Collective Stability in Structured Prediction: Generalization from One Example

Ben London[*] BLONDON@CS.UMD.EDU
Bert Huang[*] BERT@CS.UMD.EDU
Ben Taskar[†] TASKAR@CS.WASHINGTON.EDU
Lise Getoor[*] GETOOR@CS.UMD.EDU

[*] University of Maryland, College Park, MD 20742 USA
[†] University of Washington, Seattle, WA 98195 USA

Abstract

Structured predictors enable joint inference over multiple interdependent output variables. These models are often trained on a small number of examples with large internal structure. Existing distribution-free generalization bounds do not guarantee generalization in this setting, though this contradicts a large body of empirical evidence from computer vision, natural language processing, social networks and other fields. In this paper, we identify a set of natural conditions—weak dependence, hypothesis complexity and a new measure, *collective stability*—that are sufficient for generalization from even a single example, without imposing an explicit generative model of the data. We then demonstrate that the complexity and stability conditions are satisfied by a broad class of models, including marginal inference in templated graphical models. We thus obtain uniform convergence rates that can decrease significantly faster than previous bounds, particularly when each structured example is sufficiently large and the number of training examples is constant, even one.

1. Introduction

Structured prediction is the task of joint reasoning over multiple interdependent output variables. In practice, structured models are often trained on a small set of examples, each containing many dependent variables. In network analysis, training data can come from a single, massive, connected network (Taskar et al., 2002; Richardson & Domingos, 2006; Sen et al., 2008); in computer vision, object classifiers are often trained on a handful of large outdoor scenes (Munoz et al., 2009); in cross-document coreference resolution, the training data may be a single, large corpus (Singh et al., 2010). Existing generalization bounds for structured prediction cannot guarantee generalization in these settings. In contrast, intuition, empirical results (Jensen et al., 2004; Tsochantaridis et al., 2005) and recent statistical consistency analysis (Xiang & Neville, 2011) suggest that generalization is possible if the single or few examples are large enough, provided they have reasonable internal correlation decay and the models have suitably controlled capacity. In this paper, we present new generalization bounds for structured prediction that explicitly consider both the number and size of structured examples, such that even one example can guarantee generalization if certain sufficient conditions hold. Among these conditions is a new measure we refer to as *collective stability*, which parameterizes the sensitivity of structured predictors to small changes in input data. Collective stability enables finer control over the smoothness of the generalization error w.r.t. single-variable perturbations, which is nontrivial when the predictions are interdependent.

In the structured prediction literature (Taskar et al., 2004; McAllester, 2007), current distribution-free generalization bounds scale as $\mathrm{O}(\sqrt{\ln(mn)/m})$, where m is the number of examples, and n the size of each structure. For a fixed m, this term increases with n. Some assumptions on the distribution are evidently needed to turn the size of each example into an advantage. We thus adopt a finer grained analysis in which we view each structure as a set of dependent variables, but without making parametric assumptions about the dis-

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

tribution. Leveraging recent results in the concentration of dependent random variables (Chazottes et al., 2007; Kontorovich & Ramanan, 2008), we show that, if the data exhibits weak dependence within each structure, and the hypothesis class has suitable collective stability, then the empirical error estimate from a single structured example should uniformly converge to its mean as n (or m) grows. Under suitable weak dependence conditions, the effect of dependence does not affect the convergence rate.

Our specific contributions are as follows. We derive new generalization bounds for structured prediction, identifying two properties of the hypothesis class—Rademacher complexity, and collective stability—as sufficient conditions. We demonstrate that these conditions are attainable by a broad class of structured predictors, which we refer to as *templated structured models* (TSMs). TSMs subsume many graphical models used in practice. Inference with TSMs is a minimization of a convex objective, consisting of templated feature functions and a regularization term. In particular, we focus on TSMs whose objectives are *strongly* convex, examples of which include (approximate) marginal inference and some continuous or relaxed MAP inference. By exploiting the strong convexity of the inference, and the condition of bounded weight and feature norms resulting from templating, we prove that strongly convex TSMs have constant uniform collective stability. Further, using a novel covering argument, we show that the space of strongly convex inference functions can be ϵ-covered by a set whose size is polynomial (rather than exponential) in n—a result that is of independent interest. Using this, we prove that the Rademacher complexity of strongly convex TSMs asymptotically decreases to zero as n (or m) grows. We are thus able to prove $\mathrm{O}(\sqrt{(\ln n)/(mn)})$ generalization bounds for structured prediction, which decay significantly faster than previous bounds when m is constant; even for a single structured example, the empirical error uniformly converges to the true error as the size of the structure increases.

2. Preliminaries

In our learning framework, we are given a set of n dependent random variables $\mathbf{Z} \triangleq \{Z_i\}_{i=1}^n$, where each Z_i takes values in a measure space $\mathcal{Z}$. We define $\mathcal{Z}$ as the Cartesian product of a domain $\mathcal{X}$ and a codomain $\mathcal{Y}$, so Z_i can be expressed as two random variables (X_i, Y_i), taking values in $\mathcal{X}$ and $\mathcal{Y}$ respectively. We quantify the dependence within $\mathbf{Z}$ in Section 4.

The learner aims to predict $\mathbf{Y}$ given $\mathbf{X}$. To do so, it learns a discriminative hypothesis h from a specified class $\mathcal{H} \subseteq \{h : \mathcal{X}^n \to \hat{\mathcal{Y}}^n\}$, where $\hat{\mathcal{Y}} \subseteq \mathbb{R}^k$ is not necessarily the same as $\mathcal{Y}$. For example, in multiclass classification, each dimension of $\hat{\mathcal{Y}}$ could indicate a real-valued confidence in a specific label. We use $h_i(\mathbf{x})$ to denote the i^{th} prediction $\hat{y}_i$ and $h_i^j(\mathbf{x})$ to denote its j^{th} value $\hat{y}_i^j$. Similarly, we use $h^j(\mathbf{x})$ to denote the prediction vector limited to the j^{th} value of each prediction, i.e., $(\hat{y}_1^j, \ldots, \hat{y}_n^j)$, and let $\mathcal{H}^j \triangleq \{h^j : h \in \mathcal{H}\}$.

We are particularly interested in hypothesis classes that perform joint reasoning over all variables simultaneously. This means that changes to any single input variable may affect the output predictions on others. In Section 6, we discuss examples of such models.

