ABSTRACT

Title of dissertation: ON THE STABILITY OF STRUCTURED PREDICTION

Benjamin Alexei London,
Doctor of Philosophy, 2015

Dissertation directed by: Professor Lise Getoor
Department of Computer Science

Many important applications of artificial intelligence—such as image segmentation, part-of-speech tagging and network classification—are framed as multiple, interdependent prediction tasks. These structured prediction problems are typically modeled using some form of joint inference over the outputs, to exploit the relational dependencies. Joint reasoning can significantly improve predictive accuracy, but it introduces a complication in the analysis of structured models: the *stability* of inference. In optimizations involving multiple interdependent variables, such as joint inference, a small change to the input or parameters could induce drastic changes in the solution.

In this dissertation, I investigate the impact of stability in structured prediction. I explore two topics, connected by the stability of inference. First, I provide generalization bounds for learning from a limited number of examples with large internal structure. The effective learning rate can be significantly sharper than rates given in related work. Under certain conditions on the data distribution and stability of the predictor, the bounds decrease with both the number of examples and the size of each example, meaning one could potentially learn from a single giant example. Secondly, I investigate the benefits of learning with strongly convex variational inference. Using the duality between strong convexity and stability, I demonstrate, both theoretically and empirically, that learning with a strongly convex free energy can result in significantly more accurate marginal probabilities. One consequence of this work is a new technique that "strongly convexifies" many free energies used in practice. These two seemingly unrelated threads are tied by the idea that stable inference leads to lower error, particularly in the limited example setting, thereby demonstrating that inference stability is of critical importance to the study and practice of structured prediction.

ON THE STABILITY OF STRUCTURED PREDICTION

by

Benjamin Alexei London

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2015

Advisory Committee:

Prof. Lise Getoor, Univ. of Maryland	Chair/Advisor
Prof. Philip Resnik, Univ. of Maryland	Dean's Representative
Prof. Hal Daumé, III, Univ. of Maryland	Committee Member
Prof. Larry S. Davis, Univ. of Maryland	Committee Member
Prof. Dan Roth, Univ. of Illinois at Urbana-Champaign	Committee Member

To my parents, Fredda and Jack,

and my wife, Martha.

Acknowledgments

On its surface, graduate school is a solitary experience, filled with long days (and nights) of study, reflection and exploration. Yet, no one finishes graduate school on one's own. Behind every degree conferred, there is a long list of people who supported the recipient. I find myself at this moment, about to finish a doctorate, thanks to the incredible support of family, friends and colleagues.

First and foremost, I owe the deepest gratitude to my advisor, Lise Getoor. When I applied to the University of Maryland, I knew next to nothing about graphical models and structured prediction, and even less about conducting research. She took a chance on me and shaped me into the researcher I am today. She taught me how to ask the right questions, how to present with passion and clarity, and most importantly, that all lists should contain at least three items. She allowed me the freedom to work on projects that excited me, even when they did not perfectly align with the goals of her research group. She promoted my work, introduced me to the right contacts and collaborators, and fostered my career as a scientist. Her guidance and support over the past five years has been invaluable, and I am forever grateful for it.

I would also like to recognize my collaborator and unofficial second advisor, Bert Huang. Bert was integral to all of my key publications; without his technical prowess and linguistic agility, these works might have never gotten off the ground, let alone been published. Despite his busy schedule and workload—being involved in all projects within our research group—he was always available to discuss an idea or work through a problem. He patiently endured many such meetings, making a dedicated effort to follow my meandering train of thought as I scribbled some inscrutable (and likely flawed) proof on the board. He was my sounding board, editor and friend. I cannot thank him enough, and I wish him great success in his future as a professor at Virginia Tech.

For my first two major publications, I had the honor of collaborating with the late Ben Taskar. Ben's ability to immediately understand a problem and provide insightful input was remarkable. He was always thinking several steps ahead of me, and it would sometimes take weeks or months for me to understand the deep connections to related concepts and literature that came to him almost instantaneously. I was devastated to hear of his death, and of the circumstances his family then faced. I feel so fortunate to have worked with him, and I will always treasure the time I had getting to know him.

The summer before my final year of graduate school, I interned at Google Research, in New York. I am deeply grateful for this opportunity, and to my host, Afshin Rostamizadeh, for his mentorship. Even after my internship had ended, he was quick to offer help and career counseling. I also thank David Weiss, the WebTables team, and my fellow interns, who made the experience that much more enjoyable.

Over the years, I have had the pleasure of collaborating with a number of excellent researchers. Those not already mentioned include: Stephen H. Bach, James Foulds, Sameh Khamis, Galileo Namata, Jay Pujara and Theodoros "Theo" Rekatsinas. I hope I will have the opportunity to work with them again some day.

I thank my committee members—Hal Daumé, III, Larry S. Davis, Philip Resnik and Dan Roth—for their time and excellent feedback.

I thank the administrative staff at the Computer Science department and UMIACS, for helping me navigate the bureaucracy of graduate school; and the technical staff, who maintain our equipment and services.

I thank my fellow LINQS group members—and all graduate students in the Computer Science department—for the many thought-provoking discussions and reading groups over the years, as well as their camaraderie in this shared experience. I am especially grateful for my friendships with Ioana Bercea and Theo Rekatsinas, who provided some much needed moments of levity along the way.

Needless to say, I would not have arrived at this moment without the love and support of my parents. They instilled in me a love of learning, and showed me, by example, that learning is a lifelong journey, in which it is never too late to change course. They will always be my archetypal examples of success, both professionally and in life.

Finally, I dedicate this dissertation to my wife, my partner and my best friend, Martha. Every day, she inspires me to strive to be a better, more thoughtful person—which is partly why I went back to school in the first place. This degree truly is a joint accomplishment: we embarked on this expedition together and, fortunately, landed on the other side, that much stronger and ready to begin the next adventure.

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

Table of Contents

Chapter 1: Introduction

Many important applications of artificial intelligence involve reasoning about multiple interdependent unknowns, which give the problem an inherent internal *structure*. Often, structure can be exploited to improve accuracy or efficiency. A classic example of such a problem is labeling the nodes of a graph, such as a social network or publication database. Connected users tend to share similar traits, and papers typically cite related papers, so a link between two users or documents suggests that they have the same label (Getoor and Taskar, 2007). Structure need not be explicit; it can be imposed on a problem by assigning semantics to proximity. For instance, when segmenting an image, pixels that are within a certain neighborhood of each other are likely to belong to the same object (Anguelov et al., 2005). Moreover, if the goal is scene understanding, one can use proximity of actors to reason about their activities; if one actor is talking, a nearby actor is likely listening (Khamis et al., 2012; London et al., 2013b).

These types of problems and approaches are broadly categorized as *structured prediction*. Structured prediction problems are generally characterized by (learning) a mapping to an exponentially large output space, such as labelings of n points or parses of length-n sentences. The output space is endowed with an implicit or explicit structure, which is represented by a graph. Inference involves searching the output space to find

the "best" assignment (usually, conditioned on some observations), by some measure, or score, defined by the model. It is assumed that the scoring function does not readily decompose over the individual assignments, meaning inference necessitates a *global* optimization (e.g., Pearl, 1988; Roth and Yih, 2005; Daumé, III et al., 2009). Indeed, in the aforementioned examples, the relational dependencies (e.g., social links, citations, proximities) prevent one from decomposing the joint optimization into a set of independent, local optimizations.

Due to the collective nature of inference, perturbing the observations or model parameters, however slightly, may result in drastic changes to the predictions. As a result, a performance measurement (i.e., *loss function*) of joint inference can have high variance. Measuring this effect is the focus of *stability* analysis. In a general sense, an algorithm is stable if small changes to its input induce proportionally small changes in its output. The stability of learning algorithms has been shown to play a key role in how well they generalize from a finite training sample (Bousquet and Elisseeff, 2002). *In this dissertation, I explore the impact of inference stability in learning structured models.*

One insight is that stability is particularly important when learning from examples with large internal structure. In certain applications of structured prediction, each example consists of many interdependent inputs and outputs. For instance, a document may contain thousands of words to be assigned a part-of-speech tag; a digital image may contain millions of pixels to be segmented; and a social network may contain millions of users to be categorized. Obtaining fully annotated examples can be time-consuming and expensive, due to the number of variables. It is therefore common to train a structured predictor using far fewer examples than are used in the unstructured setting. In the extreme

(yet not atypical) case, the training set consists of a single example, with large internal structure. *The central message of my work is that inference stability leads to lower error, which means that one can do more with fewer examples.*

1.1 Contributions

My contributions are grouped into two topics, tied by the idea of stability. The first topic concerns providing better theoretical guarantees for learning from a few large, structured examples. The second topic investigates the benefits of strongly convex variational free energies when learning with approximate inference. The common thread between these seemingly disparate subjects is that the stability of inference influences the learning rate. The first line of research demonstrates how stable inference can improve generalization in the limited data setting; the second line proposes a criteria and method for promoting stable inference, thereby improving the accuracy of the learned model.

The content of this dissertation is derived from research I led, and to which I was the primary contributor. Certain portions have been published elsewhere (see London et al., 2013a, 2014, 2015a).

1.1.1 Generalization Bounds for Learning from Limited Examples

A fundamental question in statistical learning theory is *generalization*; that is, whether the expected error at test time will be reasonably close to the empirical error measured during training. Canonical learning-theoretic results for structured prediction (e.g., Taskar et al., 2004; Bartlett et al., 2005; McAllester, 2007) only guarantee generalization when the

number of training examples is high. Yet, this pessimism contradicts a wealth of experimental results (e.g., Taskar et al., 2002; Tsochantaridis et al., 2005), which indicate that training on a few large examples is sufficient. In Chapters 5 and 6, I address the question of when generalization is possible in this setting. I derive new generalization bounds for structured prediction that are far more optimistic than previous results. When sufficient conditions hold, these bounds guarantee generalization from a few large examples—even just one.

The intuition behind the analysis is motivated by a common practice known alternatively as *templating* or *parameter-tying*. At a high level, templating shares parameters across substructures (e.g., nodes, edges, etc.) with identical local structure. (Templating is explained in detail in Section 2.3.3.) Originally proposed for relational learning as a way of dealing with non-uniformly-structured examples, templating has an additional benefit in that it effectively limits the complexity of the hypothesis class by reducing the number of parameters to be learned. Each instance of a substructure within an example acts as a kind of "micro example" of a template. Since each example may contain many micro examples, it is plausible that generalization could occur from even a single example.

Part of the difficulty when formalizing this intuition is that the micro examples are interdependent. Like all statistical arguments, generalization bounds must show that the empirical error concentrates around the expected error, and analyzing the concentration of functions of dependent random variables is nontrivial. Moreover, inference in a structured predictor is typically formulated as a global optimization over all outputs simultaneously. Due to model-induced dependencies, changes to one input may affect many of the outputs, which affects the loss differently than in binary or multiclass prediction. Thus, this

problem cannot be viewed as simply learning from interdependent data, which has been studied extensively (e.g., Usunier et al., 2006; Mohri and Rostamizadeh, 2010; Ralaivola et al., 2010).

There are therefore two obstacles: the dependence in the data distribution and the dependence induced by the the predictor. I characterize the former dependence using concepts from measure concentration theory (Kontorovich and Ramanan, 2008), and I view the latter dependence as the stability of inference. Chapter 5 reviews my earlier work in this area (London et al., 2013a), which was based on the notion of *collective stability* (see Section 3.2). Collectively stability isolates the stability of the predictions from the stability of the loss function. The results of Chapter 5 use a *uniform* definition of collective stability, which is guaranteed for predictors whose inference objectives are strongly convex. In Chapter 6 (which presents work from London et al., 2015b), I use a more general form of stability (see Section 3.1) that analyzes the loss function directly. This definition of stability accommodates a broader range of loss functions and predictors, and eliminates the previous reliance on strong convexity. Moreover, it supports functions that are *locally* stable over some subset of their domain, and random functions that are stable with high probability.[1]

The primary contributions of Chapters 5 and 6 are generalization bounds for structured prediction that decrease when either the number of examples, m, or the size of each example, n, increase. Under suitable conditions on the data distribution, hypothesis class and loss function, the bounds can be as tight as $\tilde{\mathrm{O}}\left(1/\sqrt{mn}\right)$, which decreases as either

[1] I introduced local, probabilistic versions of collective stability in 2014, though these will not be covered, since they are subsumed by the more general results from 2015b.

m or n increase. This rate is much tighter than previous results, which only guarantee $\tilde{\mathrm{O}}\left(1/\sqrt{m}\right)$. Chapter 5 uses an approach based on the collective stability and covering number of the hypothesis class. Chapter 6 uses PAC-Bayesian analysis, which is particularly well suited for the probabilistic definitions of stability. These latter results apply to any composition of loss function and hypothesis class that satisfies the local stability conditions, which includes a broad range of modeling regimes used in practice. I also propose a novel view of PAC-Bayesian "derandomization," based on the principle of stability, which provides a general proof technique for converting a generalization bound for a randomized structured predictor into a bound for a deterministic structured predictor.

To enable the PAC-Bayesian analysis, I derive a new bound on the moment-generating function of a locally stable functional (see Section 4.2). This result implies a new tail inequality, which is used in the covering number-based analysis. The tightness of these bounds (hence, the generalization bounds) hinges on a measure of the aggregate dependence between the random variables within each example. The bounds are meaningful when the dependence is sub-logarithmic in the number of variables. In Section 4.3, I provide two examples of stochastic processes for which this condition holds. All of these results may be of independent interest for those in the learning theory and measure concentration community.

I apply the generalization bounds to several common models and learning scenarios, including *posterior decoding*, *max-margin Markov networks* and *soft-max* training of *conditional random fields*. To demonstrate the benefit of local stability analysis, I also consider a specific generative process that induces unbounded stability in certain predictors, given certain inputs. These examples suggest several factors to be considered when

modeling structured data, in order to obtain the fast generalization rate: (1) templating is crucial; (2) the norm of the parameters contributes to the stability of inference, and should be controlled via regularization; and (3) limiting local interactions in the model can improve stability, hence, generalization. All of these considerations can be summarized by the classic tension between representational power and overfitting, applied to the structured setting. Most importantly, these examples confirm that generalization from limited training examples is indeed possible for many structured prediction techniques used in practice.

1.1.2 Benefits of Learning with Strongly Convex Variational Inference

Though marginal inference in general graphical models is an intractable problem, many approximations have been proposed using the *variational free energy* (see Section 2.3.2). Much of this research has focused on the convexity of the free energy. When it is convex, convergence to a global minimum is guaranteed. Less attention has been paid to when the free energy is *strongly* convex (i.e., has curvature), and what benefits this offers. In Chapter 7 (which presents work from London et al., 2015a), I show, both theoretically and empirically, that learning with a strongly convex free energy results in more accurate marginal probabilities.

My theoretical analysis is based on the stability of the marginals. The marginals of a log-linear distribution (such as the one described in Section 2.3.1) form the gradient of the *log-partition function* with respect to the *potentials*. Thus, one can characterize the stability of the marginals by the *Lipschitz constant* of the gradient. The Lipschitz gradient

condition (Hiriart-Urruty and Lemaréchal, 2001) is the dual of strong convexity. Using this duality, and the variational form of the log-partition, I show that strongly convex free energies result in more stable marginals. Further, I argue that a simply convex free energy cannot satisfy this stability guarantee. Finally, using the stability of the marginals, I prove an error bound for the marginals of a model that is learned using strongly convex variational inference. The error bound is inversely proportional to the *modulus* of convexity (i.e., amount of curvature) of the free energy, thereby highlighting an important consideration for strongly convex free energies: the modulus should be constant with respect to the size of the graph, $|G|$, particularly when $|G|$ is large relative to the number of examples.

Based on the above insights, I proceed to identify free energies that are strongly convex, and when their respective moduli of convexity are constant with respect to $|G|$. I consider two popular variational methods: tree-reweighted (Wainwright et al., 2005) and counting number (Heskes, 2006) entropies. Using the notion of *contraction*, I give model-dependent conditions under which the negative tree-reweighted entropy is $\Omega(1)$-strongly convex. I then propose new sufficient conditions to characterize the modulus of convexity for counting number entropies. I use this to derive a novel counting number optimization that yields κ-strongly convex free energies, for any $\kappa > 0$, independent of the model parameters. This optimization can "strongly convexify" any entropy approximation that can be expressed via counting numbers, which includes many used in practice (e.g., Bethe and tree-reweighted).

I demonstrate the practical impact of the theory in a set of experiments on challenging grid-structured models. The empirical results suggest that strongly convex free

energies can dramatically improve the quality of marginal inference, and that the counting number optimization reduces the error of learned marginals by over 40%. These findings indicate that having a tunable modulus can offer substantial benefit in practice.

