mathbb{E}}_{h \sim \mathbb{Q}}\left[\Phi(h, \hat{\mathbf{Z}}) \,|\, h \in \mathcal{B}_\mathcal{H}\right] \\ &\quad + \mathop{\mathbb{E}}_{h \sim \mathbb{Q}}\left[\Phi'(h, \hat{\mathbf{Z}})\right]. \end{aligned}$$

[2] The η-mixing coefficients were introduce by Kontorovich and Ramanan (2008), and are related to the *maximal coupling coefficients* used by Chazottes et al. (2007).

Further, for any a free parameter $u \in \mathbb{R}$, and any prior and posterior distributions, $\mathbb{H}$ and $\mathbb{Q}$, on $\mathcal{H}$, we have via Donsker and Varadhan's *change of measure* inequality (see Appendix B.1) that

$$\begin{aligned}\mathop{\mathbb{E}}_{h\sim\mathbb{Q}}\left[\Phi'(h,\hat{\mathbf{Z}})\right] &= \frac{1}{u}\mathop{\mathbb{E}}_{h\sim\mathbb{Q}}\left[u\,\Phi'(h,\hat{\mathbf{Z}})\right] \\ &\leq \frac{1}{u}\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H}) + \ln\mathop{\mathbb{E}}_{h\sim\mathbb{H}}\left[e^{u\Phi'(h,\hat{\mathbf{Z}})}\right]\right).\end{aligned}$$

Combining these expressions and applying Markov's inequality, we have, with probability at least $1-\delta$,

$$\begin{aligned}&\overline{L}(\mathbb{Q}) - L(\mathbb{Q},\hat{\mathbf{Z}}) \\ &\quad\leq \mathbb{Q}\{h\in\mathcal{B}_{\mathcal{H}}\}\mathop{\mathbb{E}}_{h\sim\mathbb{Q}}\left[\Phi(h,\hat{\mathbf{Z}})\,|\,h\in\mathcal{B}_{\mathcal{H}}\right] \\ &\qquad+\frac{1}{u}\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H}) + \ln\mathop{\mathbb{E}}_{h\sim\mathbb{H}}\mathop{\mathbb{E}}_{\hat{\mathbf{Z}}\sim\mathbb{P}^m}\left[\frac{1}{\delta}\,e^{u\Phi'(h,\hat{\mathbf{Z}})}\right]\right).\end{aligned}$$

We can then upper-bound $\mathbb{E}_{\hat{\mathbf{Z}}\sim\mathbb{P}^m}\left[e^{u\Phi'(h,\hat{\mathbf{Z}})}\right]$, using Theorem 1, and optimize u. However, if we optimize u for a *particular* posterior $\mathbb{Q}$, the bound might not hold for *all* posteriors simultaneously. We therefore adopt a technique due to Seldin et al. (2012) in which we discretize the space of u and assign each posterior to a value that *approximately* optimizes the bound. Using the union bound to upper-bound the probability that the bound fails for some discrete value of u, we ensure that the bound holds for all posteriors simultaneously with high probability.

To isolate the collective stability of the hypothesis class, our bounds are stated in terms of the following properties of the loss function.

Definition 8. We say that a loss function $\ell : \mathcal{Y}\times\hat{\mathcal{Y}} \to \mathbb{R}_+$ is (M,Λ)-*admissible* if:

1. ℓ is M-uniformly range-bounded;
2. for all $y\in\mathcal{Y}$ and $\hat{y},\hat{y}'\in\hat{\mathcal{Y}}$,

$$|\ell(y,\hat{y}) - \ell(y,\hat{y}')| \leq \Lambda\,\|\hat{y}-\hat{y}'\|_1\,.$$

6.1 $\mathbb{Q}$ Collective Stability Bounds

In the following theorem, we use $(\mathbb{Q},\eta,\beta)$ collective stability to obtain a new PAC-Bayes bound. This is a weaker requirement than the uniform condition used by London et al. (2013), in that it allows the hypothesis class to contain a subset $\mathcal{B}_{\mathcal{H}}$ with "bad" collective stability—that is, Equation 3 does not hold for some desired β. Provided the posterior places suitably low measure, η, on this set, we obtain the same asymptotic convergence rate as the uniform case.