In the canonical learning framework of structured prediction, we are given m independent draws from $\mathbb{P}(\mathbf{Z})$—i.e., m *realizations* of $\mathbf{Z}$. Such is the case in many computer vision tasks, in which the training set consists of multiple images of identical dimensions. Note that any number of realizations can be represented as a single realization of a set of mn random variables, whose distribution factorizes over the (identical) marginal distributions of m subsets of size n. We are interested in the scenario in which n is much larger than m, or where n grows and $m = \mathrm{O}(1)$. For example, in network analysis, it is not unusual to learn from a single structured example. Thus, unless otherwise specified, we assume that the training data consists of a single realization.

Let $\ell : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R}$ be a *loss* function. Define the *empirical* loss of a hypothesis h w.r.t. $\mathbf{Z}$ as $L(h, \mathbf{Z}) \triangleq \frac{1}{n}\sum_{i=1}^n \ell(Y_i, h_i(\mathbf{X}))$. The quantity of interest is the *expected* loss $\overline{L}(h) \triangleq \mathbb{E}[L(h, \mathbf{Z})]$ (also known as the *risk*) over realizations of $\mathbf{Z}$, which corresponds to the error h will incur on future predictions. In the event that $\mathbf{Z}$ represents m realizations of the same set of variables, the risk is the expected loss on a single realization; using the previous computer vision example, the test example would be a single image.

3. Related Work

Our analysis departs from that of traditional structured prediction (Taskar et al., 2004; McAllester, 2007) in that we explicitly consider the dependence between the variables in each example, similar to learning with interdependent data. There is a large body of work in learning *local* (i.e., non-structured) predictors from interdependent data. Usunier et al. (2006) analyze learning with a specific type of dependence in which each variable depends on a finite number of variables and is unconditionally independent of all others. Representing the dependence as a graph, the authors use graph

coloring to derive Rademacher-based risk bounds for local predictors. Ralaivola et al. (2010) use a similar technique to derive PAC-Bayes bounds for this setting. Mohri & Rostamizadeh develop risk bounds for ϕ- and β-mixing time series data, using both Rademacher complexity (2009) and algorithmic stability (2010), though the hypotheses they consider predict each time step independently. Other authors (e.g., McDonald et al., 2011; Alquier & Wintenburger, 2012) provide risk bounds for autoregressive forecasting models, in which the prediction at time t depends on a moving window of previous observations. We study a more general setting that allows hypotheses to perform joint inference over arbitrarily structured examples.

Xiang & Neville (2011) examine the asymptotic properties of collective inference in the *one-network* learning paradigm, in which data is generated by an infinite Markov random field, with certain labels observed during training. They show that maximum likelihood and pseudo-likelihood estimation are *asymptotically consistent*, which suggests that non-asymptotic convergence is possible.

Our condition of uniform collective stability is a form of global Lipschitz stability. Wainwright (2006) analyze the Lipschitz stability of approximate marginal inference w.r.t. changes in the model parameters, using this to bound the error of an inconsistent estimator w.r.t. the Bayes optimum. Similarly, (Honorio, 2011) show that the log-likelihood of many graphical models is also Lipschitz w.r.t. the parameters. To our knowledge, ours is the first work to identify the connection between predictive stability, w.r.t. changing inputs, and the generalization of structured prediction, particularly in the limited example setting.

4. Concentration Inequality

In this section, we review some supporting definitions and a theorem on the concentration of dependent random variables. We use this theorem to show that the generalization error uniformly converges to zero as the size of the structure grows.

For probability measures $\mathbb{P}$ and $\mathbb{Q}$ on a σ-algebra Σ over a sample space Ω, define the *total variation distance* as

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

We fix an ordering of the variables $\mathbf{Z} \triangleq \{Z_i\}_{i=1}^n$ and define a measure of dependence w.r.t. the ordering. For $i \in [n]$, $j > i$, let

$$\eta_{i,j} \triangleq \sup \|\mathbb{P}(\mathbf{Z}_{j:n} \,|\, \mathbf{z}_{1:i-1}, z_i) - \mathbb{P}(\mathbf{Z}_{j:n} \,|\, \mathbf{z}_{1:i-1}, z_i')\|_{\text{TV}} ,$$

where the supremum runs over all $\mathbf{z}_{1:i-1} \in \mathcal{Z}^{i-1}$ and $z_i, z_i' \in \mathcal{Z}$. Define the upper triangular *dependency matrix* $\mathbf{\Theta}_n^\pi \in \mathbb{R}^{n \times n}$ as

$$\theta_{i,j}^\pi \triangleq \begin{cases} 1 & \text{for } i = j, \\ \eta_{i,j} & \text{for } i < j, \\ 0 & \text{for } i > j. \end{cases}$$

Finally, recall the standard definition of the matrix infinity norm, $\|\mathbf{\Theta}_n^\pi\|_\infty \triangleq \max_{i \in [n]} \sum_{j=1}^n \left|\theta_{i,j}^\pi\right|$. With these definitions, we recall the following bound, due to Kontorovich & Ramanan (2008, Theorem 1.1).

Theorem 1. *Let $f : \mathcal{Z}^n \to \mathbb{R}$ be a measurable function for which there exists a constant c such that, for any $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}^n$ that differ only at a single coordinate, $|f(\mathbf{z}) - f(\mathbf{z}')| \le c/n$. Then for any $\epsilon > 0$,*

$$\mathbb{P}\left\{f(\mathbf{Z}) - \mathbb{E}[f(\mathbf{Z})] \ge \epsilon\right\} \le \exp\left(\frac{-2n\epsilon^2}{c^2 \left\|\mathbf{\Theta}_n^\pi\right\|_\infty^2}\right).$$

Like (Mohri & Rostamizadeh, 2010, Theorem 8), Theorem 1 achieves a slight improvement over the original by using a general form of McDiarmid's inequality instead of Azuma's inequality. A short proof is given in the supplementary materials.

It can be shown that the above bound holds for any ordering of $\mathbf{Z}$, which has a strong impact on the growth of $\|\mathbf{\Theta}_n^\pi\|_\infty$ w.r.t. n. Note that we do not assume that $\mathbf{Z}$ is a temporal process. In general, given a graph topology and an ordering of the vertices, $\|\mathbf{\Theta}_n^\pi\|_\infty$ measures the decay of dependence over graph distance. For instance, for Markov a tree process, Kontorovich (2007) 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. In both of these instances, under suitable contraction or temperature regimes, the authors show that $\|\mathbf{\Theta}_n^\pi\|_\infty$ is bounded independent of n (i.e., $\|\mathbf{\Theta}_n^\pi\|_\infty = \mathrm{O}(1)$). We posit that the same holds for any graph with bounded degree when the mixing coefficients exhibit geometric decay.