1.2 Related Work

One of the earliest explorations of generalization in structured prediction is by Collins (2001), who developed risk bounds for language parsers using various classical tools, such as the Vapnik-Chervonenkis dimension and margin theory. In Taskar et al.'s (2004) landmark paper on max-margin Markov networks, the authors use covering numbers to derive risk bounds for their proposed class of models. Bartlett et al. (2005) improved this result using PAC-Bayesian analysis.[2] McAllester (2007; 2011) provided a comprehensive PAC-Bayesian study of various structured losses and learning algorithms. Recently, Hazan et al. (2013) proposed a PAC-Bayes bound with a form often attributed to Catoni (2007), which can be minimized directly using gradient descent. Giguère et al. (2013) used PAC-Bayesian analysis to derive risk bounds for the kernel regression approach to structured prediction. In a similar vein as the above literature, yet taking a significantly different approach, Bradley and Guestrin (2012) derived finite sample complexity bounds for learning conditional random fields using the composite likelihood estimator.

All of the above works have approached the problem from the traditional viewpoint, that the generalization error should decrease proportionally to the number of examples. In a previous publication (London et al., 2013a), I proposed the first bounds that decrease

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

with both the number of examples and the size of each example (given suitably weak dependence within each example). I later refined these results using PAC-Bayesian analysis (London et al., 2014). My more recent work (London et al., 2015b) builds upon this foundation to derive similarly optimistic generalization bounds, while accommodating a broader range of loss functions and hypothesis classes.

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

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

The idea of learning from one example is related to the "one-network" learning

paradigm, in which data is generated by a (possibly infinite) random field, with certain labels observed for training. The underlying model is estimated from the partially observed network, and the learned model is used to predict the missing labels, typically with some form of joint inference. Xiang and Neville (2011) examined maximum likelihood and pseudo-likelihood estimation in this setting, proving that are asymptotically consistent. Note that this is a *transductive* setting, in that the network data is fixed (i.e., realized), so the learned hypothesis is not expected to generalize to other network data. In contrast, I analyze *inductive* learning, wherein the model is applied to future draws from a distribution over network data.

Connections between stability and generalization have been explored in various forms. Bousquet and Elisseeff (2002) proposed the stability of a learning algorithm as a tool for analyzing generalization error. This landmark work paved the way for a number of related results (e.g., Kutin and Niyogi, 2002; Elisseeff et al., 2005; Mukherjee et al., 2006; Cortes et al., 2008; Mohri and Rostamizadeh, 2010; Shalev-Shwartz et al., 2010). Taking a significantly different approach, Wainwright (2006) analyzed the stability of marginal probabilities in variational inference, identifying the relationship between stability and strong convexity (similar to my work in London et al., 2013a, 2014, 2015a). He used this result to show that an *inconsistent* estimator, which uses approximate inference during training, can asymptotically yield lower regret (relative to the optimal Bayes least squares estimator) than using the *true* model with approximate inference. In a another different, yet related, work, Honorio (2011) showed that the Bayes error rate of various graphical models is related to the stability of their log-likelihood functions with respect to changes in the model parameters.

The study of convex free energies in approximate inference has a long history. Approaches can be broadly categorized by their approximation of the negative entropy term. Wainwright et al.'s (2005) tree-reweighted approximation decomposes the entropy into a convex combination of tree entropies, each of which is convex. Wainwright (2006) later showed that this approximation is in fact strongly convex, though his lower bound on the modulus decreases as a function of the size of the graph. Another decomposition approach, due to Globerson and Jaakkola (2007), replaces the entropy with a sum of conditional entropies. This approximation is provably convex, but not strongly convex. Heskes (2006) proposed general sufficient conditions, based on counting (or, "over-counting") numbers, to establish the convexity of the Bethe and Kikuchi approximations. This work inspired a wave of research in counting number-based approximations (e.g., Weiss et al., 2007; Hazan and Shashua, 2008; Meltzer et al., 2009; Meshi et al., 2009). Hazan and Shashua (2008) used a slight modification of Heskes's conditions to guarantee strict convexity, which guarantees a unique global minimum, but does not identify a modulus. To my knowledge, the sufficient conditions I identify in Section 7.2.2 are the first to identify when the counting number entropy is strongly convex, with a known modulus.

1.3 Organization

The remainder of this document is organized as follows.

- Chapter 2 introduces the notation and core concepts used throughout the paper, including a review of probabilistic graphical models, variational inference and strong convexity.

- Chapter 3 introduces the concept of stability, beginning with a general definition for a generic functional (e.g., the composition of a loss function and hypothesis), followed by the more specific definition for collective stability, and concluding with a discussion of related notions in the literature.

- Chapter 4 introduces the statistical quantities and identities used in my analysis, as well as some examples of "nice" dependence conditions.

- Chapter 5 summarizes my early work in learning a structured predictor from limited examples. I derive a generalization bound based on the collective stability and covering number of the hypothesis class, then apply it to a class of predictors that use strongly convex variational inference.

- Chapter 6 presents more recent bounds, using PAC-Bayesian analysis and local stability. I apply these bounds to max-margin and soft-max learning of (conditional) Markov random fields.

- Chapter 7 explores the connections between strongly convex free energies, stability and the accuracy of learned marginal distributions. The theoretical insights of this section result in a new technique that can "strongly convexify" a wide range of variational free energies. Empirical results suggest that this approach can yield significantly more accurate marginals in practice.

- Chapter 8 concludes the document with a review of the main results and directions for future research.

Chapter 2: Preliminaries

In this chapter, I introduce the core concepts covered in this document. I begin with notational conventions. I then define structured prediction and the learning setup. Following this, I review some background on probabilistic graphical models, introducing the class of *templated Markov random fields* that I will use throughout this document. Finally, I review the concept of *strong convexity*, which will be used in Chapters 5 and 7.

2.1 Notational Conventions

Let $\mathcal{X} \subseteq \mathbb{R}^k$ denote a domain of observations. Let $\mathcal{Y}$ denote a finite set of categorical labels, represented by the standard basis (i.e., "one-hot") vectors, $\{\mathbf{e}_1, \ldots, \mathbf{e}_{|\mathcal{Y}|}\}$. Let $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$ denote the cross product of the two.

For a graph $G \triangleq (\mathcal{V}, \mathcal{E})$, with nodes $\mathcal{V} \triangleq \{1, \ldots, n\}$ and edges $\mathcal{E}$, I will use $|G| \triangleq |\mathcal{V}| + |\mathcal{E}|$ to denote the total number of nodes and edges in G. I will sometimes refer to this as the *size* of G.

Let $\mathbf{Z} \triangleq (Z_i)_{i=1}^n$ denote a set of random variables with joint distribution $\mathbb{D}$ on $\mathcal{Z}^n$. I will denote *realizations* of $\mathbf{Z}$ by $\mathbf{z} \in \mathcal{Z}^n$. I will use $\Pr_{\mathbf{Z} \sim \mathbb{D}}\{\,\cdot\,\}$ to denote the probability of an event over realizations of $\mathbf{Z}$, distributed according to $\mathbb{D}$. Similarly, I will use $\mathbb{E}_{\mathbf{Z} \sim \mathbb{D}}[\,\cdot\,]$ to specify an expectation over $\mathbf{Z}$. When it is clear from context which variable(s) and dis-

tribution the probability (or expectation) is taken over, I may omit the subscript notation. I will occasionally employ the shorthand $\mathbb{D}(\mathcal{S})$ to denote the measure of a subset $\mathcal{S} \subseteq \mathcal{Z}^n$ under $\mathbb{D}$; i.e., $\mathbb{D}(\mathcal{S}) = \Pr_{\mathbf{Z}\sim\mathbb{D}}\{\mathbf{Z} \in \mathcal{S}\}$. With a slight abuse of notation, which should be clear from context, I also use $\mathbb{D}(\mathbf{Z}_{i:j} \mid \Sigma)$ to denote the distribution of some subset of the variables, $(Z_i, \ldots, Z_j)$, conditioned on a σ-algebra Σ.

2.2 Structured Prediction

At its core, *structured prediction* (sometimes referred to as *structured output prediction* or *structured learning*) is about learning concepts that have a natural internal structure. In the framework I consider, each example of the concept contains n interdependent random variables, $\mathbf{Z} \triangleq (Z_i)_{i=1}^n$, with joint distribution $\mathbb{D}$. Each $Z_i \triangleq (X_i, Y_i)$ is an input-output pair, taking values in $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.[1] An example is associated with an implicit dependency graph, $G \triangleq (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} \triangleq \{1, \ldots, n\}$, and $\mathcal{E}$ captures the dependencies in $\mathbf{Z}$. Unless otherwise stated, assume that the edge structure is imposed by the modeler. The edge structure may be obvious from context, or may be inferred beforehand. To simplify the analysis, I assume that each example uses the same structure, and that it has been fixed *a priori*.

The learning algorithm's goal is to learn to predict $\mathbf{Y} \triangleq (Y_i)_{i=1}^n$, conditioned on $\mathbf{X} \triangleq (X_i)_{i=1}^n$. Let $\mathcal{H} \subseteq \{h : \mathcal{X}^n \to \mathcal{Y}^n\}$ denote a class of hypotheses, where each hypothesis possesses some internal representation. I am interested in hypotheses that perform joint reasoning over all variables simultaneously, according to some prior knowl-

[1] To minimize bookkeeping, I 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.

edge about the structure of the data. I therefore assume that computing $h(\mathbf{X})$ implicitly involves a global optimization that does not decompose over the outputs, due to dependencies. Note that I do not assume that the data is generated according to some target concept in $\mathcal{H}$. Indeed, $\mathcal{H}$ may be misspecified.

Predictors are evaluated using a *loss function* of the form $L : \mathcal{H} \times \mathcal{Z}^n \to \mathbb{R}_+$, where L may have access to the internal representation of h. For a loss function L and hypothesis h, denote the average loss on a set of m structured examples, $\hat{\mathbf{Z}} \triangleq (\mathbf{Z}^{(l)})_{l=1}^m$, by

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

Most training algorithms minimize (an upper bound on) the empirical loss, so if $h^\star$ is the learned hypothesis, then $\hat{L}(h, \hat{\mathbf{Z}})$ indicates of how well $h^\star$ fit the training data. Let $\overline{L}(h) \triangleq \mathbb{E}_{\mathbf{Z} \sim \mathbb{D}}[L(h, \mathbf{Z})]$ denote the expected loss (also known as the *risk*) over realizations of a single example $\mathbf{Z}$. This quantity corresponds to the error h will incur on future predictions.

The goal of *generalization* analysis is to bound the difference of the expected and empirical loss measures, $\overline{L}(h) - \hat{L}(h, \hat{\mathbf{Z}})$. I will refer to this quantity as the *generalization error*.[2] Given a learned hypothesis, with empirical loss $\hat{L}(h, \hat{\mathbf{Z}})$, a generalization bound provides an upper bound on the expected loss, $\overline{L}(h)$.

[2]This definition of generalization error differs from some literature, in which the term is used to refer to the expected loss.

2.3 Probabilistic Graphical Models

Arguably, most models used for structured prediction can be viewed as *probabilistic graphical models* (PGMs). A PGM is a statistical model in which the conditional independence structure is represented by a graph. I will focus on a popular class of undirected[3] PGMS known as *Markov random fields* (MRFs). MRFs generalize many models used in practice, such as (relational) Markov networks (Taskar et al., 2002), conditional random fields (Lafferty et al., 2001), and Markov logic networks (Richardson and Domingos, 2006).

2.3.1 Markov Random Fields

Recall that each example is associated with a dependency graph, $G \triangleq (\mathcal{V}, \mathcal{E})$. In this case, the edge set is assumed to be undirected. The parameters of an MRF are organized according to the *cliques* (i.e., complete subgraphs), $\mathcal{C}$, contained in G. For each clique, $c \in \mathcal{C}$, there is an associated real-valued *potential* function, $\theta_c(\mathbf{y}; \mathbf{w})$, parameterized by a vector of weights, $\mathbf{w} \in \mathbb{R}^d$, for some $d \geq 1$. This function indicates the score for $\mathbf{Y}_c$ being in state $\mathbf{y}_c$. The potentials define a log-linear probability distribution,

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

[3]This does not limit the applicability of my work, since there exists a straightforward conversion from directed PGMs to undirected PGMs (Koller and Friedman, 2009).

where

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

is a normalizing function known as the *log-partition*. The potential functions may additionally be conditioned on the observation $\mathbf{X} = \mathbf{x}$, in which case they are denoted by $\theta_c(\mathbf{y} \,|\, \mathbf{x}; \mathbf{w})$, and

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

Since the label space, $\mathcal{Y}$, is represented by the standard basis vectors, the joint state of a clique c is represented by a vector, $\mathbf{y}_c = \bigotimes_{i \in c} y_i$, of length $|\mathcal{Y}_c| \triangleq |\mathcal{Y}|^{|c|}$. With a slight abuse of notation, I overload the potential functions so that $\boldsymbol{\theta}_c(\mathbf{w}) \in \mathbb{R}^{|\mathcal{Y}_c|}$ (alternatively, $\boldsymbol{\theta}_c(\mathbf{x}; \mathbf{w}) \in \mathbb{R}^{|\mathcal{Y}_c|}$) denotes a vector of potentials, and

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

Thus, with

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

we have that

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

I refer to $\hat{\mathbf{y}}$ as the *full* representation of $\mathbf{y}$. To reduce clutter, I will sometimes denote the potentials, $\boldsymbol{\theta}(\mathbf{w})$ or $\boldsymbol{\theta}(\mathbf{x}; \mathbf{w})$, by simply $\boldsymbol{\theta}$, and denote the distribution under $\boldsymbol{\theta}$ by $p(\mathbf{Y}; \boldsymbol{\theta})$.

2.3.2 Inference

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. In general, both tasks are intractable—MAP inference is NP-hard and marginal inference is #P-hard (Roth, 1996)—though there are some useful special cases for which inference is tractable, and many approximation algorithms for the general case.

One class of approximate inference techniques uses a well-known *variational* form of the log-partition function:

$$\Phi(\boldsymbol{\theta}) = \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\theta} \cdot \boldsymbol{\mu} - \Phi^*(\boldsymbol{\mu}), \tag{2.1}$$

where $\mathcal{M}$ is the *marginal polytope*—the set of all consistent marginal vectors—and Φ^* is the *convex conjugate* of Φ. For any $\boldsymbol{\mu} \in \mathcal{M}$, there is a corresponding distribution, $p_{\boldsymbol{\mu}}$, such that $\boldsymbol{\mu}_c \cdot \mathbf{y}_c = p_{\boldsymbol{\mu}}(\mathbf{Y}_c = \mathbf{y}_c)$ for each clique, $c \in \mathcal{C}$, and clique state, $\mathbf{y}_c$. In the model I consider, $\Phi^*(\boldsymbol{\mu})$ is equal to the negative entropy of $p_{\boldsymbol{\mu}}$.[4] The negative of the quantity being maximized is often referred to as the *free energy*,

$$E(\boldsymbol{\mu}; \boldsymbol{\theta}) \triangleq -\boldsymbol{\theta} \cdot \boldsymbol{\mu} + \Phi^*(\boldsymbol{\mu}). \tag{2.2}$$

The gradient of $\Phi(\boldsymbol{\theta})$ is the maximizing $\boldsymbol{\mu}$ (i.e., minimizer of E), which corresponds to

[4] I omit some details of the conjugate function for simplicity of exposition. See Wainwright and Jordan (2008) for a precise definition.

the marginal distributions of $Y_1, \ldots, Y_n$. I denote this by

$$\boldsymbol{\mu}(\boldsymbol{\theta}) \triangleq \arg\min_{\boldsymbol{\mu} \in \mathcal{M}} E(\boldsymbol{\mu}; \boldsymbol{\theta}) = \nabla \Phi(\boldsymbol{\theta}). \tag{2.3}$$

Each vertex of $\mathcal{M}$ corresponds to a full labeling, $\hat{\mathbf{y}}$. If one removes the conjugate function from the free energy, then the inference objective is a linear function of $\mathcal{M}$. The minimizer of this function is a vertex of $\mathcal{M}$, so minimizing the free energy without Φ^* is equivalent to MAP inference.