Theorem 2. *Fix any $m\geq 1$, $n\geq 1$, $\boldsymbol{\pi}\in\Pi(n)$ and $\delta\in(0,1)$. Let $\mathcal{H}$ denote a class of hypotheses, ℓ an (M,Λ)-admissible loss function, and $\hat{\mathbf{Z}} \triangleq ((Z_i^{(l)})_{i=1}^n)_{l=1}^m$ a training set. For any prior $\mathbb{H}$ on $\mathcal{H}$, with probability at least $1-\delta$ over realizations of $\hat{\mathbf{Z}}$, the following holds simultaneously for all posteriors $\mathbb{Q}$ such that $\mathcal{H}$ has $(\mathbb{Q},\eta,\beta)$ collective stability:*

$$\begin{aligned}&\overline{L}(\mathbb{Q}) - L(\mathbb{Q},\hat{\mathbf{Z}}) \leq \\ &\quad\eta M + \frac{2(M+\Lambda\beta)\,\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\|_\infty}{\sqrt{2mn}}\sqrt{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H}) + \ln\frac{2}{\delta}}.\end{aligned} \quad (5)$$

Suppose $\mathcal{H}$ has $\left(\mathbb{Q}, \mathrm{O}\left((mn)^{-1/2}\right), \mathrm{O}(1)\right)$ collective stability and $D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H}) = \mathrm{O}\left(\log(mn)\right)$. If the data distribution is weakly dependent, with $\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\|_\infty = \mathrm{O}(1)$, then Equation 5 decreases with both m and n. This decays much faster than bounds that ignore the intra-example dependence when each structured example is large and the number of examples is small. Even for $m=1$, Equation 5 goes to zero as n increases, meaning one can generalize from a single, large example.

Theorem 2 is easily extended to classes with uniform collective stability (see Section 7.1), since $\eta=0$, making it strictly more general than London et al. (2013). We also note that, unlike some previous PAC-Bayes bounds for structured prediction (e.g., Bartlett et al., 2005; McAllester, 2007; Keshet et al., 2011), ours do not have $\ln m$ or $\ln n$ in the numerator—though they may be introduced when bounding the KL divergence.

6.2 $(\mathbb{P},\mathbb{Q})$ Collective Stability Bounds

In our next PAC-Bayes bound, we relax the collective stability requirements even further, to hypothesis classes with $(\mathbb{P},\nu,\mathbb{Q},\eta,\beta)$ collective stability. From Definition 5, this means that there exists a "bad" set of inputs $\mathcal{B}\subseteq\mathcal{Z}^n$ and a "bad" set of hypotheses $\mathcal{B}_{\mathcal{H}}\subseteq\mathcal{H}$. The probability of drawing a "bad" hypothesis $h\in\mathcal{B}_{\mathcal{H}}$, under the posterior $\mathbb{Q}$ on $\mathcal{H}$, is at most η; conditioned on any "good" hypothesis $h\notin\mathcal{B}_{\mathcal{H}}$, the probability of drawing a "bad" input $\mathbf{z}\in\mathcal{B}$, under $\mathbb{P}$, is at most ν. For any "good" inputs $\mathbf{z},\mathbf{z}'\notin\mathcal{B}$, and any "good" hypothesis $h\notin\mathcal{B}_{\mathcal{H}}$, the stability condition (Equation 3) holds for the given β.

Theorem 3. *Fix any $m\geq 1$, $n\geq 1$, $\boldsymbol{\pi}\in\Pi(n)$ and $\delta\in(0,1)$. For $\nu\in[0,1]$, let $\epsilon(\nu)\triangleq 2\nu(mn)^2$. Let $\mathcal{H}$ denote a class of hypotheses, ℓ an (M,Λ)-admissible loss function, and $\hat{\mathbf{Z}} \triangleq ((Z_i^{(l)})_{i=1}^n)_{l=1}^m$ a training set. For any prior $\mathbb{H}$ on $\mathcal{H}$, with probability at least $1-\delta$ over realizations of $\hat{\mathbf{Z}}$, the following holds simultaneously for all posteriors $\mathbb{Q}$ such that $\mathcal{H}$ has $(\mathbb{P},\nu,\mathbb{Q},\eta,\beta)$ collective stability, and $\delta>\epsilon(\nu)$:*

$$\begin{aligned}&\overline{L}(\mathbb{Q}) - L(\mathbb{Q},\hat{\mathbf{Z}}) \leq \\ &\quad\eta M + \frac{4(M+\Lambda\beta)\,\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\|_\infty}{\sqrt{2mn}}\sqrt{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H}) + \ln\frac{2}{\delta-\epsilon(\nu)}}.\end{aligned}$$