5. Generalization Bounds

In this section, we derive *probably approximately correct* (PAC) generalization bounds for structured prediction and identify the associated sufficient conditions.

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 & Elisseeff, 2002). We apply this concept to arbitrary vector-valued functions.

Definition 1. Let $\mathcal{F}$ be a class of vector-valued functions from $\mathcal{Z}^n$ to $\mathbb{R}^N$, where N does not necessarily equal n. We say that $\mathcal{F}$ has *uniform collective stability* β if, for any two inputs $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}^n$ that differ only at a single coordinate, $\sup_{f\in\mathcal{F}} \|f(\mathbf{z}) - f(\mathbf{z}')\|_1 \leq \beta$.

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

In addition to stability, we use the *Rademacher complexity* to measure hypothesis complexity. We adapt the canonical definition from Bartlett & Mendelson (2003) for structured prediction and remove the assumption that $Z_1, \ldots, Z_n$ are i.i.d.

Definition 2. Let $\mathbf{Z} \triangleq \{Z_i\}_{i=1}^n$ be a set of random variables. Let $\{\sigma_i\}_{i=1}^n$ be a set of independent, uniformly distributed, $\{\pm 1\}$-valued random variables, referred to as *Rademacher variables*. Define the *empirical Rademacher complexity* of $\mathcal{F} \subseteq \{f : \mathcal{Z}^n \to \mathbb{R}^n\}$ as

$$\mathfrak{R}(\mathcal{F}, \mathbf{Z}) \triangleq \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f_i(\mathbf{Z}) \,|\, \mathbf{Z}\right].$$

Define the *Rademacher complexity* of $\mathcal{F}$, w.r.t. realizations of $\mathbf{Z}$, as $\overline{\mathfrak{R}}_n(\mathcal{F}) \triangleq \mathbb{E}[\mathfrak{R}(\mathcal{F}, \mathbf{Z})]$.

To accommodate a variety of loss functions, we require the following generic properties.

Definition 3. A loss function $\ell : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R}$ is (M, λ)-*admissible* if there exist constants $M < \infty$ and $\lambda < \infty$ such that: (1) for any $y, y' \in \mathcal{Y}$ and $\hat{y} \in \hat{\mathcal{Y}}$, $|\ell(y, \hat{y}) - \ell(y', \hat{y})| \leq M$; (2) for any $y \in \mathcal{Y}$ and $\hat{y}, \hat{y}' \in \hat{\mathcal{Y}}$, $|\ell(y, \hat{y}) - \ell(y, \hat{y}')| \leq \lambda \|\hat{y} - \hat{y}'\|_1$.

We provide an example of an admissible loss function in Section 7. We now state our main result.

Theorem 2. *Let $\mathcal{H} \subseteq \{h : \mathcal{X}^n \to \hat{\mathcal{Y}}^n\}$ be a class of hypotheses, where $\hat{\mathcal{Y}} \subseteq \mathbb{R}^k$, and suppose $\mathcal{H}$ has uniform collective stability β. Let ℓ be a loss function that is (M, λ)-admissible. Then, for any $n \geq 1$ and $\delta \in (0, 1)$, with probability $\geq 1 - \delta$ over realizations of $\mathbf{Z} \triangleq \{Z_i\}_{i=1}^n$, every $h \in \mathcal{H}$ satisfies*

$$\overline{L}(h) \leq L(h, \mathbf{Z}) + 2\lambda \sum_{j=1}^k \overline{\mathfrak{R}}_n(\mathcal{H}^j) + (M + \lambda\beta) \|\boldsymbol{\Theta}_n^\pi\|_\infty \sqrt{\frac{\ln(1/\delta)}{2n}}. \quad (1)$$

We can directly apply Theorem 2 to the setting in which the training set is m i.i.d. structured examples.

Corollary 1. *Let $\mathbf{Z}' \triangleq \{\mathbf{Z}'_l\}_{l=1}^m$ be a set of random variables representing m realizations of $\mathbf{Z}$. If $h(\mathbf{X}') = (h(\mathbf{X}'_l))_{l=1}^m$, then, for any $m \geq 1$, $n \geq 1$ and $\delta \in (0, 1)$, with probability $\geq 1 - \delta$ over realizations of $\mathbf{Z}'$, every $h \in \mathcal{H}$ satisfies*

$$\overline{L}(h) \leq L(h, \mathbf{Z}') + 2\lambda \sum_{j=1}^k \overline{\mathfrak{R}}_{mn}(\mathcal{H}^j) + (M + \lambda\beta) \|\boldsymbol{\Theta}_n^\pi\|_\infty \sqrt{\frac{\ln(1/\delta)}{2mn}}. \quad (2)$$

We prove Theorem 2 via a series of technical lemmas. The first lemma establishes the uniform collective stability of ℓ in terms of the stability of $\mathcal{H}$. In the interest of space, we defer all intermediate proofs to the supplemental materials.

Lemma 1. *If a hypothesis class $\mathcal{H}$ has uniform collective stability β, and a loss function ℓ is (M, λ)-admissible, then $\ell \circ \mathcal{H}$ has uniform collective stability $(M + \lambda\beta)$.*

For the following, let $\mathcal{F}$ be an arbitrary class of functions from $\mathcal{Z}^n$ to $\mathbb{R}^n$. For any particular $f \in \mathcal{F}$, let

$$F(\mathbf{Z}) \triangleq \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{Z}) \;;\; \overline{F} \triangleq \mathbb{E}[F(\mathbf{Z})] \;;$$

$$\Phi(\mathcal{F}, \mathbf{Z}) \triangleq \sup_{f\in\mathcal{F}} \overline{F} - F(\mathbf{Z}) \;;\; \overline{\Phi}(\mathcal{F}) \triangleq \mathbb{E}[\Phi(\mathcal{F}, \mathbf{Z})].$$

The following lemma shows that, with high probability, Φ uniformly converges to its expected value. The key insight is that uniform collective stability enables concentration.