Unfortunately, for general graph structures, $\mathcal{M}$ may require an exponential number of constraints, and Φ^* is difficult to compute. Many variational methods address these problems by relaxing $\mathcal{M}$ to an outer bound that uses a polynomial number of "local" constraints, and replacing Φ^* with a tractable approximation, $\tilde{\Phi}^*$. The *local* marginal polytope, $\tilde{\mathcal{M}} \supseteq \mathcal{M}$, is typically defined as follows:

$$\tilde{\mathcal{M}} \triangleq \left\{ \tilde{\boldsymbol{\mu}} \; : \; \begin{array}{c} \forall v \in \mathcal{V}, \; \sum_{j=1}^{|\mathcal{Y}|} \tilde{\mu}_v^j = 1 \; ; \\ \forall e \in \mathcal{E}, \forall v \in e, \; \sum_{i=1}^{|\mathcal{Y}|} \tilde{\mu}_e^{ij} = \tilde{\mu}_v^j \end{array} \right\}. \tag{2.4}$$

We call each $\tilde{\boldsymbol{\mu}} \in \tilde{\mathcal{M}}$ a set of *pseudomarginals*. With a slight abuse of notation, let

$$\tilde{E}(\tilde{\boldsymbol{\mu}}; \boldsymbol{\theta}) \triangleq -\boldsymbol{\theta} \cdot \tilde{\boldsymbol{\mu}} + \tilde{\Phi}^*(\tilde{\boldsymbol{\mu}})$$

denote a variational free energy for $\tilde{\Phi}^*$ and $\tilde{\mathcal{M}}$, let $\tilde{\Phi}(\boldsymbol{\theta}) \triangleq \max_{\tilde{\boldsymbol{\mu}} \in \tilde{\mathcal{M}}} -\tilde{E}(\tilde{\boldsymbol{\mu}}; \boldsymbol{\theta})$ denote

the convex conjugate of $\tilde{\Phi}^*$ (i.e., the approximate log-partition), and let

$$\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}) \triangleq \arg\min_{\tilde{\boldsymbol{\mu}} \in \tilde{\mathcal{M}}} \tilde{E}(\tilde{\boldsymbol{\mu}}; \boldsymbol{\theta}) = \nabla \tilde{\Phi}(\boldsymbol{\theta}) \tag{2.5}$$

denote the pseudomarginals of the variational distribution,

$$\tilde{p}(\mathbf{Y} = \mathbf{y}; \boldsymbol{\theta}) \triangleq \exp\big(\boldsymbol{\theta} \cdot \hat{\mathbf{y}} - \tilde{\Phi}(\boldsymbol{\theta})\big). \tag{2.6}$$

To perform approximate MAP inference, one removes $\tilde{\Phi}^*$ from the inference objective. Since $\tilde{\mathcal{M}}$ is an outer bound on $\mathcal{M}$, this approximation may result in fractional solutions (Wainwright and Jordan, 2008). When this happens, the fractional solution can be rounded to the nearest integral solution by selecting the highest scoring label for each node.

A related form of inference is *posterior decoding*. Posterior decoding selects the labeling that maximizes the marginal probability at each node. Given a vector of (pseudo)marginals, $\boldsymbol{\mu}$, from (approximate) marginal inference, one decodes the label for node v as $y_v \triangleq \arg\max_{y \in \mathcal{Y}} y \cdot \mu_v$. (This formula assumes a basis vector representation of $\mathcal{Y}$.) Posterior decoding is more robust to outliers than MAP inference because it maximizes the expected per-label accuracy rather than the accuracy of the joint labeling (Gross et al., 2006).

2.3.3 Templating

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

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

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

In practice, templated models typically consist of unary and pairwise templates. I will refer to these as pairwise models. Higher-order templates can capture certain induc-

tive rules that pairwise models cannot. For example, for a binary relation r, the transitive closure $r(A, B) \wedge r(B, C) \implies r(A, C)$ requires triadic templates. Rules like this are sometimes used for link prediction and entity resolution. Of course, this additional expressivity comes at a cost, as will become apparent later.

2.3.4 Defining the Potential Functions

In this section, I describe a common formulation of the potential functions based on simple, multilinear functions of $(\mathbf{w}, \mathbf{x}, \mathbf{y})$. Assume that each node i has local observations $x_i \in \mathcal{X}$ and label $y_i \in \mathcal{Y}$. Define a vector of local *features*,

$$f_i(\mathbf{x}, \mathbf{y}) \triangleq x_i \otimes y_i, \tag{2.7}$$

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

$$f_{ij}(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{2} \begin{bmatrix} x_i \\ x_j \end{bmatrix} \otimes (y_i \otimes y_j). \tag{2.8}$$

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

The potential functions are then defined as weighted feature functions. For the following, I will assume that the weights are templated, as described in Section 2.3.3. The node features are associated with a set of singleton weights, $\mathbf{w}_s \in \mathbb{R}^{d_s}$, and the edge

features with a set of pairwise weights, $\mathbf{w}_p \in \mathbb{R}^{d_p}$, where d_s and d_p denote the respective lengths of the feature vectors. Then,

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

and, with

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

we have that

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

Using this parameterization of the potentials, one obtains two key technical lemmas, which will be used later on. These are deferred to Appendix A.

2.4 Strong Convexity

Various results in this document rely on the concept of *strong convexity*. The following is the most general definition.

Definition 1. A function, $\varphi : \mathcal{S} \to \mathbb{R}$, of a convex set, $\mathcal{S}$, is κ-*strongly convex* with respect to a norm[5], $\|\cdot\|$, if and only if, for all $s, s' \in \mathcal{S}$ and $\tau \in [0, 1]$,

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

The *modulus* of convexity, κ, measures the curvature of φ.

[5]Unless specified, assume strong convexity with respect to the 2-norm.

Differentiable functions have the following simplified definition.

Definition 2. A differentiable function, $\varphi : \mathcal{S} \to \mathbb{R}$, of a convex set, $\mathcal{S}$, is κ-strongly convex with respect to a norm, $\|\cdot\|$, if and only if, for all $s, s' \in \mathcal{S}$,

$$\frac{\kappa}{2}\|s - s'\|^2 + \langle \nabla\varphi(s),\, s' - s\rangle \leq \varphi(s') - \varphi(s).$$

Strong convexity of can be characterized in a number of ways. The following facts provide some conditions that are equivalent to Definition 2.

Fact 1. *A differentiable function, $\varphi : \mathcal{S} \to \mathbb{R}$, of a convex set, $\mathcal{S}$, is κ-strongly convex w.r.t. a norm, $\|\cdot\|$, if and only if, for all $s, s' \in \mathcal{S}$,*

$$\kappa\|s - s'\|^2 \leq \langle \nabla\varphi(s) - \nabla\varphi(s'),\, s - s'\rangle\,.$$

Fact 2. *A twice-differentiable function, $\varphi : \mathcal{S} \to \mathbb{R}$, of a convex set, $\mathcal{S}$, is κ-strongly convex w.r.t. a norm, $\|\cdot\|$, if and only if, for all $s, s' \in \mathcal{S}$,*

$$\kappa\|s\|^2 \leq \left\langle s, \nabla^2\varphi(s')\, s\right\rangle.$$

For the 2-norm, Fact 5 means that the minimum eigenvalue of the Hessian is lower-bounded by κ.

One convenient property of a strongly convex function is that one can upper-bound the squared distance between the minimizer and any other point in the domain by a function of the difference between their respective evaluations. This identity is illustrated in

Figure 2.1, and formalized in the following lemma.

Lemma 1. *Let $\varphi : \mathcal{S} \to \mathbb{R}$ be κ-strongly convex, and let $\dot{s} \triangleq \arg\min_{s \in \mathcal{S}} \varphi(s)$. Then, for any $s \in \mathcal{S}$,*

$$\|s - \dot{s}\|^2 \leq \frac{2}{\kappa} \left(\varphi(s) - \varphi(\dot{s})\right).$$

Proof Let $\Delta s \triangleq s - \dot{s}$. By Definition 1, for any $\tau \in [0, 1]$,

$$\frac{\kappa}{2} \tau(1 - \tau) \|\Delta s\|^2 + \varphi(\dot{s} + \tau \Delta s) - \varphi(\dot{s}) \leq \tau \left(\varphi(s) - \varphi(\dot{s})\right).$$

Since $\dot{s}$ is the unique minimizer of φ, it follows that $\varphi(\dot{s} + \tau \Delta s) - \varphi(\dot{s}) \geq 0$; so the above inequality is preserved when this term is dropped. Then, dividing both sides by $\tau\kappa/2$, we have that

$$(1 - \tau) \|\Delta s\|^2 \leq \frac{2}{\kappa} \left(\varphi(s) - \varphi(\dot{s})\right).$$

Since this holds for all $\tau \in [0, 1]$, it holds for $\tau = 1$, which completes the proof. ■

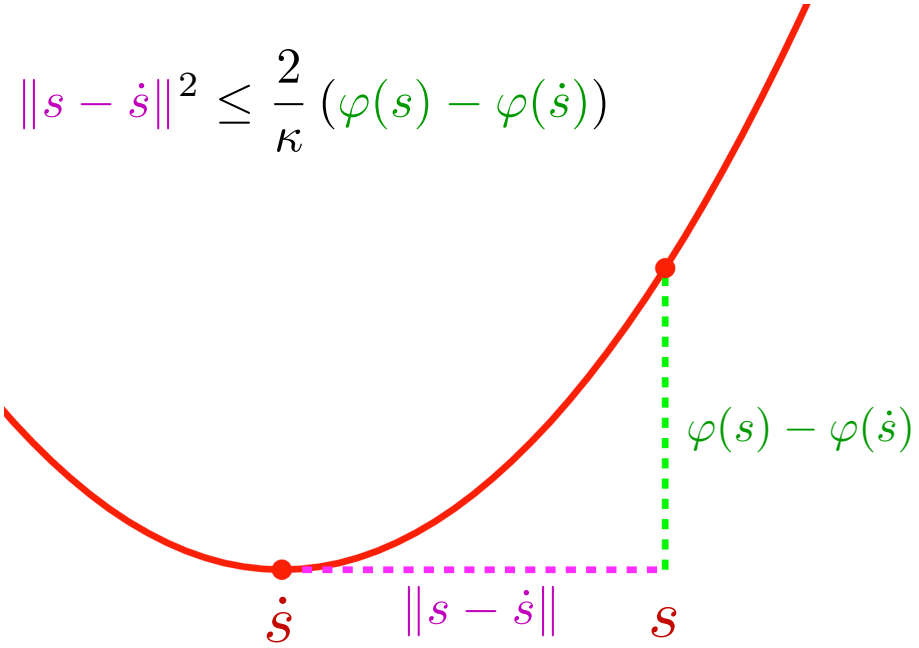


Figure 2.1: Illustration of Lemma 1. For a κ-strongly convex function, $\varphi : \mathcal{S} \to \mathbb{R}$, the squared distance between the minimizer, $\dot{s}$, and any other point, $s \in \mathcal{S}$, is upper-bounded by $\frac{2}{\kappa} (\varphi(s) - \varphi(\dot{s}))$.

Chapter 3: Stability

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

and relate them to other forms found in the literature.

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

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

3.1 A General Definition of Stability

In this section, I introduce definitions of stability for functionals; i.e., functions that map multiple inputs to a single scalar output. Loss functions are examples of functionals. The following definitions will be stated in terms of an arbitrary class of functionals, $\mathcal{F} \triangleq \{\varphi : \mathcal{Z}^n \to \mathbb{R}\}$.

The simplest form of stability is the uniform condition, in which stability must hold uniformly over all inputs.

Definition 3. A function $\varphi \in \mathcal{F}$ is β-*uniformly stable* if, for any inputs $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}^n$,

$$|\varphi(\mathbf{z}) - \varphi(\mathbf{z}')| \leq \beta \, D_{\mathrm{H}}(\mathbf{z}, \mathbf{z}'). \tag{3.1}$$

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

Put differently, a uniformly stable function is Lipschitz under the Hamming norm.

Uniform stability over the entire domain can be a strong requirement. Sometimes, stability only holds for a certain subset of inputs, such as points contained in a Euclidean ball of a certain radius. I refer to the set of inputs for which stability holds as the "good"

set; all other inputs are "bad." The precise meaning of good and bad depends on the hypothesis class. Given some delineation of good and bad, one obtains the following localized notion of stability.

Definition 4. For a subset $\mathcal{B}_{\mathcal{Z}} \subseteq \mathcal{Z}^n$, a function $\varphi \in \mathcal{F}$ is $(\beta, \mathcal{B}_{\mathcal{Z}})$-*locally stable* if Equation 3.1 holds for all $\mathbf{z}, \mathbf{z}' \notin \mathcal{B}_{\mathcal{Z}}$. The class $\mathcal{F}$ is $(\beta, \mathcal{B}_{\mathcal{Z}})$-locally stable if every $\varphi \in \mathcal{F}$ is $(\beta, \mathcal{B}_{\mathcal{Z}})$-locally stable.

Definition 4 has an alternate probabilistic interpretation. If $\mathbb{D}$ is a distribution on $\mathcal{Z}^n$, then β-uniform stability holds with some probability over draws of $\mathbf{z}, \mathbf{z}' \sim \mathbb{D}$. If the bad set $\mathcal{B}_{\mathcal{Z}}$ has measure $\mathbb{D}(\mathcal{B}_{\mathcal{Z}}) \leq \nu$, then $(\beta, \mathcal{B}_{\mathcal{Z}})$-local stability is similar to, though slightly weaker than, the *strongly difference-bounded* property proposed by Kutin (2002). If φ is strongly difference-bounded, then Equation 3.1 must hold for any $\mathbf{z} \notin \mathcal{B}_{\mathcal{Z}}$ and $\mathbf{z}' \in \mathcal{Z}^n$ (which could be in $\mathcal{B}_{\mathcal{Z}}$). All functions that are strongly difference-bounded are locally stable, but the converse is not true.

The notion of probabilistic stability can be extended to distributions on the function class. For any stability parameter β (and bad inputs $\mathcal{B}_{\mathcal{Z}}$), the function class is partitioned into functions that satisfy Equation 3.1, and those that do not. Therefore, for any distribution $\mathbb{Q}$ on $\mathcal{F}$, uniform (or local) stability holds with some probability over draws of $\varphi \sim \mathbb{Q}$. This idea motivates the following definition.

Definition 5. Fix some $\beta \geq 0$ and $\mathcal{B}_{\mathcal{Z}} \subseteq \mathcal{Z}^n$, and let $\mathcal{B}_{\mathcal{F}} \subseteq \mathcal{F}$ denote the subset of functions that are not $(\beta, \mathcal{B}_{\mathcal{Z}})$-locally stable. A distribution $\mathbb{Q}$ on $\mathcal{F}$ is $(\beta, \mathcal{B}_{\mathcal{Z}}, \eta)$-*locally stable* if $\mathbb{Q}(\mathcal{B}_{\mathcal{F}}) \leq \eta$.

Note the taxonomical relationship between these definitions. Definition 3 is the

strongest condition, since it implies Definitions 4 and 5. Clearly, if $\mathcal{F}$ is β-uniformly stable, then it is $(\beta, \emptyset)$-locally and $(\beta, \emptyset, 0)$-locally stable. Definition 4 extends Definition 3 by accommodating broader domains. Definition 5 extends this even further, by accommodating classes in which only some functions satisfy local stability.

Stability analyzes the sensitivity of a function (or class) to single variable perturbations. A related property is its sensitivity to changes in *any* of the variables, which we formalize in the following.

Definition 6. A function $\varphi \in \mathcal{F}$ is α-*uniformly range-bounded* if, for any $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}^n$,

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

I will use this as a "fall-back" property for when stability does not hold.

3.2 Collective Stability

My first generalization bounds for structured prediction (London et al., 2013a) crucially rely on a property I referred to as *uniform collective stability*. Whereas the previous definitions of stability apply to functionals (multiple inputs, scalar output), collective stability applies to vector-valued functions (multiple inputs, multiple outputs). Collective stability measures the stability of the predictor. The following definition is for the uniform case.[1]

Definition 7. A class of vector-valued functions, $\mathcal{G} \triangleq \left\{g : \mathcal{Z}^n \to \mathbb{R}^N\right\}$, has β-*uniform*

[1] Non-uniform definitions of collective stability have been proposed (London et al., 2014), but they will not be used in this document.

collective stability if, for any $g \in \mathcal{G}$, and any $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}^n$,

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

I later relaxed this requirement to various non-uniform definitions of collective stability (London et al., 2014). The results presented in this document will only use the uniform definition.

3.3 Connections to Other Notions of Stability

In the learning theory literature, the word "stability" has traditionally been associated with a *learning* algorithm, rather than an inference algorithm. A learning algorithm is said to be stable with respect to a loss function if the loss of a learned hypothesis varies by a bounded amount upon replacing (or deleting) examples from the training set. This property has been used to derive generalization bounds (e.g., Bousquet and Elisseeff, 2002), in the same way I use stability of inference.

My definitions of stability should also be contrasted with *sensitivity analysis*. Since the terms are often used interchangeably, I distinguish the two as follows: stability measures the amount of change induced in the *output* of a function upon perturbing its input within a certain range, and sensitivity analysis measures the amount of perturbation one can apply to the *input* such that its output remains within a certain range. By these definitions, one is the dual of the other. In the context of probabilistic inference, sensitivity analysis has been used to determine the maximum amount one can perturb the model parameters (or evidence) such that the likelihood of a query stays within a given toler-

ance, or such that the most likely assignment does not change (Chan and Darwiche, 2005, 2006). Stability measures *how much* the likelihood or most likely assignment changes.