Theorem 3 implies generalization when ν is sufficiently small; e.g., $\mathrm{o}\left((mn)^{-K}\right)$, for some order $K > 2$. Assume that the learning algorithm has full knowledge of the hypothesis class and which inputs are "bad" for each hypothesis. The learner can designate a "good" set of hypotheses based on some criteria that map to collective stability. The only unknown is the distribution of the bad inputs, $\mathcal{B}$, which can be estimated from the training data. Given an empirical estimate of $\mathbb{P}\{\mathbf{Z} \in \mathcal{B} \,|\, h \notin \mathcal{B}_{\mathcal{H}}\}$, the learner can construct a posterior that allocates mass to hypotheses proportionally to the mass of their bad inputs, thereby effectively shrinking ν. As the size of the data grows, the learner's estimate of $\mathbb{P}\{\mathbf{Z} \in \mathcal{B} \,|\, h \notin \mathcal{B}_{\mathcal{H}}\}$ improves, allowing it to reduce ν accordingly. We plan to explore this strategy of learning and posterior construction in future work.

7 EXAMPLES

In this section, we derive generalization bounds for a generic collective classification problem. To represent multiclass label assignment for $k \geq 2$ labels, we use the standard basis vectors, wherein each $y \in \mathcal{Y}$ has exactly one nonzero entry, set to 1, whose ordinal corresponds to a label; similarly, the predictor outputs a nonnegative vector, $\hat{y} \in \hat{\mathcal{Y}} \subseteq \mathbb{R}^k_+$, wherein each dimension indicates a score for a particular label. We measure multiclass prediction error using a *margin loss*,

$$\ell_\gamma(y, \hat{y}) \triangleq \mathbb{1}\left\{ \left(\langle y, \hat{y} \rangle - \max_{y' \in \mathcal{Y}: y \neq y'} \langle y', \hat{y} \rangle \right) \leq \gamma \right\},$$

for some $\gamma \geq 0$. Thus, an error is incurred whenever the score of the true label does not exceed a margin of γ over any competing label. Note that ℓ_0 is equivalent to the standard 0-1 loss.

The bounds presented in this section are *derandomized*, in that the loss is stated in terms of a deterministic predictor. We use the PAC-Bayes framework as an analytic tool. Our motivation for this decision is that derandomized bounds offer greater insight in practice, where one typically uses a deterministic predictor. The derandomized bounds are easily rerandomized by a simple modification of the proof technique. Proofs for this section are provided in Appendix D.

7.1 Strongly Convex TSMs

Certain classes of TSMs satisfy the condition of uniform collective stability; in particular, TSMs whose inference objectives are *strongly convex.* (See Appendix C.1 for our precise definition of strong convexity, which we specialize for the 1-norm.) In this subsection, we apply our PAC-Bayes bounds to an instance of this class, and obtain generalization bounds that decay faster than previous results.

Consider a TSM with a convex search space $\mathcal{S}$, weights $\mathbf{w}$, features $\mathbf{f}$ and regularizer Ψ. Let

$$\phi_{\mathbf{w}}(\mathbf{x}, \mathbf{s}) \triangleq -\langle \mathbf{w}, \mathbf{f}(\mathbf{x}, \mathbf{s}) \rangle\,,$$

and note that ϕ is convex in $\mathcal{S}$ if either (a) the features are linear, or (b) the features are concave in $\mathcal{S}$ and the weights are nonnegative (such as in a hinge-loss MRF (Bach et al., 2013)). Assuming ϕ is convex in $\mathcal{S}$, if one further assumes that Ψ is κ-strongly convex, then it is readily verified that the negative energy, $-E_{\mathbf{w}}(\mathbf{x}, \mathbf{s}) = \phi_{\mathbf{w}}(\mathbf{x}, \mathbf{s}) + \Psi(\mathbf{s})$, is at least κ-strongly convex in $\mathcal{S}$. Using this fact, London et al. (2013) proved an upper bound on the uniform collective stability of strongly-convex, bounded TSMs.

Definition 9. Denote by $\mathcal{H}^{\text{SC}}_{\mathcal{T}}$ a class of strongly convex TSMs with bounded features, where:

1. $\mathcal{S}$ is a convex set, ϕ is convex in $\mathcal{S}$ and $\exists \kappa > 0$ such that Ψ is κ-strongly convex;
2. $\exists b \geq 1$ such that $\forall t \in \mathcal{T}$, $\|f_t(\cdot, \cdot)\|_b \leq 1$;
3. Γ has induced 1-norm $\|\Gamma\|_1 \leq 1$.