Lemma 2. *If $\mathcal{F}$ has uniform collective stability β, then, for any $n \geq 1$ and $\delta \in (0, 1)$, with probability $\geq 1 - \delta$ over realizations of $\mathbf{Z}$,*

$$\Phi(\mathcal{F}, \mathbf{Z}) \leq \overline{\Phi}(\mathcal{F}) + \beta \|\boldsymbol{\Theta}_n^\pi\|_\infty \sqrt{\frac{\ln(1/\delta)}{2n}}.$$

Using a symmetry argument, we now upper-bound $\overline{\Phi}(\mathcal{F})$ by the Rademacher complexity $\overline{\mathfrak{R}}_n(\mathcal{F})$. Unlike typical symmetry arguments, our analysis does not require the individual random variables to be mutually independent; all we require is that the train and test sets be identically distributed. (This analysis also holds for local inference.)

Lemma 3. *For any $n \geq 1$, $\overline{\Phi}(\mathcal{F}) \leq 2\overline{\mathfrak{R}}_n(\mathcal{F})$.*

Lemma 4. *Let $\mathcal{H} \subseteq \{h : \mathcal{X}^n \to \hat{\mathcal{Y}}^n\}$, with $\hat{\mathcal{Y}} \subseteq \mathbb{R}^k$. If ℓ is (M, λ)-admissible, then*

$$\overline{\mathfrak{R}}_n(\ell \circ \mathcal{H}) \leq \lambda \sum_{j=1}^k \overline{\mathfrak{R}}_n(\mathcal{H}^j).$$

We are now ready to prove Theorem 2. We start with the simple observation that $\overline{L}(h) \leq L(h, \mathbf{Z}) + \Phi(\mathcal{F}, \mathbf{Z})$, where we let $\mathcal{F} \triangleq \ell \circ \mathcal{H}$. By Lemma 1, $\mathcal{F}$ has uniform collective stability $(M + \lambda\beta)$. We therefore have from Lemma 2 that, with probability $\geq 1 - \delta$,

$$\overline{L}(h) \leq L(h, \mathbf{Z}) + \overline{\Phi}(\mathcal{F}) + (M + \lambda\beta) \left\| \mathbf{\Theta}_n^{\pi} \right\|_{\infty} \sqrt{\frac{\ln(1/\delta)}{2n}}.$$

To bound $\overline{\Phi}(\mathcal{F})$, we apply Lemmas 3 and 4 to $\overline{\Phi}(\mathcal{F})$, which establishes Equation 1.

The bounds in this section imply sufficient conditions for generalization of structured learning and explicitly consider both the number and size of structured examples. When these conditions are met, Equation 2 can be tighter than existing bounds when the number of structured examples is few or one.

6. Templated Structured Models

In this section, we discuss a broad class of hypotheses that satisfy $\beta = \mathrm{O}(1)$ and $\overline{\mathfrak{R}}_n(\mathcal{H}) = \mathrm{O}(\sqrt{\ln(n)/n})$ (or $\mathrm{O}(\sqrt{\ln(n)/(mn)})$). Before doing so, we briefly review some related models that are typically used for structured prediction.

One such model is a *Markov random field* (MRF). An MRF is defined by a graph $G \triangleq (\mathcal{V}, \mathcal{E})$, a set of cliques $\mathcal{Q}$, a set of *feature functions* $\{f_q : \mathcal{Z}^{|q|} \to \mathbb{R}^{d_q}\}_{q \in \mathcal{Q}}$ and a set of weights $\{w_q \in \mathbb{R}^{d_q}\}_{q \in \mathcal{Q}}$, where $|q|$ is the size of clique q and d_q is the number of possible assignments. For now, assume that a feature function outputs the *overcomplete representation* of its clique's assignment; i.e., $f_q^j(\mathbf{z}_q) = 1$ if $\mathbf{z}_q$ is in the j^{th} state. One typically denotes the weights by a single vector $\mathbf{w} \triangleq (w_q)_{q \in \mathcal{Q}}$ and the features by a single function $\mathbf{f}(\mathbf{z}) \triangleq (f_q(\mathbf{z}_q))_{q \in \mathcal{Q}}$, both of which have (output) length $d \triangleq \sum_{q \in \mathcal{Q}} d_q$. An MRF defines a distribution $P_{\mathbf{w}}$ over a set of random variables $\mathbf{Z} \triangleq \{Z_i\}_{i \in \mathcal{V}}$ as

$$P_{\mathbf{w}}(\mathbf{Z} = \mathbf{z}) \triangleq \frac{1}{\Pi(\mathbf{w})} \exp\left(\langle \mathbf{w}, \mathbf{f}(\mathbf{z}) \rangle\right),$$

where $\Pi(\mathbf{w})$ is a normalizing constant. If each Z_i is actually a tuple (X_i, Y_i) of input-output pairs, then a *conditional random field* (CRF) defines a distribution,

$$P_{\mathbf{w}}(\mathbf{Y} = \mathbf{y} \,|\, \mathbf{X} = \mathbf{x}) \triangleq \frac{1}{\Pi(\mathbf{w}, \mathbf{x})} \exp(\langle \mathbf{w}, \mathbf{f}(\mathbf{x}, \mathbf{y}) \rangle),$$

There are two canonical inference problems in M/CRFs: *marginal* inference, which estimates the marginal distribution of each clique, and *maximum a posteriori* (MAP) inference, which estimates the most likely global assignment. We denote the marginals by a vector $\boldsymbol{\mu} \triangleq \mathbb{E}_{\mathbf{w}}[\mathbf{f}(\mathbf{Z})]$, where μ_q^j indicates the marginal probability that clique q is in state $j \in [d_q]$. It is well known that $\boldsymbol{\mu} = \arg\max_{\boldsymbol{\mu}' \in \mathcal{M}} \langle \mathbf{w}, \boldsymbol{\mu}' \rangle + H(\boldsymbol{\mu}')$, where $\mathcal{M} \triangleq \{\boldsymbol{\mu}' \in \mathbb{R}^d \,|\, \exists \mathbf{w}' : \boldsymbol{\mu}' = \mathbb{E}_{\mathbf{w}'}[\mathbf{f}(\mathbf{Z})]\}$ is the *marginal polytope* and $H(\boldsymbol{\mu}')$ is the entropy of the distribution whose marginals are $\boldsymbol{\mu}'$. This identity is commonly used to perform approximate marginal inference, by relaxing the marginal polytope and using a convex surrogate for $-H$ (Wainwright, 2006). Since we are primarily concerned with the marginals of individual variables, we assume that the higher-order marginals are discarded. For a given observation $\mathbf{x} \in \mathcal{X}^n$, the MAP state is the $\mathbf{y} \in \mathcal{Y}^n$ that maximizes $P_{\mathbf{w}}(\mathbf{Y} = \mathbf{y} \,|\, \mathbf{X} = \mathbf{x})$. (Assume some deterministic, consistent tie-breaking mechanism.) Since the logarithm is strictly increasing and $\Pi(\mathbf{w}, \mathbf{x})$ is constant, this is equivalent to $\arg\max_{\mathbf{y} \in \mathcal{Y}^n} \langle \mathbf{w}, \mathbf{f}(\mathbf{x}, \mathbf{y}) \rangle$. Thus far, we have assumed that $\mathcal{Z}$ is a finite set, though these inference methods have equivalent forms when $\mathcal{Z}$ is continuous, such as in Gaussian random fields.