Chapter 4: Statistical Tools

Before presenting the generalization bounds, I will review some supporting definitions and introduce a new moment-generating function inequality for locally stable functions of interdependent random variables. This inequality will be used in the PAC-Bayes bounds (Chapter 6). It can also be used to derive a concentration inequality for uniformly stable functions, which will be used in the covering number-based bounds (Chapter 5). I conclude this section with some example conditions under which the dependence is bounded (Section 4.3), thereby supporting improved generalization bounds.

4.1 Dependency Matrix

The main results of this section leverage a data structure that quantifies dependence between random variables. Let π be a permutation of $[n] \triangleq \{1, \ldots, n\}$, where $\pi(i)$ denotes the i^{th} element in the sequence and $\pi(i:j)$ denotes a subsequence of elements i through j. Used to index variables $\mathbf{Z} \triangleq (Z_i)_{i=1}^n$, denote by $Z_{\pi(i)}$ the i^{th} variable in the permutation and $\mathbf{Z}_{\pi(i:j)}$ the subsequence $(Z_{\pi(i)}, \ldots, Z_{\pi(j)})$.

Definition 8. 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, the *total variation distance* is

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

To clarify notation of suprema (or infima) over $\mathcal{Z}$, when probabilities are conditioned on Σ, I will occasionally write $\mathcal{Z}_\Sigma$ to denote the subset of $\mathcal{Z}$ that is consistent with Σ.

Definition 9. Fix a filtration $\boldsymbol{\pi} \in \Pi(n)$ and a σ-algebra Σ on $\mathcal{Z}^n$. For $i \in [n]$, $j > i$, $\mathbf{z} \in \mathcal{Z}^{i-1}$ and $z, z' \in \mathcal{Z}$, define the *ϑ-mixing coefficients*[1] as

$$\vartheta_{ij}^{\boldsymbol{\pi}}(\mathbf{z}, z, z') \triangleq \left\| \mathbb{D}\left(\mathbf{Z}_{\pi_i(j:n)} \,|\, \Sigma, \mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z)\right) - \mathbb{D}\left(\mathbf{Z}_{\pi_i(j:n)} \,|\, \Sigma, \mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z')\right) \right\|_{\text{TV}},$$

where it is assumed that $\mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z)$ and $\mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z')$ are consistent with Σ. These coefficients define an upper-triangular *dependency matrix* $\boldsymbol{\Gamma}_\Sigma^{\boldsymbol{\pi}} \in \mathbb{R}^{n \times n}$, with entries

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

When Σ is the full σ-algebra of $\mathbf{Z}$, I will simply omit the subscript notation.

The bounds in the following sections will use the induced matrix infinity norm of

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

$\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\Sigma}$, denoted

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

to measure the aggregate dependence in the distribution. Observe that, if $Z_1, \ldots, Z_n$ are mutually independent, then $\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\Sigma}$ is the identity matrix and $\left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\Sigma}\right\|_{\infty} = 1$.

The ordering of the variables in each row of $\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\Sigma}$ can have a strong impact on $\left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\Sigma}\right\|_{\infty}$. Since we do not assume that $\mathbf{Z}$ corresponds to a temporal process, there may not be any natural ordering of the variables. In general, given an arbitrary graph topology, $\left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\Sigma}\right\|_{\infty}$ 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 $\left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\Sigma}\right\|_{\infty}$ is bounded independently of n (i.e., $\left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\Sigma}\right\|_{\infty} = \mathrm{O}(1)$). By exploiting filtrations, one can show that the same holds for Markov random fields of any bounded-degree structure, provided the distribution exhibits suitable mixing. I discuss these conditions in Section 4.3.

4.2 Statistical Inequalities

With the supporting definitions in mind, I now present a new moment-generating function inequality. The proof is provided in Appendix B.3.

Proposition 1. *Let* $\mathbf{Z} \triangleq (Z_i)_{i=1}^n$ *be random variables with joint distribution* $\mathbb{D}$ *on* $\mathcal{Z}^n$. *Let* $\mathcal{B}_{\mathcal{Z}} \subseteq \mathcal{Z}^n$ *denote a set of "bad" inputs. Let* $\overline{\mathcal{B}}$ *denote the* σ*-algebra corresponding to*

$\mathbf{Z} \notin \mathcal{B}_{\mathcal{Z}}$. *Let* $\varphi : \mathcal{Z}^n \to \mathbb{R}$ *be a measurable function with* $(\beta, \mathcal{B}_{\mathcal{Z}})$*-local stability. Then, for any* $\tau \in \mathbb{R}$ *and filtration* $\boldsymbol{\pi} \in \Pi(n)$*,*

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

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

Proposition 1 yields a novel concentration inequality for uniformly stable functions of interdependent random variables. The proof is given in Appendix B.4.

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

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

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

inequality.

4.3 Bounded Dependence Conditions

The infinity norm of the dependency matrix, $\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\Sigma}$, has a trivial upper bound, $\left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\Sigma}\right\|_{\infty} \leq n$. However, we are interested in bounds that are sub-logarithmic in n. In this section, I describe some general settings in which $\left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\Sigma}\right\|_{\infty}$ has a nontrivial upper bound. For the remainder of this section, fix a graph, $G \triangleq (\mathcal{V}, \mathcal{E})$, for the *data* distribution. For any two subsets, $\mathcal{S}, \mathcal{T} \subseteq \mathcal{V}$, define their graph distance, $D_{\mathrm{G}}(\mathcal{S}, \mathcal{T})$, as the length of the shortest path from any node in $\mathcal{S}$ to any node in $\mathcal{T}$. I will use the following notion of distance-based dependence.

Definition 10. For a random field $\mathbf{Z}$ on a graph G, with distribution $\mathbb{D}$, and a σ-algebra Σ on $\mathcal{Z}^n$, define the *distance-based ϑ-mixing coefficients* as

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

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

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

$n)) \geq k$. Therefore, for each upper-triangular entry of $\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\Sigma}$, we have that

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

Proposition 2. *Let* $\mathbf{Z}$ *be a random field on a graph* G*, with maximum degree* Δ_G*. For any positive constant* $\epsilon > 0$*, if* $\mathbf{Z}$ *admits a distance-based* ϑ*-mixing sequence such that, for all* $k \geq 1$*,* $\vartheta(k) \leq (\Delta_G + \epsilon)^{-k}$*, then there exists a filtration* $\boldsymbol{\pi}$ *such that*

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

The proof is provided in Appendix B.5.

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

Another important setting is when the graph is a chain. Chain-structured stochastic processes (usually temporal) under various mixing assumptions have been well-studied (see Bradley, 2005 for a comprehensive survey). It can be shown that any *contracting*

Markov chain has $\|\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\Sigma}\|_{\infty} = \mathrm{O}(1)$ (Kontorovich, 2012). Here, I provide an alternate condition, using distance-based ϑ-mixing, under which the dependency matrix of a chain has suitably low norm. The key property of a chain is that the number of nodes at distance k from any starting node is constant. One can therefore relax the assumption of geometric decay used in the previous result.

Proposition 3. *Let* $\mathbf{Z}$ *be a chain-structured random field, of length* n*. For any constants* $\epsilon > 0$ *and* $p \geq 1$*, if* $\mathbf{Z}$ *admits a distance-based* ϑ*-mixing sequence such that, for all* $k \geq 1$*,* $\vartheta(k) \leq \epsilon k^{-p}$*, then there exists a filtration* $\boldsymbol{\pi}$ *such that*

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

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

The proof is provided in Appendix B.6.

For $p > 1$, the Riemann function converges to a constant. For example, $\zeta(2) = \pi^2/6 \approx 1.645$. However, even $p = 1$ yields a sufficiently low growth rate. In Chapters 5 and 6, I prove generalization bounds of the form $\mathrm{O}\left(\|\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\Sigma}\|_{\infty} / \sqrt{mn}\right)$, which still converges if $\|\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\Sigma}\|_{\infty} = \mathrm{O}(\ln n)$, albeit at a slower rate.

Chapter 5: Generalization Bounds via Collective Stability

In many applications of structured prediction, each example has large internal structure, so obtaining labeled examples can be expensive. It is therefore common to train a structured model on a few large examples—sometimes even just one example. Since previous learning guarantees (e.g., Taskar et al., 2004; Bartlett et al., 2005; McAllester, 2007) are vacuous for small training samples, the goal of this chapter is to show that one can indeed obtain good generalization in this setting.

I first present a generalization bound based on the collective stability of the class of predictors. This bound reviews (and corrects) my early work on this topic (London et al., 2013a). I then apply the bound to a class of MRFs that use posterior decoding with strongly convex variational inference. To my knowledge, the results reviewed in this chapter were the first to show that one could potentially generalize from a single large example, and the first to point out the importance of inference stability.

5.1 Covering Number

The generalization bound presented in the following section is stated in terms of the *covering number* of the hypothesis class. The covering number is an approximate bound on the number of functions in the class. The *uniform* covering number effectively discretizes

the space of hypotheses, allowing one to efficiently apply a union bound over all hypotheses in the class. Low covering number equates with low complexity, which therefore leads to better generalization.

Definition 11. Let $\mathcal{S}$ be a pseudometric space with pseudometric $\rho : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_+$. A set $\mathcal{C} \subseteq \mathcal{S}$ is an *ϵ-cover* of $\mathcal{A} \subseteq \mathcal{S}$ under ρ if, for any $a \in \mathcal{A}$, there exists a $c \in \mathcal{C}$ such that $\rho(a, c) \leq \epsilon$.

Definition 12. Let $\mathcal{F}$ be a class of functions from $\mathcal{X}^n$ to $\mathbb{R}^N$. For $\varphi, \varphi' \in \mathcal{F}$, let

$$\rho_\infty(\varphi, \varphi') \triangleq \sup_{\mathbf{x} \in \mathcal{X}^n} \frac{1}{N} \left\| \varphi(\mathbf{x}) - \varphi'(\mathbf{x}) \right\|_1 .$$

The *uniform covering number*, $\mathcal{N}_\infty(\epsilon, \mathcal{F}, n)$, is the cardinality of the minimal class, $\mathcal{F}' \subseteq \mathcal{F}$, needed to ϵ-cover $\mathcal{F}$ under ρ_∞.

This definition differs slightly from the canonical definition used in the literature. Covering-based analyses typically use the *empirical* covering number, which requires a symmetrization step to employ (Pollard, 1984). The above *uniform* covering number is stronger; however, it does not require symmetrization. Modulo notation, this definition can be viewed as a structured extension of the ∞-norm covering number used by Bartlett (1998).

5.2 Combining Collective Stability and Covering Number

In this section, I present a general risk bound for structured prediction, stated in terms of the collective stability and covering number of the hypothesis class. This bound is

an adaptation of my first published bounds (London et al., 2013a). It corrects a critical mistake in the analysis by replacing Rademacher complexity with the covering number. Since the applications of the original bounds used the covering number, the corrected bound yields similar results.

For the following, I will assume an admissible *decomposable* loss function.

Definition 13. An *decomposable* is a loss function, $L : \mathcal{H} \times \mathcal{Z}^n \to \mathbb{R}_+$, such that, for a hypothesis, $h : \mathcal{X}^n \to \hat{\mathcal{Y}}^n$,

$$L(h, \mathbf{Z}) = \frac{1}{n} \sum_{i=1}^{n} c(Y_i, h_i(\mathbf{X})),$$

for some *cost function*, $c : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R}_+$.

An example of a decomposable loss function is given in Section 5.3. Note that the output of the hypothesis is not assumed to be $\mathcal{Y}^n$; in the examples given, $\hat{\mathcal{Y}}$ is a continuous relaxation of $\mathcal{Y}$.

The stability of inference affects the stability of the loss; for a decomposable loss, the cost function also affects the stability. In this work, certain cost functions are considered *admissible*.

Definition 14. A cost function, c, is (M, λ)-*admissible* if:

1. c is M-uniformly range-bounded;
2. for any $y \in \mathcal{Y}$ and $\hat{y}, \hat{y}' \in \hat{\mathcal{Y}}$, $\left|c(y, \hat{y}) - c(y, \hat{y}')\right| \leq \lambda \left\|\hat{y} - \hat{y}'\right\|_1$.

Lemma 2. *If c is (M, λ)-admissible, and $\mathcal{H}$ has β-uniform collective stability, then $c \circ \mathcal{H}$ has $(M + \lambda\beta)$-uniform collective stability.*

The proof is given in Appendix C.1.

Theorem 1. *Fix $m \geq 1$, $n \geq 1$, $\boldsymbol{\pi} \in \Pi(n)$, $\delta \in (0,1)$. Fix a distribution, $\mathbb{D}$, on $\mathcal{Z}^n$, and let $\boldsymbol{\Gamma}^{\boldsymbol{\pi}}$ denote the dependency matrix induced by $\mathbb{D}$ and $\boldsymbol{\pi}$. Fix a hypothesis class, $\mathcal{H}$, with β-uniform collective stability, and a decomposable loss function, L, with a (M,λ)-admissible cost function, c. Then, with probability at least $1-\delta$ over realizations of a training set, $\hat{\mathbf{Z}} \triangleq (\mathbf{Z}^{(l)})_{l=1}^{m}$, drawn according to $\mathbb{D}^m$, every $h \in \mathcal{H}$ satisfies*

$$\overline{L}(h) \leq \hat{L}(h,\hat{\mathbf{Z}}) + \inf_{\epsilon}\; 2\left|\mathcal{Y}\right|\lambda\,\epsilon + (M+\lambda\beta)\left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}\right\|_{\infty}\sqrt{\frac{\ln\left(\mathcal{N}_{\infty}(\epsilon,\mathcal{H},n)/\delta\right)}{2mn}}. \tag{5.1}$$

Proof For the following, let $\mathcal{F}$ be a finite class of functions from $\mathcal{Z}^n$ to $\mathbb{R}^n$. For any particular $\varphi \in \mathcal{F}$, let

$$\overline{\varphi}(\hat{\mathbf{Z}}) \triangleq \frac{1}{mn}\sum_{l=1}^{m}\sum_{i=1}^{n}\varphi_i(\mathbf{Z}^{(l)}),$$

and

$$\phi(\varphi,\hat{\mathbf{Z}}) \triangleq \mathbb{E}[\overline{\varphi}(\hat{\mathbf{Z}})] - \overline{\varphi}(\hat{\mathbf{Z}}) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\varphi_i(\mathbf{Z})\right] - \overline{\varphi}(\hat{\mathbf{Z}}).$$

The last equality follows from linearity of expectation, and the fact that $\mathbf{Z}^{(1)},\ldots,\mathbf{Z}^{(m)}$ are identically distributed. Note that $\mathbb{E}[\phi(\varphi,\hat{\mathbf{Z}})] = 0$.

Now, suppose $\mathcal{F}$ has uniform collective stability β. Then, for any two realized

training sets, $\hat{\mathbf{z}}$ and $\hat{\mathbf{z}}'$, that differ only in the l^{th} example,

$$\begin{aligned}\left|\phi(\varphi,\hat{\mathbf{z}})-\phi(\varphi,\hat{\mathbf{z}}')\right| &= \left|\overline{\varphi}(\hat{\mathbf{z}})-\overline{\varphi}(\hat{\mathbf{z}}')\right| \\ &= \left|\frac{1}{mn}\sum_{i=1}^{n}\varphi_i(\mathbf{z}^{(l)})-\varphi_i(\mathbf{z}'^{(l)})\right| \\ &\leq \frac{1}{mn}\left\|\varphi(\mathbf{z}^{(l)})-\varphi(\mathbf{z}'^{(l)})\right\|_1 \\ &\leq \frac{\beta}{mn}D_{\text{H}}(\mathbf{z}^{(l)},\mathbf{z}'^{(l)}).\end{aligned}$$

Using the triangle inequality, it can then be shown that, for any $\hat{\mathbf{z}}$ and $\hat{\mathbf{z}}'$, which may differ at multiple examples,

$$\left|\phi(\varphi,\hat{\mathbf{z}})-\phi(\varphi,\hat{\mathbf{z}}')\right| \leq \frac{\beta}{mn}D_{\text{H}}(\hat{\mathbf{z}},\hat{\mathbf{z}}').$$

Thus, every $\phi(\varphi,\cdot):\varphi\in\mathcal{F}$ is $(\beta/(mn))$-uniformly stable.