Definition 10. Denote by $\mathcal{H}^{\text{SC}}_{\mathcal{T},R,\kappa} \subset \mathcal{H}^{\text{SC}}_{\mathcal{T}}$ a class of κ-strongly convex, totally bounded TSMs, where:

1. Ψ is κ-strongly convex;
2. $\exists a, b \geq 1 : 1/a + 1/b = 1$ such that $\|\mathbf{w}\|_a \leq R$ and $\forall t \in \mathcal{T}$, $\|f_t(\cdot, \cdot)\|_b \leq 1$.

Theorem 4. *Fix a graph $G \triangleq (\mathcal{V}, \mathcal{E})$ on n nodes. For a set of clique templates $\mathcal{T}$, let*

$$C_G \triangleq \max_{i \in \mathcal{V}} \sum_{t \in \mathcal{T}} \sum_{c \in t(G)} \mathbb{1}\{i \in c\}$$

denote the maximum number of groundings involving any node in G. Then, any $h \in \mathcal{H}^{\text{SC}}_{\mathcal{T}}$ has $(2\sqrt{\|\mathbf{w}\|_a C_G / \kappa})$-uniform collective stability.

Corollary 1. *The class $\mathcal{H}^{\text{SC}}_{\mathcal{T},R,\kappa}$ has $(2\sqrt{R C_G / \kappa})$-uniform collective stability.*

The proof (in Appendix C.2) leverages the strong convexity of the inference objective and the bounded norm properties. For graphs with *bounded degree* (i.e., the maximum degree is independent of n), it can be shown C_G is upper-bounded by a constant. This is further improved when $|\mathcal{T}| = \mathrm{O}(1)$. An important special case is a *pairwise* TSM, in which $\mathcal{T}$ contains only the unary and pairwise templates.

In our first example, we apply Theorem 2 to a subclass of $\mathcal{H}^{\text{SC}}_{\mathcal{T},R,\kappa}$ for approximate marginal inference. An example of this class is the "convexified" Bethe approximation (Wainwright, 2006).

Definition 11. Denote by $\mathcal{H}^{\text{PAM}}_{R,\kappa} \subset \mathcal{H}^{\text{SC}}_{\mathcal{T},R,\kappa}$ a class of pairwise TSMs that perform approximate marginal inference, where:

1. $\mathcal{S}$ is the local marginal polytope; Ψ is a κ-strongly convex surrogate for the negative entropy;
2. $\mathcal{T}$ contains the unary and pairwise templates;
3. $\|\mathbf{w}\|_\infty \leq R$ and $\forall t \in \mathcal{T}$, $\|f_t(\cdot,\cdot)\|_1 \leq 1$.

Theorem 5. *Fix any $m \geq 1$, $n \geq 1$, $\boldsymbol{\pi} \in \Pi(n)$, $\delta \in (0,1)$ and $\gamma > 0$. Also, fix a graph G on n nodes, with maximum degree $\Delta_G = \mathrm{O}(1)$. Then, with probability at least $1-\delta$ over realizations of $\hat{\mathbf{Z}} \triangleq ((Z_i^{(l)})_{i=1}^n)_{l=1}^m$, the following holds simultaneously for all $h \in \mathcal{H}_{R,\kappa}^{\text{PAM}}$:*

$$\overline{L}^0(h) - L^\gamma(h, \hat{\mathbf{Z}}) \leq \frac{2\left\|\mathbf{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty}{\sqrt{2mn}} \left(1 + \frac{6}{\gamma}\sqrt{\frac{R(\Delta_G+1)}{\kappa}}\right) \times \sqrt{d \ln\left(\frac{18Rn(\Delta_G+2)}{\kappa\gamma^2}\right) + \ln\frac{2}{\delta}}.$$

Since the model is templated, and the templates are bounded, it is reasonable to assume that d and R do not grow with n. Thus, the effective convergence rate is $\mathrm{O}\left(\left\|\mathbf{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty \sqrt{\frac{\ln n}{mn}}\right)$. This is an improvement over London et al.'s uniform collective stability risk bounds (2013) in that it avoids the $\ln m$ term in the numerator.

7.2 Variable-Convexity TSMs

We now consider an interesting new class of pairwise TSMs that have *variable convexity*. This example highlights the benefits of using $(\mathbb{Q}, \eta, \beta)$ collective stability instead of uniform collective stability.

Definition 12. Denote by $\mathcal{H}^{\text{PVC}} \subset \mathcal{H}_{\mathcal{T}}^{\text{SC}}$ a class of pairwise TSMs with variable convexity, where:

1. Ψ is 1-strongly convex, $\kappa > 0$ is a parameter and
$$E_{\mathbf{w},\kappa}(\mathbf{x}, \mathbf{s}) \triangleq \langle \mathbf{w}, f(\mathbf{x}, \mathbf{y}) \rangle - \kappa \Psi(\mathbf{s});$$
2. $\mathcal{T}$ contains the unary and pairwise templates;
3. $\forall t \in \mathcal{T}$, $\|f_t(\cdot,\cdot)\|_1 \leq 1$.