A common technique for defining M/CRFs is *templating* (sometimes referred to 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 (e.g., Taskar et al., 2002; Neville & Jensen, 2004; Richardson & Domingos, 2006; Broecheler et al., 2010). We will later show that templating also enables uniform convergence, due to the fact that the number of parameters does not grow with the size of the data.

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

Definition 4. For the following, let $\mathcal{A}$ be a convex set and $\hat{\mathcal{Y}} \subseteq \mathbb{R}^k$, for some $k \geq 1$. A *templated structured model* (TSM) is defined by: 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* $\Psi : \mathcal{A} \to \mathbb{R}$; and a linear *projection* $\Gamma : \mathcal{A} \to \hat{\mathcal{Y}}^n$. Given a graph G, let $t(G)$ denote the set of cliques matching template t. As before, let $\mathbf{w} \triangleq (w_t)_{t \in \mathcal{T}}$ and $\mathbf{f}(\mathbf{x}, \mathbf{a}) \triangleq (\sum_{q \in t(G)} f_t(\mathbf{x}_q, \mathbf{a}_q))_{t \in \mathcal{T}}$, both of which have (output) length $d \triangleq \sum_{t \in \mathcal{T}} d_t$. Define the *energy function* $E_{\mathbf{w}}(\mathbf{x}, \mathbf{a}) \triangleq \phi_{\mathbf{w}}(\mathbf{x}, \mathbf{a}) - \Psi(\mathbf{a})$, where $\phi_{\mathbf{w}}(\mathbf{x}, \mathbf{a}) \triangleq \langle \mathbf{w}, \mathbf{f}(\mathbf{x}, \mathbf{a}) \rangle$. All TSMs must satisfy

a convexity property, where $\phi_{\mathbf{w}}(\mathbf{x}, \mathbf{a})$ is concave in $\mathcal{A}$, and Ψ is convex, implying $-E$ is convex in $\mathcal{A}$.[1] For a given input $\mathbf{x} \in \mathcal{X}^n$, a TSM hypothesis h outputs

$$h(\mathbf{x}) \triangleq \Gamma(\arg\max_{\mathbf{a} \in \mathcal{A}} E_{\mathbf{w}}(\mathbf{x}, \mathbf{a})). \tag{3}$$

We denote by $\mathcal{H}_{\mathcal{T},R,B}$ the class of TSMs that satisfy the following boundedness conditions: $|\mathcal{T}| = \mathrm{O}(1)$; $\sup_{t \in \mathcal{T}} |t| = \mathrm{O}(1)$; and there exist constants $R < \infty$ and $B < \infty$ such that $\forall \mathbf{w} \in \mathcal{H}_{\mathcal{T},R,B}$, $\|\mathbf{w}\|_2 \leq R$ and, $\forall t \in \mathcal{T}$, $\|f_t(\cdot, \cdot)\|_2 \leq B$. Since the model is templated, and the maximum size and number of templates is bounded, it is reasonable to assume that R and B do not grow with n. This might not be the case if one assigned a unique weight to each clique, or if the maximum size or number of templates were unbounded.

We typically assume that the graph, clique templates, feature functions and energy function are given *a priori*. Thus, learning a TSM amounts to learning the template weights.

Though the above representation is very abstract, one can show that inference in TSMs is equivalent to inference in the previous models. To recreate (approximate) marginal inference in a templated CRF, we define $\mathcal{A}$ as the (relaxed) marginal polytope and Ψ as (a convex surrogate for) the negative entropy; each feature function f_t simply returns the overcomplete value of $\mathbf{a}_q$; the projection Γ zeros out the non-singleton marginals and then discards any entries for which $\mathbf{X} \neq \mathbf{x}$ (which are, by definition, zero). Note that the resulting output is of the length nk. We can also recover MAP inference by letting $\Psi(\mathbf{a}) \triangleq 0$ for all $\mathbf{a} \in \mathcal{A}$. This will return an integral solution, but it will not satisfy the conditions necessary for the rest of our analysis.

Suppose we wanted to perform inference on $\hat{\mathcal{Y}}^n$ directly; in other words, let $\mathcal{A} \triangleq \hat{\mathcal{Y}}^n$ and optimize over $E(\mathbf{x}, \hat{\mathbf{y}})$, using some convex regularizer. In this case, f_t is an arbitrary linear or concave function of $(\mathbf{x}_q, \hat{\mathbf{y}}_q)$, and Γ is the identity. For models of continuous domains (e.g., Broecheler et al., 2010), this is equivalent to MAP inference with a convex prior. For discrete domains, if $\hat{\mathcal{Y}}$ is the simplex $\{\hat{y} \in [0,1]^k : \|\hat{y}\|_1 = 1\}$, then the optimal $\hat{\mathbf{y}}$ can be considered a relaxation of the true MAP state, where each $\hat{y}_i^j$ indicates a score for variable Y_i being in state j.

In the following subsections, we show that certain TSMs satisfy the sufficient conditions for generalization given in Theorem 2; specifically, TSMs whose inference objectives are *strongly convex*.

[1] This is satisfied when the features are linear, or when they are concave in $\mathcal{A}$ and the weights are nonnegative.