Since each example, $\mathbf{Z}^{(l)}$, is independent and identically distributed, the dependency matrix induced by $\hat{\mathbf{Z}}$ is block diagonal, with each block equal to $\mathbf{\Gamma}^{\boldsymbol{\pi}}$; i.e.,

$$\begin{bmatrix}\mathbf{\Gamma}^{\boldsymbol{\pi}} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Gamma}^{\boldsymbol{\pi}} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{\Gamma}^{\boldsymbol{\pi}}\end{bmatrix}$$

The infinity norm of this matrix is clearly the infinity norm of $\mathbf{\Gamma}^{\boldsymbol{\pi}}$. We therefore apply the

union bound and Corollary 1, with $N \triangleq mn$ and $\bar{\beta} \triangleq \beta/N$, and have that

$$\begin{aligned}
\Pr\left\{\sup_{\varphi\in\mathcal{F}} \phi(\varphi, \hat{\mathbf{Z}}) \geq \tau\right\} &= \Pr\left\{\exists\varphi \in \mathcal{F} \,:\, \phi(\varphi, \hat{\mathbf{Z}}) \geq \tau\right\} \\
&\leq \sum_{\varphi\in\mathcal{F}} \Pr\left\{\phi(\varphi, \hat{\mathbf{Z}}) \geq \tau\right\} \\
&\leq |\mathcal{F}| \exp\left(\frac{-2\epsilon^2}{N\bar{\beta}^2 \left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}\right\|_\infty^2}\right) \\
&= |\mathcal{F}| \exp\left(\frac{-2mn\epsilon^2}{\beta^2 \left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}\right\|_\infty^2}\right).
\end{aligned}$$

Assigning δ probability to this event and solving for τ, we have that with probability at least $1-\delta$,

$$\sup_{\varphi\in\mathcal{F}} \phi(\varphi, \mathbf{Z}) \leq \beta \left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}\right\|_\infty \sqrt{\frac{\ln(|\mathcal{F}|/\delta)}{2mn}}. \tag{5.2}$$

Now, suppose we had a finite hypothesis class, $\mathcal{H}' \subseteq \mathcal{H}$, that ϵ-covered $\mathcal{H}$ under ρ_∞. By the admissibility of c, we would have that, for any $h \in \mathcal{H}$, there exists a corresponding $h' \in \mathcal{H}$ where

$$\begin{aligned}
|L(h, \mathbf{Z}) - L(h', \mathbf{Z})| &\leq \frac{1}{n}\sum_{i=1}^{n} |c(Y_i, h_i(\mathbf{X})) - c(Y_i, h'_i(\mathbf{X}))| \\
&\leq \frac{\lambda}{n} \left\|h(\mathbf{X}) - h'(\mathbf{X})\right\|_1 \\
&= |\mathcal{Y}|\,\lambda\,\frac{1}{|\mathcal{Y}|\,n} \left\|h(\mathbf{X}) - h'(\mathbf{X})\right\|_1 \\
&\leq |\mathcal{Y}|\,\lambda\,\epsilon.
\end{aligned}$$

(The prediction vectors, $h(\mathbf{x})$ and $h'(\mathbf{x})$, have length $(|\mathcal{Y}|\,n)$, so the term $(|\mathcal{Y}|\,n)$ in the

denominator becomes the normalizing factor for the pseudometric, ρ_∞.) This means that

$$\left|\overline{L}(h) - \overline{L}(h')\right| \leq \mathbb{E}\left[|L(h, \mathbf{Z}) - L(h', \mathbf{Z})|\right] \leq |\mathcal{Y}|\,\lambda\,\epsilon, \tag{5.3}$$

and

$$\left|\hat{L}(h, \hat{\mathbf{Z}}) - \hat{L}(h', \hat{\mathbf{Z}})\right| \leq \frac{1}{m}\sum_{l=1}^{m}\left|L(h, \mathbf{Z}^{(l)}) - L(h', \mathbf{Z}^{(l)})\right| \leq |\mathcal{Y}|\,\lambda\,\epsilon. \tag{5.4}$$

Take $\mathcal{H}'$ to be the minimal ϵ-cover of $\mathcal{H}$, and let $\mathcal{F} \triangleq c \circ \mathcal{H}'$. Every $h' \in \mathcal{H}'$ has a corresponding $\varphi \in \mathcal{F}$ such that

$$\overline{\varphi}(\hat{\mathbf{Z}}) = \frac{1}{m}\sum_{l=1}^{m}\frac{1}{n}\sum_{i=1}^{n} c(Y_i^{(l)}, h_i'(\mathbf{X}^{(l)})) = \frac{1}{m}\sum_{l=1}^{m} L(h', \mathbf{Z}^{(l)}) = \hat{L}(h', \hat{\mathbf{Z}}),$$

and

$$\phi(\varphi, \hat{\mathbf{Z}}) \triangleq \mathbb{E}[\hat{L}(h', \hat{\mathbf{Z}})] - \hat{L}(h', \hat{\mathbf{Z}}) = \overline{L}(h') - \hat{L}(h', \hat{\mathbf{Z}}).$$

Note that $|\mathcal{F}| = |\mathcal{H}'| = \mathcal{N}_\infty(\epsilon, \mathcal{H}, n)$. Further, since $\mathcal{H}$ has β-uniform collective stability, so does $\mathcal{H}'$; and by Lemma 2, $\mathcal{F}$ has $(M + \lambda\beta)$-uniform collective stability. Using Equation 5.2, we therefore have that, with probability at least $1-\delta$, every $h' \in \mathcal{H}'$ satisfies

$$\begin{aligned}
\overline{L}(h') &\leq \hat{L}(h', \hat{\mathbf{Z}}) + \sup_{\varphi \in \mathcal{F}} \phi(\varphi, \hat{\mathbf{Z}}) \\
&\leq \hat{L}(h', \hat{\mathbf{Z}}) + (M + \lambda\beta)\left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}\right\|_\infty \sqrt{\frac{\ln(|\mathcal{F}|/\delta)}{2mn}} \\
&= \hat{L}(h', \hat{\mathbf{Z}}) + (M + \lambda\beta)\left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}\right\|_\infty \sqrt{\frac{\ln\left(\mathcal{N}_\infty(\epsilon, \mathcal{H}, n)/\delta\right)}{2mn}}.
\end{aligned}$$

Combining this with Equations 5.3 and 5.4, with probability at least $1 - \delta$, every $h \in \mathcal{H}$ (and corresponding $h' \in \mathcal{H}'$), satisfies

$$\begin{aligned}
\overline{L}(h) &\leq |\mathcal{Y}|\, \lambda\, \epsilon + \overline{L}(h') \\
&\leq |\mathcal{Y}|\, \lambda\, \epsilon + \hat{L}(h', \hat{\mathbf{Z}}) + (M + \lambda\beta)\, \|\mathbf{\Gamma}^{\boldsymbol{\pi}}\|_{\infty} \sqrt{\frac{\ln\left(\mathcal{N}_{\infty}(\epsilon, \mathcal{H}, n)/\delta\right)}{2mn}} \\
&\leq 2\, |\mathcal{Y}|\, \lambda\, \epsilon + \hat{L}(h, \hat{\mathbf{Z}}) + (M + \lambda\beta)\, \|\mathbf{\Gamma}^{\boldsymbol{\pi}}\|_{\infty} \sqrt{\frac{\ln\left(\mathcal{N}_{\infty}(\epsilon, \mathcal{H}, n)/\delta\right)}{2mn}}.
\end{aligned}$$

Taking the infimum over ϵ completes the proof. ■

Clearly, if $\|\mathbf{\Gamma}^{\boldsymbol{\pi}}\|_{\infty} = \mathrm{O}(1)$, $\beta = \mathrm{O}(1)$, $\epsilon = \mathrm{O}(1/\sqrt{mn})$ and $\mathcal{N}_{\infty}(\epsilon, \mathcal{H}, n)$ is not too big, then the empirical risk will uniformly converge to the true risk. The dependency matrix is a function of the data distribution and cannot be controlled; however, the remaining conditions depend on the loss function and hypothesis class. In the following section, I will present an example of a loss function and hypothesis class for which these conditions hold.

5.3 Application of Covering Number Bound

In this section, I will apply Theorem 1 to derive risk bounds for collective classification by posterior decoding. Recall from Section 2.3.2 that posterior decoding selects the labels with the highest marginal probability under the model distribution. I will examine a specific class of templated MRFs that perform approximate marginal inference using a free energy that strongly convex with respect to the 1-norm. I will show that the strong

convexity of the inference objective enables two key results: (1) marginal inference in this class of models has $\mathrm{O}(1)$-uniform collective stability; (2) the covering number of the class is exponential in the number of weights, which is small because of templating.

Recall that each label, $y \in \mathcal{Y}$, is a basis vector, and the local marginals, $\mu \in [0,1]^{|\mathcal{Y}|} : \|\mu\|_1 = 1$, obey the simplex constraint. Therefore, to select the marginal probability of any label, one can simply take the dot product, $y \cdot \mu$. When using posterior decoding, a correct prediction requires that the correct label have the highest marginal probability. (To avoid tie-breaking, assume this must be a strict inequality.) Thus, a natural cost function is whether there is another label with equal or higher marginal probability:

$$c_0(y, \mu) \triangleq \mathbb{1}\left\{\langle y, \mu\rangle \leq \max_{y' \in \mathcal{Y}: y \neq y'} \langle y', \mu\rangle\right\}.$$

This is equivalent to the multiclass 0-1 cost for posterior decoding.

The problem with c_0 is that it fails to satisfy the second admissibility requirement for any finite λ. As such, to apply the risk bound, I will use a *ramp* cost:

$$c_\rho(y, \mu) \triangleq r_\rho\left(\langle y, \mu\rangle - \max_{y' \in \mathcal{Y}: y \neq y'} \langle y', \mu\rangle\right),$$

where $\rho \geq 0$ and

$$r_\rho(\delta) \triangleq \begin{cases} 1 & \text{for } \delta \leq 0, \\ 1 - \delta/\rho & \text{for } 0 < \delta < \rho, \\ 0 & \text{for } \delta \geq \rho. \end{cases}$$

Note that c_ρ generalizes c_0; the costs are equivalent when $\rho = 0$. Moreover, when $\rho > 0$,

c_ρ dominates c_0, and is admissible for a finite λ.

Lemma 3. *The ramp cost c_ρ is $(1, 1/\rho)$-admissible.*

The proof is deferred to Appendix C.2.

Let L_ρ denote the decomposable loss function for c_ρ. Note that L_0 is equivalent to the Hamming loss (introduced in Section 6.5.1) for posterior decoding. To avoid confusion with the structured ramp loss (introduced in Section 6.5.1.1), I will refer to L_ρ as the decomposable ramp loss. The following theorem upper-bounds the expected posterior decoding Hamming loss, $\overline{L}_0$, by the empirical decomposable ramp loss, $\hat{L}_\rho$.

Example 1. Fix any $m \geq 1$, $n \geq 1$, $\boldsymbol{\pi} \in \Pi(n)$, $\rho > 0$ and $\delta \in (0, 1)$. Fix a graph, $G \triangleq (\mathcal{V}, \mathcal{E})$, with maximum degree Δ_G. Assume that $\sup_{x \in \mathcal{X}} \|x\|_2 \leq 1$. Let $\mathcal{H}_{\text{sc}}$ denote the class of templated MRFs, with potentials defined in Section 2.3.4, that perform approximate marginal inference using a variational free energy whose conjugate function, $\tilde{\Phi}^*$, is κ-strongly convex w.r.t. the 1-norm. Let $d \triangleq |\mathbf{w}|$ denote the number of weights. Then, with probability at least $1 - \delta$ over realizations of $\hat{\mathbf{Z}} \triangleq (\mathbf{Z}^{(l)})_{l=1}^m$, for all $h \in \mathcal{H}_{\text{sc}}$ with $\|\mathbf{w}\|_2 \leq 1$,

$$\overline{L}_0(h) \leq \hat{L}_\rho(h, \hat{\mathbf{Z}}) + \frac{2}{\rho\sqrt{mn}} + \left(1 + \sqrt{\frac{2\Delta_G + 4}{\rho^2 \kappa}}\right) \|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}\|_\infty \sqrt{\frac{d \ln^+\left(\frac{2d\Delta_G m}{\kappa}\right) + \ln \frac{e}{\delta}}{2mn}},$$

where $\ln^+(\alpha) \triangleq \max\{0, \ln \alpha\}$.

This example is limited to weights in the unit hypercube, but could be extended to any bounded hypercube. Further, using a covering argument (similar to the ones used in Chapter 6), the bound can be made to hold simultaneously for all bounded hypercubes.

To prove Example 1, I will prove two technical lemmas. The first upper-bounds the uniform collective stability of the class $\mathcal{H}^1_{\mathrm{sc}} \triangleq \{h \in \mathcal{H}_{\mathrm{sc}} : \|\mathbf{w}\|_2 \leq 1\}$; the second upper-bounds the covering number of this class.

Lemma 4. *Fix a graph, $G \triangleq (\mathcal{V}, \mathcal{E})$, with maximum degree Δ_G, and assume that $\sup_{x \in \mathcal{X}} \|x\|_2 \leq 1$. Then the hypothesis class $\mathcal{H}^1_{\mathrm{sc}}$ has $\left(\sqrt{(2\Delta_G + 4)/\kappa}\right)$-uniform collective stability.*

Lemma 5. *Fix a graph, $G \triangleq (\mathcal{V}, \mathcal{E})$, with maximum degree Δ_G, and assume that $\sup_{x \in \mathcal{X}} \|x\|_2 \leq 1$. Let $d \triangleq |\mathbf{w}|$ denote the number of weights. Then, for any $\epsilon > 0$,*

$$\mathcal{N}_\infty(\epsilon, \mathcal{H}^1_{\mathrm{sc}}, n) \leq \left\lceil \left(\frac{2d\Delta_G}{\kappa n \left|\mathcal{Y}\right|^2 \epsilon^2} \right)^d \right\rceil.$$

Proofs are given in Appendices C.3 and C.4, respectively.

We can now prove Example 1.

Proof (Example 1) Since L_ρ dominates L_0 for all $\rho > 0$, it follows that $\overline{L}_\rho$ dominates $\overline{L}_0$. One can therefore apply Theorem 1 to $\overline{L}_\rho$, with $(M, \lambda) = (1, 1/\rho)$ (via Lemma 3), using Lemma 4 to upper-bound the uniform collective stability, and Lemma 5, with $\epsilon \triangleq (|\mathcal{Y}| \sqrt{mn})^{-1}$, for the covering number. We simplify the covering number using the fact that

$$\ln \lceil \alpha \rceil \leq \ln(\alpha + 1) \leq \ln^+(\alpha) + 1,$$

for any $\alpha \geq 0$. ■

5.4 Discussion

In this chapter, I presented a generalization bound based on the uniform collective stability and uniform covering number of the hypothesis class. The bound decreases as a function of both the number of examples, m, and the size of each example, n. I applied the bound to a class of templated MRFs that use strongly convex variational inference. The strong convexity of the inference objective enables good collective stability and low covering number. If the modulus of convexity does not decrease with n, and the data distribution exhibits suitably weak dependence, then the learning rate is $\tilde{\mathrm{O}}(1/\sqrt{mn})$, which is much sharper than previous bounds.

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

Example 1 applies to posterior decoders whose variational free energy is strongly convex with respect to the 1-norm. Unfortunately, this form of strong convexity is difficult to satisfy with an $\Omega(1)$ modulus. It is easier to prove $\Omega(1)$-strong convexity with respect to the 2-norm. (For instance, the strong convexity guarantees for variational inference discussed in Chapter 7 are all stated for the 2-norm.) Using the equivalence of norms identity,

$$\|\mathbf{u}\|_1 \leq \sqrt{|\mathbf{u}|}\,\|\mathbf{u}\|_2\,,$$

one can relate 2-norm strong convexity to 1-norm strong convexity. If a function, $\varphi : \mathbb{R}^n \to \mathbb{R}$, is κ-strongly convex with respect to the 2-norm, then φ is (κ/n)-strongly convex with respect to the 1-norm. However, for a conjugate function that is κ-strongly convex with respect to the 2-norm, using this identity yields $\Omega(\kappa/\,|G|)$-strong convexity with respect to the 1-norm; meaning, the modulus decreases with the size of the graph. Substituting this modulus into Example 1 yields a bound that is $\tilde{\mathrm{O}}\left(\sqrt{\frac{|G|}{\kappa mn}}\right)$, which is at least $\tilde{\mathrm{O}}\left(\sqrt{\frac{1}{\kappa m}}\right)$.