Learning a TSM from $\mathcal{H}^{\text{PVC}}$ involves learning a weight vector $\mathbf{w} \in \mathbb{R}^d$ and convexity parameter $\kappa > 0$.

The first thing to note is that this class does not place any restrictions on the norm of $\mathbf{w}$. Secondly, this class contains hypotheses for which $-E$ has arbitrarily low convexity, as $\kappa \to 0$. Thus, the convexity parameter facilitates a continuum of inference functions; for example, from (approximate) marginal inference to (approximate) MAP inference. This smoothing between marginal and MAP inference has been explored by a number of authors (e.g., Hazan and Urtasun, 2010; Meshi et al., 2012). Another interpretation is that κ controls the amount of *hedging*, discounting extreme points in the inference optimization.

We use Theorem 2 to derive a risk bound for $\mathcal{H}^{\text{PVC}}$.

Theorem 6. *Fix any $m \geq 1$, $n \geq 2$, $\boldsymbol{\pi} \in \Pi(n)$, $\delta \in (0,1)$, $\gamma \in (0, \sqrt{n}]$ and G with $\Delta_G = \mathrm{O}(1)$. Then, with probability at least $1-\delta$ over realizations of $\hat{\mathbf{Z}} \triangleq ((Z_i^{(l)})_{i=1}^n)_{l=1}^m$, the following holds simultaneously for all $h \in \mathcal{H}^{\text{PVC}}$ with parameters $(\mathbf{w}, \kappa)$:*

$$\overline{L}^0(h) - L^\gamma(h, \hat{\mathbf{Z}}) \leq \frac{2}{\sqrt{mn}} + \frac{2d}{mn} + \frac{2\left\|\mathbf{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty}{\sqrt{2mn}} \left(1 + \frac{6}{\gamma}\sqrt{\left(\frac{\|\mathbf{w}\|_\infty}{\kappa} + 1\right)(\Delta_G + 1)}\right) \times \sqrt{d \ln\left(\frac{9n(\Delta_G+2)}{\gamma^2}\sqrt{\ln(mn)}\right) + \frac{\|\mathbf{w}\|_2^2}{2\kappa^2} + \ln\frac{2}{\delta}}.$$

Unlike Theorem 5—which uses a uniform upper bound for $\|\mathbf{w}\|$ and a prescribed convexity κ—Theorem 6 is stated in terms of the parameters of given (learned) hypothesis, with no such restrictions. This makes the bound closer to actual learning practices. Moreover, it implies a learning objective that minimizes the empirical margin loss, $L^\gamma(h, \hat{\mathbf{Z}})$, while also controlling the complexity and stability by minimizing $\|\mathbf{w}\|/\kappa$. Optimizing κ effectively learns the amount of hedging, such as the tradeoff between uniform and peaked marginals. Thus, Theorem 6 yields a new approach to learning structured predictors.

8 CONCLUSION

We have shown that $\tilde{\mathrm{O}}\left(1/\sqrt{mn}\right)$ generalization is indeed possible *without* requiring uniform collective stability. We derived two new PAC-Bayes bounds, based on probabilistic notions of collective stability, and illustrated how they yield generalization bounds for a broad class of structured predictors. These bounds suggest a novel learning objective that optimizes collective stability in addition to minimizing empirical risk. In future work, we plan to design learning algorithms based on these insights.

Acknowledgements

This paper is dedicated to the memory of our friend and coauthor, Ben Taskar. Support for this work was provided by NSF CAREER grants 0746930 and 1054215, NSF grant IIS1218488, and IARPA via 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.

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.

S. Bach, B. Huang, B. London, and L. Getoor. Hinge-loss Markov random fields: Convex inference for structured prediction. In *Uncertainty in Artificial Intelligence*, 2013.

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.

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. 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.

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

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

J. Keshet, D. McAllester, and T. Hazan. PAC-Bayesian approach for minimization of phoneme error rate. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2224–2227, 2011.

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. 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.

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

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. McDonald, C. Shalizi, and M. Schervish. Generalization error bounds for stationary autoregressive models. arXiv:1103.0942, 2011.

O. Meshi, A. Globerson, and T. Jaakkola. Convergence rate analysis of MAP coordinate minimization algorithms. In *Neural Information Processing Systems*, 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.

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 Peter 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.

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.

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