Definition 5. A function $\varphi : \mathcal{A} \to \mathbb{R}$ is κ-strongly convex (w.r.t. the 1-norm) if $\mathcal{A}$ is a convex set and, for any $a, a' \in \mathcal{A}$ and $\tau \in [0,1]$,

$$\tau(1-\tau)\frac{\kappa}{2}\|a - a'\|_1^2 + \varphi(\tau a + (1-\tau)a') \leq \tau\varphi(a) + (1-\tau)\varphi(a').$$

The negative energy function, $-E$, is, by design, convex; however, to ensure *strong* convexity, we consider a class of TSMs whose regularizers are strongly convex. This includes the previous example of (approximate) marginal inference, since the negative entropy, as well as many surrogates, are strongly convex (Wainwright, 2006; Shalev-Schwartz, 2007). If $\phi_{\mathbf{w}}(\mathbf{x}, \mathbf{a})$ is concave in $\mathcal{A}$, and $\Psi(\mathbf{a})$ is κ-strongly convex, then $-E_{\mathbf{w}}(\mathbf{x}, \mathbf{a}) = \Psi(\mathbf{a}) - \phi_{\mathbf{w}}(\mathbf{x}, \mathbf{a})$ is at least κ-strongly convex in $\mathcal{A}$. Thus, we now show that TSMs with κ-strongly convex regularizers, which we denote by $\mathcal{H}_{\mathcal{T},R,B,\kappa}$, have good collective stability and low Rademacher complexity.

6.1. Collective Stability

To prove the collective stability of TSMs, we begin with two technical lemmas.

Lemma 5. *Let $\varphi : \mathcal{A} \to \mathbb{R}$ be κ-strongly convex, and let $\dot{a} \triangleq \arg\min_{a \in \mathcal{A}} \varphi(a)$. Then, for any $a \in \mathcal{A}$*

$$\|a - \dot{a}\|_1^2 \leq \frac{2}{\kappa}(\varphi(a) - \varphi(\dot{a})).$$

Lemma 6. *Let $\varphi : \Omega \times \mathcal{A} \to \mathbb{R}$ be κ-strongly convex in $\mathcal{A}$. If, for any $\omega, \omega' \in \Omega$ and $a \in \mathcal{A}$, $|\varphi(\omega, a) - \varphi(\omega', a)| \leq \lambda$, then*

$$\left\| \arg\min_{a \in \mathcal{A}} \varphi(\omega, a) - \arg\min_{a' \in \mathcal{A}} \varphi(\omega', a') \right\|_1 \leq \sqrt{2\lambda/\kappa}.$$

Lemma 6 implies that the maximum of the energy function has uniform collective stability if the negative energy function is strongly convex. To apply this requires a type of Lipschitz stability, which we show in the following lemma.

Lemma 7. *For a graph G and a set of clique templates $\mathcal{T}$, let $Q_i \triangleq \sum_{t \in \mathcal{T}} \sum_{q \in t(G)} \mathbb{1}[i \in q]$ denote the number of cliques involving node i, and let $Q_G \triangleq \max_{i \in \mathcal{V}} Q_i$. Then, for any G, $\mathbf{w} \in \mathcal{H}_{\mathcal{T},R,B}$, $\mathbf{a} \in \mathcal{A}$ and $\mathbf{x}, \mathbf{x}' \in \mathcal{X}^n$ that differ at a single coordinate i,*

$$|E_{\mathbf{w}}(\mathbf{x}, \mathbf{a}) - E_{\mathbf{w}}(\mathbf{x}', \mathbf{a})| \leq 2RBQ_G.$$

The proof (given in the supplemental materials) leverages the boundedness property of $\mathcal{H}_{\mathcal{T},R,B}$. We now bound the collective stability of TSMs.

Theorem 3. *For any G, with Q_G defined in Lemma 7, $\mathcal{H}_{\mathcal{T},R,B,\kappa}$ has uniform collective stability $(2\sqrt{RBQ_G/\kappa})$.*

Proof. As discussed, if Ψ is κ-strongly convex, then $-E$ is at least κ-strongly convex in $\mathcal{A}$. Fix any $\mathbf{w} \in \mathcal{H}_{\mathcal{T},R,B,\kappa}$ and $\mathbf{x}, \mathbf{x}' \in \mathcal{X}^n$, and let $\mathbf{a}, \mathbf{a}' \in \mathcal{A}$ denote their respective maximizers of $E_{\mathbf{w}}$. Via the additive property of linear transformations,

$$\begin{aligned}\|h(\mathbf{x}) - h(\mathbf{x}')\|_1 &= \|\Gamma(\mathbf{a}) - \Gamma(\mathbf{a}')\|_1 = \|\Gamma(\mathbf{a} - \mathbf{a}')\|_1 \\ &\leq \|\Gamma\| \, \|\mathbf{a} - \mathbf{a}'\|_1 \leq 2\sqrt{RBQ_G/\kappa},\end{aligned}$$

where the last inequality follows from Lemmas 6 and 7, and the fact that a projection has norm 1. □

For G with bounded degree, one can show that $Q_G = \mathrm{O}(1)$; thus, $\mathcal{H}_{\mathcal{T},R,B,\kappa}$ has uniform collective stability $\mathrm{O}(1/\sqrt{\kappa})$.

Recall, when $\kappa = 0$, $\mathcal{Y}$ is discrete and $\mathcal{A}$ is the marginal polytope, that maximizing E will return an integral MAP state. In this case, when $-E$ is not strongly convex, the remaining boundedness conditions are insufficient for good collective stability. In fact, it can be shown by counterexample that discrete MAP inference, under nontrivial conditions, has uniform collective stability $\mathrm{O}(n)$. If, for example, $\mathcal{T}$ contains the unary and pairwise templates, one can always select inputs $\mathbf{x}$ and weights $\mathbf{w}$ such that changing a single input x_i causes every coordinate in the prediction to change. Thus, in order to obtain useful collective stability for discrete MAP inference, further restrictions on the domain or hypothesis class are necessary.

6.2. Rademacher Complexity

We now bound the Rademacher complexity of $\mathcal{H}_{\mathcal{T},R,B,\kappa}$. We do so by first bounding the *covering number* of $\mathcal{H}_{\mathcal{T},R,B,\kappa}$.

Definition 6. Let $\mathcal{S} \subseteq \mathbb{R}^d$ be a set of vectors in $\mathbb{R}^d$, for some $d \geq 1$. We say that a set $\mathcal{C} \subseteq \mathbb{R}^d$ is an ϵ-cover of $\mathcal{S}$ under a norm $\|\cdot\|$ if, for any $\mathbf{s} \in \mathcal{S}$, there exists a $\mathbf{c} \in \mathcal{C}$ such that $\|\mathbf{s} - \mathbf{c}\| \leq \epsilon$.