Then again, suppose the conjugate function were scaled by $|G|$. Then the free energy would be $(\kappa\,|G|)$-strongly convex with respect to the 2-norm; hence, κ-strongly convex with respect to the 1-norm, and Example 1 becomes $\tilde{\mathrm{O}}\left(\sqrt{\frac{1}{\kappa mn}}\right)$. Since the minimizer of the free energy is invariant to scaling,

$$\operatorname*{arg\,min}_{\tilde{\boldsymbol{\mu}}\in\tilde{\mathcal{M}}} \; -\mathbf{w}\cdot\mathbf{f}(\mathbf{x},\tilde{\boldsymbol{\mu}}) + |G|\,\tilde{\Phi}^*(\tilde{\boldsymbol{\mu}}) = \operatorname*{arg\,min}_{\tilde{\boldsymbol{\mu}}\in\tilde{\mathcal{M}}} \; -\mathbf{w}\cdot\frac{\mathbf{f}(\mathbf{x},\tilde{\boldsymbol{\mu}})}{|G|} + \tilde{\Phi}^*(\tilde{\boldsymbol{\mu}}).$$

Via Lemma 14 (in Appendix A), the features, $\mathbf{f}(\mathbf{x},\tilde{\boldsymbol{\mu}})$, have norm at most $|G|$. Therefore,

scaling up $\tilde{\Phi}^*$ is equivalent to scaling down the features—which seems natural, given that the graph could be arbitrarily large, so the unscaled features could have arbitrarily large norm. Indeed, feature scaling is common practice in machine learning, and has been shown to improve generalization for certain models (e.g., Herbrich and Graepel, 2001; Juszczak et al., 2002; Graf et al., 2003). The above argument suggests that rescaling the features can improve generalization for posterior decoding in MRFs.

How does this modification affect marginal inference? The conjugate function (which is typically a negative entropy) acts as a regularizer, favoring pseudomarginals that are more uniform. Thus, increasing the influence of this function in the free energy minimization has a "flattening," or "smoothing," effect. The posterior decoding is not affected by this flattening, but the decomposable ramp loss, L_ρ, may be.

Chapter 6: Generalization Bounds via Local Stability

The generalization bounds of the previous chapter rely on collective stability. This analysis can be limiting, since it only accommodates element-wise loss functions. To handle more sophisticated loss functions (such as those discussed in Section 6.5) requires the more general definitions of stability from Chapter 3. Another benefit of using these definitions is that they accommodate hypothesis classes that do not satisfy *uniform* stability, but satisfy *local* stability.

In this chapter, I present generalization bounds based on local stability using the *PAC-Bayes framework*. PAC-Bayes is an analytical framework in which prediction is stochastic. Let $\mathbb{P}$ denote a predetermined prior distribution on $\mathcal{H}$, and let $\mathbb{Q}$ denote a posterior distribution, the parameters of which are typically learned from training data. Given a realized input, $\mathbf{x} \in \mathcal{X}^n$, one first draws a hypothesis, $h \in \mathcal{H}$, according to $\mathbb{Q}$, then computes the prediction, $h(\mathbf{x})$. Since prediction in the PAC-Bayes framework is randomized, the loss quantities become expectations over draws of h, which I denote by

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

respectively. A typical analysis involves upper-bounding the difference of these quantities (with high probability), then "derandomizing" the bounds so that they hold for a learned,

deterministic predictor.

PAC-Bayesian analysis is particularly well suited for the local stability condition described in Definition 5. With prior knowledge of the hypothesis class (and data distribution), a posterior can be constructed so as to place low mass on predictors that do not satisfy good stability. As shown in Section 6.5, this technique lets one relax certain restrictions on the hypothesis class and data domain.

The remainder of this chapter is structured as follows. First, I provide a sketch of my PAC-Bayesian approach, to introduce the key techniques used in the following proofs. I then present two PAC-Bayesian theorems for structured prediction based on local stability. The first theorem is stated for a given stability parameter, β. I then generalize this to hold for all β simultaneously, meaning β can depend on the posterior. I also introduce a novel technique to derandomize the bounds based on the stability of the loss function. To demonstrate the application of the bounds, I give several examples, using max-margin and soft-max training as motivation. Like the application presented in the previous chapter, the new bounds also tighten as the effective size of the training set (number of examples times size of each example) increases.

6.1 Analysis Sketch

The following is a high-level sketch of my PAC-Bayesian analysis, which I will specialize to various settings in Sections 6.2 and 6.3. It will help to view the composition of the loss function, L, and the hypothesis class, $\mathcal{H}$, as a family of functions, $L \circ \mathcal{H} = \{L(h, \cdot) : h \in \mathcal{H}\}$. If $\mathbb{Q}$ is a distribution on $\mathcal{H}$, it is also a distribution on $L \circ \mathcal{H}$. Each member of

$L \circ \mathcal{H}$ is a random function, determined by the draw of $h \sim \mathbb{Q}$. Further, when $L(h, \cdot)$ is composed with a training set $\hat{\mathbf{Z}} \sim \mathbb{D}^m$ in $\hat{L}(h, \cdot)$, the generalization error, $\overline{L}(h) - \hat{L}(h, \hat{\mathbf{Z}})$, becomes a centered random variable. Part of the analysis involves bounding the moment-generating function of this random variable, and to do so requires the notions of stability from Chapter 3. The stability of $L(h, \cdot)$ is determined by h, so the "bad" members of $L \circ \mathcal{H}$ are in fact the "bad" hypotheses (for the given loss function).

Let $\hat{\mathbf{Z}} \triangleq (\mathbf{Z}^{(l)})_{l=1}^m$ denote a training set of m structured examples, distributed according to $\mathbb{D}^m$. Fix some $\beta \geq 0$ and a set of bad inputs $\mathcal{B}_{\mathcal{Z}}$, with measure $\nu \triangleq \mathbb{D}(\mathcal{B}_{\mathcal{Z}})$. Implicitly, the pair $(\beta, \mathcal{B}_{\mathcal{Z}})$ fixes a set of hypotheses $\mathcal{B}_{\mathcal{H}} \subseteq \mathcal{H}$ for which $L(h, \cdot)$ does not satisfy Equation 3.1 with $\beta' \triangleq \beta/n$ and $\mathcal{B}_{\mathcal{Z}}$. For the time being, $\mathcal{B}_{\mathcal{H}}$ is independent of $\mathbb{Q}$. Fix a prior $\mathbb{P}$ and posterior $\mathbb{Q}$ on $\mathcal{H}$. (We will later consider all posteriors.) Define a convenience function,

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

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

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

This identity can be used to show that, if $L \circ \mathcal{H}$ is α-uniformly range-bounded, and $\mathbb{Q}$ is

$(\beta/n, \mathcal{B}_{\mathcal{Z}}, \eta)$-locally stable, then

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

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

Lemma 6. *For any measurable function $\varphi : \Omega \to \mathbb{R}$, and any two distributions, $\mathbb{P}$ and $\mathbb{Q}$, on Ω,*

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

(McAllester (2003) provides a straightforward proof.) Using Lemma 6, for any free parameter $u \geq 0$, we have that

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

Combining the above inequalities yields

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

The remainder of the analysis concerns how to bound $\mathbb{E}_{h\sim\mathbb{P}}\left[e^{u\tilde{\phi}(h, \hat{\mathbf{Z}})}\right]$ and how to optimize u. For the first task, one combines Markov's inequality with the moment-generating function bound from Chapter 4. Optimizing u takes some care, since the bounds should hold simultaneously for all posteriors. To do so, I will adopt a discretization technique (Seldin et al., 2012) that approximately optimizes the bound for all poste-

riors. I use a similar technique to obtain bounds that hold for all β.

6.2 Fixed Stability Bounds

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

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

$$\overline{L}(\mathbb{Q}) \leq \hat{L}(\mathbb{Q}, \hat{\mathbf{Z}}) + \alpha(\eta + \nu) + 2\beta \left\| \boldsymbol{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}} \right\|_{\infty} \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln \frac{2}{\delta}}{2mn}}. \tag{6.1}$$

To interpret the bound, suppose $\alpha = \mathrm{O}(1)$, $\beta = \mathrm{O}(1)$, and that the data distribution is weakly dependent, with $\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}\|_{\infty} = \mathrm{O}(1)$. We would then have that the generalization error decreases at a rate of $\mathrm{O}\left(\eta + \nu + (mn)^{-1/2}\right)$. Since η is a function of the posterior, we can reasonably assume that $\eta = \mathrm{O}\left((mn)^{-1/2}\right)$. (Section 6.5 provides examples of this.) However, while ν may be proportional to n, it is unreasonable to believe that ν will decrease with m, since $\mathbb{D}$ is almost certainly agnostic to the number of training examples.

Thus, Theorem 2 is interesting when either ν is negligible, or when m is a small constant.

It can be shown that any hypothesis class with *collective* stability, composed with an element loss function and admissible cost function (see Section 5.2), satisfies the conditions of the bound. Thus, Theorem 2 is strictly more general than my prior PAC-Bayes bounds (London et al., 2014). Moreover, Theorem 2 easily applies to compositions with uniform stability, since $\mathbb{Q}(\mathcal{B}_{\mathcal{H}}) = 0$ for all posteriors. This insight yields the following corollary.

Corollary 2. *Suppose $L \circ \mathcal{H}$ is (β/n)-uniformly stable. Then, with probability at least $1 - \delta$ over realizations of $\hat{\mathbf{Z}}$, for all $\mathbb{Q}$,*

$$\overline{L}(\mathbb{Q}) \leq \hat{L}(\mathbb{Q}, \hat{\mathbf{Z}}) + 2\beta \left\| \mathbf{\Gamma}^{\boldsymbol{\pi}} \right\|_{\infty} \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q} \| \mathbb{P}) + \ln \frac{2}{\delta}}{2mn}}. \tag{6.2}$$

As shown in Section 6.5.1.2, Corollary 2 is useful when the hypothesis class and instance space are uniformly bounded. Even when this property does not hold, we obtain an identical bound for all posteriors with $(\beta/n, \emptyset, 0)$-local stability, meaning the support of the posterior is (β/n)-uniformly stable. However, this condition is less useful, since it is assumed that the posterior construction puts nonzero density on a learned hypothesis, which may not satisfy uniform stability for a fixed β.

I now prove Theorem 2.

Proof (Theorem 2) Begin by defining two convenience functions,

$$\phi(h, \hat{\mathbf{Z}}) \triangleq \overline{L}(h) - \hat{L}(h, \hat{\mathbf{Z}}) \tag{6.3}$$

$$\text{and} \quad \tilde{\phi}(h, \hat{\mathbf{Z}}) \triangleq \begin{cases} \mathbb{E}_{\mathbf{Z}\sim\mathbb{D}}\left[L(h, \mathbf{Z}) \,|\, \overline{\mathcal{B}}\right] - \hat{L}(h, \hat{\mathbf{Z}}) & \text{if } h \notin \mathcal{B}_{\mathcal{H}}, \\ 0 & \text{otherwise,} \end{cases} \tag{6.4}$$

If $L \circ \mathcal{H}$ is α-uniformly range-bounded (Definition 6), then, for any $h \in \mathcal{H}$,

$$\begin{aligned} \phi(h, \hat{\mathbf{Z}}) &= \frac{1}{m}\sum_{l=1}^{m} \overline{L}(h) - L(h, \mathbf{Z}^{(l)}) \\ &\leq \frac{1}{m}\sum_{l=1}^{m} \sup_{\mathbf{z}\in\mathcal{Z}^n} \left| L(h, \mathbf{z}) - L(h, \mathbf{Z}^{(l)}) \right| \\ &\leq \frac{1}{m}\sum_{l=1}^{m} \alpha = \alpha. \end{aligned} \tag{6.5}$$

It follows that

$$\begin{aligned} \phi(h, \hat{\mathbf{Z}}) &= \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathbb{D}}\left[L(h, \mathbf{Z}) - \hat{L}(h, \hat{\mathbf{Z}})\right] \\ &= \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathbb{D}}\left[\left(L(h, \mathbf{Z}) - \hat{L}(h, \hat{\mathbf{Z}})\right) \mathbb{1}\{\mathbf{Z} \notin \mathcal{B}_{\mathcal{Z}}\}\right] + \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathbb{D}}\left[\left(L(h, \mathbf{Z}) - \hat{L}(h, \hat{\mathbf{Z}})\right) \mathbb{1}\{\mathbf{Z} \in \mathcal{B}_{\mathcal{Z}}\}\right] \\ &\leq \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathbb{D}}\left[\left(L(h, \mathbf{Z}) - \hat{L}(h, \hat{\mathbf{Z}})\right) \mathbb{1}\{\mathbf{Z} \notin \mathcal{B}_{\mathcal{Z}}\}\right] + \alpha \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathbb{D}}\left[\mathbb{1}\{\mathbf{Z} \in \mathcal{B}_{\mathcal{Z}}\}\right] \\ &\leq \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathbb{D}}\left[\left(L(h, \mathbf{Z}) - \hat{L}(h, \hat{\mathbf{Z}})\right) \mathbb{1}\{\mathbf{Z} \notin \mathcal{B}_{\mathcal{Z}}\}\right] + \alpha\nu \\ &= \Pr_{\mathbf{Z}\sim\mathbb{D}}\{\mathbf{Z} \notin \mathcal{B}_{\mathcal{Z}}\}\left(\mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathbb{D}}\left[L(h, \mathbf{Z}) \,|\, \overline{\mathcal{B}}\right] - \hat{L}(h, \hat{\mathbf{Z}})\right) + \alpha\nu \\ &\leq \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathbb{D}}\left[L(h, \mathbf{Z}) \,|\, \overline{\mathcal{B}}\right] - \hat{L}(h, \hat{\mathbf{Z}}) + \alpha\nu. \end{aligned} \tag{6.6}$$

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

$$\begin{aligned}
\overline{L}(\mathbb{Q}) - \hat{L}(\mathbb{Q}, \hat{\mathbf{Z}}) &= \mathop{\mathbb{E}}_{h\sim\mathbb{Q}}\left[\phi(h, \hat{\mathbf{Z}})\right] \\
&= \mathop{\mathbb{E}}_{h\sim\mathbb{Q}}\left[\phi(h, \hat{\mathbf{Z}})\,\mathbb{1}\{h \in \mathcal{B}_{\mathcal{H}}\}\right] + \mathop{\mathbb{E}}_{h\sim\mathbb{Q}}\left[\phi(h, \hat{\mathbf{Z}})\,\mathbb{1}\{h \notin \mathcal{B}_{\mathcal{H}}\}\right] \\
&\leq \alpha \mathop{\mathbb{E}}_{h\sim\mathbb{Q}}\left[\mathbb{1}\{h \in \mathcal{B}_{\mathcal{H}}\}\right] + \mathop{\mathbb{E}}_{h\sim\mathbb{Q}}\left[\phi(h, \hat{\mathbf{Z}})\,\mathbb{1}\{h \notin \mathcal{B}_{\mathcal{H}}\}\right] \\
&\leq \alpha\eta + \mathop{\mathbb{E}}_{h\sim\mathbb{Q}}\left[\phi(h, \hat{\mathbf{Z}})\,\mathbb{1}\{h \notin \mathcal{B}_{\mathcal{H}}\}\right] \\
&\leq \alpha\eta + \alpha\nu + \mathop{\mathbb{E}}_{h\sim\mathbb{Q}}\left[\tilde{\phi}(h, \hat{\mathbf{Z}})\right]. \qquad (6.7)
\end{aligned}$$

Then, for any $u \in \mathbb{R}$, using Lemma 6, we have that

$$\begin{aligned}
\overline{L}(\mathbb{Q}) - \hat{L}(\mathbb{Q}, \hat{\mathbf{Z}}) &\leq \alpha\eta + \alpha\nu + \frac{1}{u}\mathop{\mathbb{E}}_{h\sim\mathbb{Q}}\left[u\,\tilde{\phi}(h, \hat{\mathbf{Z}})\right] \\
&\leq \alpha\eta + \alpha\nu + \frac{1}{u}\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln \mathop{\mathbb{E}}_{h\sim\mathbb{P}}\left[e^{u\tilde{\phi}(h,\hat{\mathbf{Z}})}\right]\right). \qquad (6.8)
\end{aligned}$$

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

$$u_j \triangleq 2^j \sqrt{\frac{8mn \ln \frac{2}{\delta}}{\beta^2 \left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}\right\|_\infty^2}}, \qquad (6.9)$$

and define an event,

$$E_j \triangleq \mathbb{1}\left\{\mathop{\mathbb{E}}_{h\sim\mathbb{P}}\left[e^{u_j\tilde{\phi}(h,\hat{\mathbf{Z}})}\right] \geq \frac{1}{\delta_j}\exp\left(\frac{u_j^2\beta^2\left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}\right\|_\infty^2}{8mn}\right)\right\}. \qquad (6.10)$$

Note that u_j and E_j are independent of $(\eta, \mathbb{Q})$, since β (hence, $\mathcal{B}_{\mathcal{H}}$) is fixed. Let $E \triangleq$

$\bigcup_{j=0}^{\infty} E_j$ denote the event that any E_j occurs. Also, define an event

$$B \triangleq \bigcup_{l=1}^{m} \mathbb{1}\left\{\mathbf{Z}^{(l)} \in \mathcal{B}_{\mathcal{Z}}\right\}, \tag{6.11}$$

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

$$\begin{aligned}
\Pr_{\hat{\mathbf{Z}}\sim\mathbb{D}^m}\{B \cup E\} &= \Pr_{\hat{\mathbf{Z}}\sim\mathbb{D}^m}\{B\} + \Pr_{\hat{\mathbf{Z}}\sim\mathbb{D}^m}\{E \cap \neg B\} \\
&\leq \Pr_{\hat{\mathbf{Z}}\sim\mathbb{D}^m}\{B\} + \Pr_{\hat{\mathbf{Z}}\sim\mathbb{D}^m}\{E \,|\, \neg B\} \\
&\leq \sum_{l=1}^{m} \Pr_{\mathbf{Z}^{(l)}\sim\mathbb{D}}\{\mathbf{Z}^{(l)} \in \mathcal{B}_{\mathcal{Z}}\} + \sum_{j=0}^{\infty} \Pr_{\hat{\mathbf{Z}}\sim\mathbb{D}^m}\{E_j \,|\, \neg B\} \\
&\leq m\nu + \sum_{j=0}^{\infty} \Pr_{\hat{\mathbf{Z}}\sim\mathbb{D}^m}\{E_j \,|\, \neg B\}.
\end{aligned} \tag{6.12}$$

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

$$\Pr_{\hat{\mathbf{Z}}\sim\mathbb{D}^m}\{E_j \,|\, \neg B\} \leq \delta_j \exp\left(-\frac{u_j^2\beta^2 \left\|\mathbf{\Gamma}_{\overline{\mathcal{B}}}^{\boldsymbol{\pi}}\right\|_{\infty}^2}{8mn}\right) \mathop{\mathbb{E}}_{h\sim\mathbb{P}} \mathop{\mathbb{E}}_{\hat{\mathbf{Z}}\sim\mathbb{D}^m}\left[e^{u_j\tilde{\phi}(h,\hat{\mathbf{Z}})} \,|\, \neg B\right]. \tag{6.13}$$

Let

$$\varphi(h, \mathbf{Z}) \triangleq \begin{cases} \frac{1}{m}\left(\mathbb{E}_{\mathbf{Z}'\sim\mathbb{D}}\left[L(h, \mathbf{Z}') \,|\, \overline{\mathcal{B}}\right] - L(h, \mathbf{Z})\right) & \text{if } h \notin \mathcal{B}_{\mathcal{H}}, \\ 0 & \text{otherwise,} \end{cases} \tag{6.14}$$

and note that $\tilde{\phi}(h, \hat{\mathbf{Z}}) = \sum_{l=1}^{m} \varphi(h, \mathbf{Z}^{(l)})$. Then, since $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(m)}$ are independent and

identically distributed, one can write the inner expectation over $\hat{\mathbf{Z}}$ as

$$\begin{aligned}
\mathop{\mathbb{E}}_{\hat{\mathbf{Z}}\sim\mathbb{D}^m}\left[e^{u_j\tilde{\phi}(h,\hat{\mathbf{Z}})} \mid \neg B\right] &= \prod_{l=1}^{m} \mathop{\mathbb{E}}_{\mathbf{Z}^{(l)}\sim\mathbb{D}}\left[e^{u_j\varphi(h,\mathbf{Z}^{(l)})} \mid \neg B\right] \\
&= \prod_{l=1}^{m} \mathop{\mathbb{E}}_{\mathbf{Z}^{(l)}\sim\mathbb{D}}\left[e^{u_j\varphi(h,\mathbf{Z}^{(l)})} \mid \mathbf{Z}^{(l)} \notin \mathcal{B}_{\mathcal{Z}}\right] \\
&= \prod_{l=1}^{m} \mathop{\mathbb{E}}_{\mathbf{Z}^{(l)}\sim\mathbb{D}}\left[e^{u_j\varphi(h,\mathbf{Z}^{(l)})} \mid \overline{\mathcal{B}}\right]. \qquad (6.15)
\end{aligned}$$

By construction, $\varphi(h,\cdot)$ outputs zero whenever $h \in \mathcal{B}_{\mathcal{H}}$. In these cases, $\varphi(h,\cdot)$ trivially satisfies uniform stability, which implies local stability. Further, if $\mathbb{Q}$ is $(\beta/n, \mathcal{B}_{\mathcal{Z}}, \eta)$-locally stable, then every $L(h,\cdot) : h \notin \mathcal{B}_{\mathcal{H}}$ is $(\beta/n, \mathcal{B}_{\mathcal{Z}})$-locally stable, and it is easily verified that $\varphi(h,\cdot) : h \notin \mathcal{B}_{\mathcal{H}}$ is $(\beta/(mn), \mathcal{B}_{\mathcal{Z}})$-locally stable. Thus, $\varphi(h,\cdot) : h \in \mathcal{H}$ is $(\beta/(mn), \mathcal{B}_{\mathcal{Z}})$-locally stable. Since $\mathbb{E}_{\mathbf{Z}\sim\mathbb{D}}[\varphi(h,\mathbf{Z}) \mid \overline{\mathcal{B}}] = 0$, we therefore apply Proposition 1 and have, for all $h \in \mathcal{H}$,

$$\mathop{\mathbb{E}}_{\mathbf{Z}^{(l)}\sim\mathbb{D}}\left[e^{u_j\varphi(h,\mathbf{Z}^{(l)})} \mid \overline{\mathcal{B}}\right] \leq \exp\left(\frac{u_j^2\beta^2\left\|\mathbf{\Gamma}_{\overline{\mathcal{B}}}^{\boldsymbol{\pi}}\right\|_\infty^2}{8m^2n}\right). \qquad (6.16)$$

Combining Equations 6.13, 6.15 and 6.16, we have that

$$\begin{aligned}
\Pr_{\hat{\mathbf{Z}}\sim\mathbb{D}^m}\{E_j \mid \neg B\} &\leq \delta_j \exp\left(-\frac{u_j^2\beta^2\left\|\mathbf{\Gamma}_{\overline{\mathcal{B}}}^{\boldsymbol{\pi}}\right\|_\infty^2}{8mn}\right) \mathop{\mathbb{E}}_{h\sim\mathbb{P}}\left[\prod_{l=1}^{m} \mathop{\mathbb{E}}_{\mathbf{Z}^{(l)}\sim\mathbb{D}}\left[e^{u_j\varphi(h,\mathbf{Z}^{(l)})} \mid \overline{\mathcal{B}}\right]\right] \\
&\leq \delta_j \exp\left(-\frac{u_j^2\beta^2\left\|\mathbf{\Gamma}_{\overline{\mathcal{B}}}^{\boldsymbol{\pi}}\right\|_\infty^2}{8mn}\right) \mathop{\mathbb{E}}_{h\sim\mathbb{P}}\left[\prod_{l=1}^{m} \exp\left(\frac{u_j^2\beta^2\left\|\mathbf{\Gamma}_{\overline{\mathcal{B}}}^{\boldsymbol{\pi}}\right\|_\infty^2}{8m^2n}\right)\right] \\
&= \delta_j \exp\left(-\frac{u_j^2\beta^2\left\|\mathbf{\Gamma}_{\overline{\mathcal{B}}}^{\boldsymbol{\pi}}\right\|_\infty^2}{8mn}\right) \exp\left(\frac{u_j^2\beta^2\left\|\mathbf{\Gamma}_{\overline{\mathcal{B}}}^{\boldsymbol{\pi}}\right\|_\infty^2}{8mn}\right) = \delta_j. \qquad (6.17)
\end{aligned}$$

Then, combining Equations 6.12 and 6.17, and using the geometric series identity, we

have that

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

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

$$\mathop{\mathbb{E}}_{h\sim\mathbb{P}}\left[e^{u_j\tilde{\phi}(h,\hat{\mathbf{Z}})}\right]\leq\frac{1}{\delta_j}\exp\left(\frac{u_j^2\beta^2\left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}\right\|_\infty^2}{8mn}\right). \tag{6.18}$$

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

$$j^\star\triangleq\left\lfloor\frac{1}{2\ln 2}\ln\left(\frac{D_{\text{KL}}(\mathbb{Q}\|\mathbb{P})}{\ln(2/\delta)}+1\right)\right\rfloor, \tag{6.19}$$

and note that $j^\star\geq 0$. For all $v\in\mathbb{R}$, we have that $v-1\leq\lfloor v\rfloor\leq v$, and $2^{\ln v}=v^{\ln 2}$. We can apply these identities to Equation 6.19 to show that

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

implying

$$\sqrt{\frac{2mn\left(D_{\text{KL}}(\mathbb{Q}\|\mathbb{P})+\ln\frac{2}{\delta}\right)}{\beta^2\left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}\right\|_\infty^2}}\leq u_{j^\star}\leq\sqrt{\frac{8mn\left(D_{\text{KL}}(\mathbb{Q}\|\mathbb{P})+\ln\frac{2}{\delta}\right)}{\beta^2\left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}\right\|_\infty^2}}. \tag{6.20}$$

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

$$\begin{aligned} D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{1}{\delta_{j^\star}} &= D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta} + j^\star \ln 2 \\ &\leq D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta} + \frac{\ln 2}{2\ln 2}\ln\left(\frac{D_{\text{KL}}(\mathbb{Q}\|\mathbb{P})}{\ln(2/\delta)} + 1\right) \\ &= D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta} + \frac{1}{2}\ln\left(D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta}\right) - \frac{1}{2}\ln\ln\frac{2}{\delta} \\ &\leq D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta} + \frac{1}{2}\left(D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta}\right), \end{aligned} \tag{6.21}$$

for all $\delta \in (0,1)$. It can be shown that this is approximately optimal, in that the bound is at most twice what it would be for a fixed posterior.

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

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

The first inequality substitutes $u_{j^\star}$ into Equation 6.8; the second uses Equation 6.18; the third is from Equation 6.21; and the last uses the lower and upper bounds from Equation 6.20. ■

6.3 Posterior-Dependent Stability

In Theorem 2, I required β to be fixed *a priori*, meaning the user must pre-specify a desired stability. In this section, I prove bounds that hold for all $\beta \geq 1$ simultaneously, meaning the value of β can depend on the learned posterior. (The requirement of nonnegativity is not restrictive, since stability with $\beta \leq 1$ implies stability with $\beta = 1$.)

Theorem 3. *Fix $m \geq 1$, $n \geq 1$, $\boldsymbol{\pi} \in \Pi(n)$, $\delta \in (0,1)$ and $\alpha \geq 0$. Fix a distribution, $\mathbb{D}$, on $\mathcal{Z}^n$. Fix a set of bad inputs, $\mathcal{B}_{\mathcal{Z}}$, with $\nu \triangleq \mathbb{D}(\mathcal{B}_{\mathcal{Z}})$, and let $\overline{\mathcal{B}}$ denote the σ-algebra of $\mathbf{Z} \notin \mathcal{B}_{\mathcal{Z}}$. Let $\boldsymbol{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}}$ denote the dependency matrix induced by $\mathbb{D}$, $\boldsymbol{\pi}$ and $\overline{\mathcal{B}}$. Fix a prior, $\mathbb{P}$, on a hypothesis class, $\mathcal{H}$. Fix a loss function, L, such that $L \circ \mathcal{H}$ is α-uniformly range-bounded. Then, with probability at least $1 - \delta - m\nu$ over realizations of $\hat{\mathbf{Z}} \triangleq (\mathbf{Z}^{(l)})_{l=1}^m$, drawn according to $\mathbb{D}^m$, the following hold: 1) for all $l \in [m]$, $\mathbf{Z}^{(l)} \notin \mathcal{B}_{\mathcal{Z}}$; 2) for all $\beta \geq 1$, $\eta \in [0,1]$ and posteriors $\mathbb{Q}$ with $(\beta/n, \mathcal{B}_{\mathcal{Z}}, \eta)$-local stability,*

$$\overline{L}(\mathbb{Q}) \leq \hat{L}(\mathbb{Q}, \hat{\mathbf{Z}}) + \alpha(\eta + \nu) + 4\beta \left\| \boldsymbol{\Gamma}^{\boldsymbol{\pi}}_{\overline{\mathcal{B}}} \right\|_\infty \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln \frac{4}{\delta} + \ln \beta}{2mn}}. \tag{6.22}$$

The proof is similar to that of Theorem 2, so I defer it to Appendix D.1.

Theorem 3 immediately yields the following corollary by taking $\mathcal{B}_{\mathcal{Z}} \triangleq \emptyset$.

Corollary 3. *With probability at least $1-\delta$ over realizations of $\hat{\mathbf{Z}}$, for all $\beta \geq 1$, $\eta \in [0,1]$ and $\mathbb{Q}$ with $(\beta/n, \emptyset, \eta)$-local stability,*

$$\overline{L}(\mathbb{Q}) \leq \hat{L}(\mathbb{Q}, \hat{\mathbf{Z}}) + \alpha\eta + 4\beta \left\| \boldsymbol{\Gamma}^{\boldsymbol{\pi}} \right\|_\infty \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln \frac{4}{\delta} + \ln \beta}{2mn}}. \tag{6.23}$$

In Section 6.5, I apply this corollary to unbounded hypothesis classes, with bounded instance spaces. Corollary 3 trivially implies a bound for posteriors with $(\beta/n, \emptyset, 0)$-local stability, such as those with bounded support on an unbounded hypothesis class, where β may depend on a learned model.

6.4 Derandomizing the Loss using Stability

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

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

Definition 15. Fix a hypothesis class, $\mathcal{H}$, a set of inputs, $\mathcal{B}_{\mathcal{Z}} \subseteq \mathcal{Z}^n$, a loss function, L, and a posterior, $\mathbb{Q}$. The pair $(L, \mathbb{Q})$ has $(\lambda, \mathcal{B}_{\mathcal{Z}}, \eta)$-local stability if, for any $h \in \mathcal{H}$ and $\mathbf{z} \notin \mathcal{B}_{\mathcal{Z}}$,

there exists a set $\mathcal{B}_{\mathcal{H}}(h, \mathbf{z}) \subseteq \mathcal{H}$ such that $\mathbb{Q}_h(\mathcal{B}_{\mathcal{H}}(h, \mathbf{z})) \leq \eta$ and, for all $h' \notin \mathcal{B}_{\mathcal{H}}(h, \mathbf{z})$,

$$|L(h, \mathbf{z}) - L(h', \mathbf{z})| \leq \lambda. \tag{6.24}$$

This form of stability is a slightly weaker condition than the previous definitions, in that each input, $(h, \mathbf{z})$, has its own "bad" set, $\mathcal{B}_{\mathcal{H}}(h, \mathbf{z})$. This distinction means that "badness" is relative, whereas, in Definitions 4 and 5, it is absolute.

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

$$\left|\overline{L}(h) - \overline{L}(\mathbb{Q}_h)\right| \leq \alpha(\eta + \nu) + \lambda, \tag{6.25}$$

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

$$\left|\hat{L}(h, \hat{\mathbf{z}}) - \hat{L}(\mathbb{Q}_h, \hat{\mathbf{z}})\right| \leq \alpha\eta + \lambda. \tag{6.26}$$

Proof Define a convenience function

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

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

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

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

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

thus proving Equation 6.26. Furthermore,

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

which proves Equation 6.25. ■

Proposition 4 can easily be combined with the PAC-Bayes bounds from the previ-

ous sections to obtain derandomized generalization bounds. I analyze some examples in Section 6.5.

6.4.1 Normed Vector Spaces

When the hypothesis class is a normed vector space (as is the case in all of the examples in Section 6.5), Definition 15 can be decomposed into properties of the loss function and posterior separately.

Definition 16. Fix a hypothesis class, $\mathcal{H}$, equipped with a norm, $\|\cdot\|$. Fix a set of inputs, $\mathcal{B}_{\mathcal{Z}} \subseteq \mathcal{Z}^n$. A loss function, L, has $(\lambda, \mathcal{B}_{\mathcal{Z}})$-*local hypothesis stability* if, for all $h, h' \in \mathcal{H}$ and $\mathbf{z} \notin \mathcal{B}_{\mathcal{Z}}$,

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

Definition 17. Fix a hypothesis class, $\mathcal{H}$, equipped with a norm, $\|\cdot\|$. A posterior, $\mathbb{Q}$, has (β, η)-*local hypothesis stability* if, for any $h \in \mathcal{H}$, there exists a set $\mathcal{B}_{\mathcal{H}}(h) \subseteq \mathcal{H}$ such that $\mathbb{Q}_h(\mathcal{B}_{\mathcal{H}}(h)) \leq \eta$ and, for all $h' \notin \mathcal{B}_{\mathcal{H}}(h)$, $\|h - h'\| \leq \beta$.