Definition 7. Let $\mathcal{F}$ be a class of functions from $\mathcal{X}^n$ to $\mathbb{R}^N$. For a given $n \geq 1$ and $\mathbf{x} \in \mathcal{X}^n$, let $\mathcal{S}(\mathbf{x}) \triangleq \{f(\mathbf{x}) : f \in \mathcal{F}\}$. The empirical *covering number* $\mathcal{N}_p(\epsilon, \mathcal{F}, \mathbf{x})$ is the cardinality of the minimal $\mathcal{C} \subseteq \mathbb{R}^N$ that ϵ-covers $\mathcal{S}(\mathbf{x})$ under the *normalized* p-norm, $N^{-1/p}\|\cdot\|_p$. With a slight abuse of notation, let $\mathcal{N}_p(\epsilon, \mathcal{F}, n) \triangleq \sup_{\mathbf{x} \in \mathcal{X}^n} \mathcal{N}_p(\epsilon, \mathcal{F}, \mathbf{x})$.

Lemma 8. *The hypercube $[0, \Lambda]^d$ admits an ϵ-cover, under the 2-norm, of cardinality $\lceil(\sqrt{d}\Lambda/(2\epsilon))^d\rceil$.*

Theorem 4. *For a graph G, let $T_G \triangleq \sup_{t \in \mathcal{T}} |t(G)|$. Then, for any $n \geq 1$, G and $\epsilon > 0$,*

$$\mathcal{N}_2(\epsilon, \mathcal{H}_{\mathcal{T},R,B,\kappa}, n) \leq \left\lceil \left(\frac{\sqrt{d}RBT_G}{\kappa n k \epsilon^2} \right)^d \right\rceil, \qquad (4)$$

where $k \triangleq |h_i(\cdot)|$ is the cardinality of a prediction.

Proof. Fix any $\mathbf{x} \in \mathcal{X}^n$, and let $\mathcal{S}(\mathbf{x}) \triangleq \{\hat{\mathbf{y}} = h(\mathbf{x}) : h \in \mathcal{H}_{\mathcal{T},R,B,\kappa}\}$. We will show that there exists a subset $\mathcal{C} \subseteq \mathcal{S}(\mathbf{x})$ that is an ϵ-cover of $\mathcal{S}(\mathbf{x})$, under the normalized 2-norm. Fix any $\hat{\mathbf{y}}, \hat{\mathbf{y}}' \in \mathcal{S}(\mathbf{x})$, and let $\mathbf{a}, \mathbf{a}' \in \mathcal{A}$ be vectors such that $\Gamma(\mathbf{a}) = \hat{\mathbf{y}}$ and $\Gamma(\mathbf{a}') = \hat{\mathbf{y}}'$. Let $\mathbf{w}, \mathbf{w}' \in \mathcal{H}_{\mathcal{T},R,B,\kappa}$ be weight vectors such that $\mathbf{a}$ and $\mathbf{a}'$ maximize $E_{\mathbf{w}}$ and $E_{\mathbf{w}'}$ respectively. Recall that $|h_i(\cdot)| = k$, and so every output of Γ is of length nk. Since Γ is a projection with norm 1, we have that

$$\frac{1}{\sqrt{nk}} \|\hat{\mathbf{y}} - \hat{\mathbf{y}}'\|_2 \leq \frac{1}{\sqrt{nk}} \|\mathbf{a} - \mathbf{a}'\|_1 .$$

Further, since $-E$ is κ-strongly convex in $\mathcal{A}$, and every $\mathbf{w} \in \mathcal{H}_{\mathcal{T},R,B,\kappa}$ satisfies $\|\mathbf{w}\|_2 \leq R$, using Lemma 5, one can show that

$$\|\mathbf{a} - \mathbf{a}'\|_1^2 \leq \frac{2R}{\kappa} \|\mathbf{f}(\mathbf{x}, \mathbf{a}) - \mathbf{f}(\mathbf{x}, \mathbf{a}')\|_2 .$$

Now, consider the set $\mathcal{S}'(\mathbf{x}) \triangleq \{\mathbf{f}(\mathbf{x}, \mathbf{a}) : \hat{\mathbf{y}} \in \mathcal{S}(\mathbf{x}), \hat{\mathbf{y}} = \Gamma(\mathbf{a})\}$, which is convex. Since the norm of any feature function f_t is uniformly upper-bounded by B, we have that $\|\mathbf{f}(\mathbf{x}, \mathbf{a})\|_\infty \leq BT_G$ for all $\mathbf{a} \in \mathcal{A}$. Therefore, the features are contained within the hypercube $[0, BT_G]^d$. By Lemma 8, this hypercube admits an ϵ'-cover, under the 2-norm, of cardinality $\left\lceil(\sqrt{d}BT_G/(2\epsilon'))^d\right\rceil$. By extension, there exists an ϵ'-cover $\mathcal{C}' \subseteq \mathbb{R}^d$ of $\mathcal{S}'(\mathbf{x})$ of at most the same size; and since $\mathcal{S}'(\mathbf{x})$ is convex, there exists such a $\mathcal{C}'$ where every point in $\mathcal{C}'$ is also in $\mathcal{S}'(\mathbf{x})$. By the definition of $\mathcal{S}'(\mathbf{x})$, this means there is a corresponding $\hat{\mathbf{y}}' \in \mathcal{C} \subseteq \mathcal{S}(\mathbf{x})$.

We now have that, for any $\hat{\mathbf{y}} \in \mathcal{S}(\mathbf{x})$, there exists a $\hat{\mathbf{y}}' \in \mathcal{C}$ such that

$$\begin{aligned}\frac{1}{\sqrt{nk}} \|\hat{\mathbf{y}} - \hat{\mathbf{y}}'\|_2 &\leq \sqrt{\frac{2R}{\kappa nk} \|\mathbf{f}(\mathbf{x}, \mathbf{a}) - \mathbf{f}(\mathbf{x}, \mathbf{a}')\|_2} \\ &\leq \sqrt{\frac{2R\epsilon'}{\kappa nk}} \triangleq \epsilon.\end{aligned}$$

Solving for ϵ', we obtain the cardinality of $\mathcal{C}'$ (hence, $\mathcal{C}$) needed to obtain an ϵ-cover of $\mathcal{S}(\mathbf{x})$, under the normalized 2-norm, which upper-bounds the empirical covering number. Since this holds uniformly for any $\mathbf{x} \in \mathcal{X}^n$, it also upper-bounds the (non-empirical) covering number. □

Theorem 5. *For any $n \geq 1$, G and $j \in [k]$,*

$$\overline{\mathfrak{R}}_n(\mathcal{H}^j_{\mathcal{T},R,B,\kappa}) \leq \left(1 + \sqrt{2 \ln \xi}\right) / \sqrt{n}, \qquad (5)$$

where $\xi \triangleq \lceil (\sqrt{d} RBT_G/\kappa)^d \rceil$.

The proof follows from direct application of the discretization theorem (see supplemental materials). (Though Dudley's Theorem is potentially tighter, we prefer the Discretization Theorem for its simplicity.) Note that, when the dataset contains m realizations of $\mathbf{Z}$, this bound becomes $\left(1 + \sqrt{2 \ln \xi}\right) / (\sqrt{mn})$.

7. Application

In this section, we use the results from previous sections to derive risk bounds for collective classification. (We consider collective regression in the supplemental material.) In collective classification, the goal is to predict a categorical variable from a set of k labels. We represent this space using an overcomplete representation in which each $y \in \mathcal{Y}$ is a binary vector with exactly one nonzero entry, whose ordinal corresponds to the label. We assume that predictors output a real-valued vector, $\hat{y} = h_i(\mathbf{x})$, where each dimension indicates a score for a particular label, so the predicted label is the one with the highest confidence, i.e., $\arg\max_{y' \in \mathcal{Y}} \langle y', \hat{y} \rangle$. We therefore want the correct label to have the uniquely highest confidence. Since the multiclass 0-1 loss $\ell_{\mathbb{1}}$ is not admissible, we define a *margin-based* loss function that is admissible and dominates the 0-1 loss:

$$\ell_\gamma(y, \hat{y}) \triangleq r_\gamma(\langle y, \hat{y} \rangle - \max_{y' \in \mathcal{Y}: y \neq y'} \langle y', \hat{y} \rangle),$$

where $\gamma \geq 0$ and r_γ is the *ramp function* (defined in the supplemental materials).

Lemma 9. *The margin loss ℓ_γ is $(1, 1/\gamma)$-admissible.*

This allows us to bound the 0-1 classification risk $L^{\mathbb{1}}$ for (approximate) marginal inference in TSMs. For the following, we assume that a graph G has been determined *a priori*, based on the structure of the problem, and that G has maximum degree $\Delta_G = \mathrm{O}(1)$. For notational convenience, let $\Delta^1_G \triangleq \Delta_G + 1$. To make our bounds concrete, we will assume that the clique templates $\mathcal{T}$ consist only of the unary and pairwise templates. We therefore have that Q_G (from Section 6.1) is upper-bounded by Δ^1_G, and T_G (from Section 6.2) is upper-bounded $n\Delta^1_G$.

Theorem 6. *Let $\mathcal{H}_{\mathcal{T},R,B,\kappa}$ be a class of TSM classifiers that output the (approximate) marginals, where $\mathcal{A}$ is the (relaxed) marginal polytope and Ψ is (a κ-strongly convex surrogate for) the negative entropy. Then, for any $n \geq 1$ and $\delta \in (0, 1)$, there exists a constant $C < \infty$ such that, with probability $\geq 1 - \delta$ over realizations of $\mathbf{Z}$, every $h \in \mathcal{H}_{\mathcal{T},R,B,\kappa}$ satisfies*

$$\overline{L}^{\mathbb{1}}(h) \leq L^\gamma(h, \mathbf{Z}) + \frac{2kC}{\gamma} \sqrt{\frac{d \ln(\sqrt{d} Rn \Delta^1_G/\kappa)}{n}} + \left(1 + \frac{2}{\gamma}\sqrt{\frac{R\Delta^1_G}{\kappa}}\right) \|\boldsymbol{\Theta}^\pi_n\|_\infty \sqrt{\frac{\ln(1/\delta)}{2n}}. \qquad (6)$$

Proof. Since ℓ_γ dominates $\ell_{\mathbb{1}}$, it follows that the expected margin loss $\overline{L}^\gamma$ dominates the expected 0-1 loss $\overline{L}^{\mathbb{1}}$. Therefore, substituting $(M, \lambda) = (1, 1/\gamma)$ into Theorem 2, we obtain risk bounds for $\overline{L}^{\mathbb{1}}$ using the empirical margin loss L^γ. The rest of the proof follows from Theorems 3 and 5, where we have substituted upper bounds for Q_G and T_G, and leveraged the fact that $B \leq 1$ in (approximate) marginal inference. □

If $-\Psi$ is the entropy, then $\kappa = 1$, since the negative entropy is 1-strongly convex (Shalev-Schwartz, 2007).

8. Discussion

In this paper, we derive generalization bounds for structured prediction in the setting where the training set consists of few large, structured examples—possibly even one. We identify three sufficient conditions: weak dependence, low model complexity and a new measure that is specific to structured prediction, collective stability. We show that a broad class of structured models satisfy the complexity and stability conditions through templating and strongly convex regularization of the inference objective. Under suitable weak dependence conditions, when $\|\boldsymbol{\Theta}^\pi_n\|_\infty$ exhibits sub-linear growth in n, this leads to $\mathrm{O}(\sqrt{(\ln n)/(mn)})$ uniform convergence, which is significantly sharper than previous bounds for structured prediction.

Acknowledgments

This work was partially supported 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 IARPA, DoI/NBC, or the U.S. Government.

References

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

Bartlett, P. and Mendelson, S. Rademacher and gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2003.

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

Broecheler, M., Mihalkova, L., and Getoor, L. Probabilistic similarity logic. In *Uncertainty in Artificial Intelligence*, 2010.

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

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

Jensen, D., Neville, J., and Gallagher, B. Why collective inference improves relational classification. In *Knowledge Discovery and Data Mining*, 2004.

Kontorovich, L. *Measure Concentration of Strongly Mixing Processes with Applications*. PhD thesis, Carnegie Mellon University, 2007.

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

McAllester, D. Generalization bounds and consistency for structured labeling. In *Predicting Structured Data*. MIT Press, 2007.

McDonald, D., Shazili, C., and Schervish, M. Risk bounds for time series without strong mixing. arXiv:1106.0730, 2011.

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

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

Munoz, D., Bagnell, J., Vandapel, N., and Hebert, M. Contextual classification with functional max-margin Markov networks. In *Computer Vision and Pattern Recognition*, 2009.

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

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

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

Sen, P., Namata, G., Bilgic, M., Getoor, L., Gallagher, B., and Eliassi-Rad, T. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008.

Shalev-Schwartz, S. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University of Jerusalem, 2007.

Singh, S., Wick, M., and McCallum, A. Distantly labeling data for large scale cross-document coreference. arXiv:1005.4298, 2010.

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

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

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

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

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

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