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

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

The proof is provided in Appendix D.2.

6.5 Example Applications

In this section, I apply the PAC-Bayes bounds to two popular training regimes, *max-margin* and *soft-max* learning, under various assumptions about the instance space and hypothesis class. This illustrates how various modeling decisions affect the generalization error. The results in this section are stated in terms of a deterministic predictor. I use the PAC-Bayes framework as an analytic tool only. However, the derandomized bounds can be adapted for a randomized predictor, and can in fact be made considerably tighter.

6.5.1 Max-Margin Learning

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

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

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

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

Structured predictors learned with a max-margin objective are alternatively referred

to as *max-margin Markov networks* (Taskar et al., 2004) or *StructSVM* (Tsochantaridis et al., 2005), depending on the form of the hinge loss. In this section, I consider the former formulation, defining the structured hinge loss as

$$L_{\text{h}}(h, \mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n} \left(\max_{\mathbf{y}' \in \mathcal{Y}^n} D_{\text{H}}(\mathbf{y}, \mathbf{y}') + h(\mathbf{x}, \mathbf{y}') - h(\mathbf{x}, \mathbf{y}) \right), \tag{6.27}$$

where

$$h(\mathbf{x}, \mathbf{y}) \triangleq \boldsymbol{\theta}(\mathbf{x}; \mathbf{w}) \cdot \hat{\mathbf{y}} \tag{6.28}$$

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

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

$$h(\mathbf{x}) \triangleq \operatorname*{arg\,max}_{\mathbf{y} \in \mathcal{Y}^n} h(\mathbf{x}, \mathbf{y}) = \operatorname*{arg\,max}_{\mathbf{y} \in \mathcal{Y}^n} p\left(\mathbf{Y} = \mathbf{y} \,|\, \mathbf{X} = \mathbf{x}; \mathbf{w}\right) \tag{6.29}$$

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

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

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

$$\tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}) \triangleq \boldsymbol{\theta}(\mathbf{x}; \mathbf{w}) + \delta(\mathbf{y}), \tag{6.30}$$

computing the hinge loss is equivalent to performing *loss-augmented* MAP inference with $\tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w})$. Provided inference can be computed efficiently with the given class of models, so too can the hinge loss.[1]

6.5.1.1 Structured Ramp Loss

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

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

where $h(\mathbf{x}, \mathbf{y})$ is defined in Equation 6.28. The ramp loss is 1-uniformly range-bounded. Further, when $h(\mathbf{x})$ performs MAP inference (Equation 6.29),

$$L_{\mathrm{H}}(h, \mathbf{x}, \mathbf{y}) \leq L_{\mathrm{r}}(h, \mathbf{x}, \mathbf{y}) \leq L_{\mathrm{h}}(h, \mathbf{x}, \mathbf{y}). \tag{6.31}$$

Thus, one can analyze the generalization properties of the ramp loss to obtain bounds for the difference of the expected Hamming loss and empirical hinge loss. To distin-

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

guish quantities of different loss functions, I will use a subscript notation; e.g., $\overline{L}_{\mathrm{H}}$ is the expected Hamming loss, and $\hat{L}_{\mathrm{h}}$ is the empirical hinge loss.

Using the templated, linear potentials defined in Section 2.3.4, one obtains two technical lemmas for the structured ramp loss. Proofs are provided in Appendices D.3 and D.4.

Lemma 7. *Fix any $p, q \geq 1$ such that $1/p + 1/q = 1$. Fix a graph, $G \triangleq (\mathcal{V}, \mathcal{E})$, with maximum degree Δ_G. Assume that $\sup_{x \in \mathcal{X}} \|x\|_p \leq R$. Then, for any MRF h with weights $\mathbf{w}$, and any $\mathbf{z}, \mathbf{z} \in \mathcal{Z}^n$, where $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ and $\mathbf{z}' = (\mathbf{x}', \mathbf{y}')$,*

$$|L_r(h, \mathbf{z}) - L_r(h, \mathbf{z}')| \leq \frac{1}{n} \left((2\Delta_G + 4) R \|\mathbf{w}\|_q + 1 \right) D_{\mathrm{H}}(\mathbf{z}, \mathbf{z}'). \tag{6.32}$$

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

$$|L_r(h, \mathbf{z}) - L_r(h, \mathbf{z}')| \leq \frac{1}{n} \left(4R \|\mathbf{w}\|_q + 1 \right) D_{\mathrm{H}}(\mathbf{z}, \mathbf{z}'). \tag{6.33}$$

Lemma 8. *Fix any $p, q \geq 1$ such that $1/p + 1/q = 1$. Fix a graph, $G \triangleq (\mathcal{V}, \mathcal{E})$. Assume that $\sup_{x \in \mathcal{X}} \|x\|_p \leq R$. Then, for any example $\mathbf{z} \in \mathcal{Z}^n$, and any two MRFs, h, h' with weights $\mathbf{w}, \mathbf{w}'$,*

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

Lemma 8 implies that L_{r} has $(2 |G| R/n, \emptyset)$-local hypothesis stability.

6.5.1.2 Generalization Bounds for Max-Margin Learning

I now apply the PAC-Bayes bounds to the class of max-margin Markov networks that perform MAP inference, with the templated, linear potentials defined in Section 2.3.4. I denote this class by $\mathcal{H}_{\text{M3N}}$. As a warm-up, I first assume that both the observations and weights are uniformly bounded by the 2-norm unit ball. By Lemma 7, this means that the ramp loss satisfies uniform stability, meaning one can apply Corollary 2.

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

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

The proof is given in Appendix D.5. Note that, with the bounded degree assumption, $|G| \leq n\Delta_G = \text{O}(n)$.

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

Example 3. Fix any $m \geq 1$, $n \geq 1$, $\boldsymbol{\pi} \in \Pi(n)$ and $\delta \in (0, 1)$. Fix a graph, $G \triangleq (\mathcal{V}, \mathcal{E})$, with maximum degree Δ_G. Assume that $\sup_{x \in \mathcal{X}} \|x\|_2 \leq 1$. Then, with probability at

least $1-\delta$ over realizations of $\hat{\mathbf{Z}} \triangleq (\mathbf{Z}^{(l)})_{l=1}^{m}$, for all $h \in \mathcal{H}_{\text{M3N}}$,

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

where

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

Example 3 is only slightly worse than Example 2, incurring a $\mathrm{O}\left(\ln\ln(mn)\right)$ term for the Gaussian construction. Both bounds guarantee generalization when either m or n is large.

The proof of Example 3 uses a concentration inequality for vectors of Gaussian random variables, the proof of which is given Appendix D.6.

Lemma 9. *Let* $\mathbf{X} \triangleq (X_i)_{i=1}^{d}$ *be independent Gaussian random variables, with mean vector* $\boldsymbol{\mu} \triangleq (\mu_1, \ldots, \mu_d)$ *and variance* σ^2*. Then, for any* $p \geq 1$ *and* $\epsilon > 0$*,*

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

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

Proof (Example 3) Define the prior, $\mathbb{P}$, as an isotropic, standard normal distribution; that is, zero-mean, with unit variance in all directions. More precisely, let

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

denote the density of $\mathbb{P}$. Given a (learned) hypothesis, h, we construct the posterior, $\mathbb{Q}_h$,

as another isotropic Gaussian, centered at $\mathbf{w}$, with variance

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

Its density is

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

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

The proof technique involves four steps. First, we upper-bound the KL divergence between $\mathbb{Q}_h$ and $\mathbb{P}$. Then, we identify a β_h and η such that $\mathbb{Q}_h$ is $(\beta_h/n, \emptyset, \eta)$-locally stable. Combining the first two steps with Corollary 3 yields a PAC-Bayes bound for the randomized predictor. The final step is to derandomize this bound using Proposition 4.

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

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

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

Fix any $h \in \mathcal{H}_{\text{M3N}}$, and define a "bad" set of hypotheses,

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

Using Lemma 9,

$$\begin{aligned}
\mathbb{Q}_h\left(\mathcal{B}_{\mathcal{H}_{\text{M3N}}}(h)\right) &= \Pr_{h' \sim \mathbb{Q}_h}\left\{ \|\mathbf{w}' - \mathbf{w}\|_2 \geq \frac{1}{m\,|G|} \right\} \\
&\leq 2d \exp\left(-\frac{2d(m\,|G|)^2 \ln(2dmn)}{2d(m\,|G|)^2} \right) \\
&= \frac{1}{mn}. \qquad (6.34)
\end{aligned}$$

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

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

When combined with Lemma 7, with $R = 1$, we have that

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

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

The definition of β_h depends on the posterior via $\mathbf{w}$. Therefore, we must use a PAC-Bayes bound from Section 6.3. In this case, there are no "bad" inputs, since the

observations are bounded in the unit ball, so we can invoke Corollary 3. Recalling that the ramp loss is 1-uniformly difference bounded, we then have that, with probability at least $1-\delta$, every $\mathbb{Q}_h : h \in \mathcal{H}_{\text{M3N}}$ satisfies

$$\overline{L}_{\text{r}}(\mathbb{Q}_h) \leq \hat{L}_{\text{r}}(\mathbb{Q}_h, \hat{\mathbf{Z}}) + \frac{1}{mn} + 4\beta_h \left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}\right\|_\infty \sqrt{\frac{\frac{1}{2}\left\|\mathbf{w}\right\|_2^2 + \frac{d}{2}\ln\left(2d(m\left|G\right|)^2\ln(2dmn)\right) + \ln\frac{4\beta_h}{\delta}}{2mn}}. \tag{6.35}$$

Observe that $\mathcal{H}_{\text{M3N}}$ is a normed vector space, since it consists of weight vectors in $\mathbb{R}^d$. In this case, we will use the 2-norm. By Equation 6.34, it is clear that $\mathbb{Q}$ has $(1/(m\left|G\right|), 1/(mn))$-local hypothesis stability (Definition 17), since every $h \in \mathcal{H}_{\text{M3N}}$ results in the same probability bound. Further, by Lemma 8, with $R = 1$,

$$\left|L_{\text{r}}(h, \mathbf{z}) - L_{\text{r}}(h', \mathbf{z})\right| \leq \frac{2\left|G\right|}{n}\left\|\mathbf{w} - \mathbf{w}'\right\|_2, \tag{6.36}$$

meaning L_{r} has $(2\left|G\right|/n, \emptyset)$-local hypothesis stability (Definition 16). Therefore, by Proposition 5, $(L_{\text{r}}, \mathbb{Q})$ has $(2/(mn), \emptyset, 1/(mn))$-local stability. It then follows, via Proposition 4 and Equation 6.31, that

$$\overline{L}_{\text{H}}(h) \leq \overline{L}_{\text{r}}(h) \leq \overline{L}_{\text{r}}(\mathbb{Q}_h) + \frac{3}{mn}, \tag{6.37}$$

and

$$\hat{L}_{\text{r}}(\mathbb{Q}_h, \hat{\mathbf{Z}}) \leq \hat{L}_{\text{r}}(h, \hat{\mathbf{Z}}) + \frac{3}{mn} \leq \hat{L}_{\text{h}}(h, \hat{\mathbf{Z}}) + \frac{3}{mn}. \tag{6.38}$$

Combining Equations 6.35, 6.37 and 6.38 completes the proof. ■

6.5.2 Soft-Max Learning

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

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

$$L_{\text{sm}}(h, \mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n} \left(\Phi_\epsilon(\mathbf{x}, \mathbf{y}; \mathbf{w}) - h(\mathbf{x}, \mathbf{y}) \right), \tag{6.39}$$

where $h(\mathbf{x}, \mathbf{y})$ is the unnormalized log-likelihood (Equation 6.28) and

$$\begin{aligned} \Phi_\epsilon(\mathbf{x}, \mathbf{y}; \mathbf{w}) &\triangleq \epsilon \ln \sum_{\mathbf{y}' \in \mathcal{Y}^n} \exp\left(\frac{1}{\epsilon} \left(D_{\text{H}}(\mathbf{y}, \mathbf{y}') + h(\mathbf{x}, \mathbf{y}') \right) \right) \\ &= \epsilon \ln \sum_{\mathbf{y}' \in \mathcal{Y}^n} \exp\left(\frac{1}{\epsilon} \tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}) \cdot \hat{\mathbf{y}}' \right). \end{aligned} \tag{6.40}$$

is the soft-max function. I purposefully overload the notation of the log-partition function due to its relationship to the soft-max. Observe that, for $\epsilon = 1$, the soft-max becomes the log-partition of the distribution induced by the loss-augmented potentials, and Equation 6.39 is the corresponding negative log-likelihood, scaled by $1/n$. Further, as $\epsilon \to 0$, the soft-max approaches the $\max$ operator and Equation 6.39 becomes the hinge loss (Equation 6.27).

The latter equivalence can be illustrated using the variational form of the log-

partition function (Equation 2.1). The soft-max, like the log-partition, has the following variational form:

$$\begin{aligned}\Phi_\epsilon(\mathbf{x}, \mathbf{y}; \mathbf{w}) &= \max_{\boldsymbol{\mu} \in \mathcal{M}} \tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}) \cdot \boldsymbol{\mu} - \epsilon \, \Phi^*(\boldsymbol{\mu}) \\ &= \max_{\boldsymbol{\mu} \in \mathcal{M}} (\boldsymbol{\theta}(\mathbf{x}, \mathbf{y}; \mathbf{w}) + \delta(\mathbf{y})) \cdot \boldsymbol{\mu} - \epsilon \, \Phi^*(\boldsymbol{\mu}), \end{aligned} \tag{6.41}$$

where Φ^* is the convex conjugate of the loss-augmented log-partition. This maximization is equivalent to marginal inference with loss-augmented potentials. Thus, the soft-max is generally intractable to compute. In practice, one could substitute a variational technique, such as the ones discussed in Chapter 7.

Let $\boldsymbol{\mu}_{\text{u}}$ denote the marginals of the unary cliques, and observe that

$$\delta(\mathbf{y}) \cdot \boldsymbol{\mu} = \frac{1}{2} \left\| \mathbf{y} - \boldsymbol{\mu}_{\text{u}} \right\|_1 \triangleq D_1(\mathbf{y}, \boldsymbol{\mu}). \tag{6.42}$$

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

$$h_\epsilon(\mathbf{x}, \boldsymbol{\mu}) \triangleq \boldsymbol{\theta}(\mathbf{x}; \mathbf{w}) \cdot \boldsymbol{\mu} - \epsilon \, \Phi^*(\boldsymbol{\mu}). \tag{6.43}$$

Recall that each full labeling, $\hat{\mathbf{y}}$, corresponds to a vertex of the marginal polytope, so $\hat{\mathbf{y}} \in \mathcal{M}$. Further, $h_\epsilon(\mathbf{x}, \hat{\mathbf{y}}) = h(\mathbf{x}, \mathbf{y})$, since $\Phi^*(\hat{\mathbf{y}}) = 0$. Thus, combining Equations 6.41 to 6.43, we have that the soft-max loss (Equation 6.39) is equivalent to

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

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

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

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

Essentially, this just replaces the maxes over $\mathcal{Z}^n$ with maxes over $\mathcal{M}$ and uses Equation 6.43 instead of Equation 6.28. I refer to this loss as the *soft ramp loss*. Many properties of the regular ramp loss also apply to the soft ramp loss. Since each clique's marginals sum to one, it is straightforward to show that Lemmas 13 and 14 still hold. Further, the additional Φ^* term cancels out in Equations D.5, D.6, D.8 and D.9, so Lemmas 7 and 8 also hold.

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

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

This identity motivates another loss function; with

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

let

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

Note that

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

Conveniently, because the (pseudo)marginals sum to one, it can also be shown that the Hamming loss of the posterior decoding of $\boldsymbol{\mu}_\epsilon(\mathbf{x}; \mathbf{w})$ is at most twice L_1.

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

Example 4. Fix any $m \geq 1$, $n \geq 1$, $\boldsymbol{\pi} \in \Pi(n)$, $\delta \in (0, 1)$ and $G \triangleq (\mathcal{V}, \mathcal{E})$. Assume that $\sup_{x \in \mathcal{X}} \|x\|_2 \leq 1$. Then, with probability at least $1 - \delta$ over realizations of $\hat{\mathbf{Z}} \triangleq (\mathbf{Z}^{(l)})_{l=1}^m$, for all $h \in \mathcal{H}_{\text{CRF}}$,

$$\overline{L}_1(h) \leq \hat{L}_{\text{sm}}(h, \hat{\mathbf{Z}}) + \frac{7}{mn} + 4\beta_h \left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}\right\|_\infty \sqrt{\frac{\frac{1}{2}\|\mathbf{w}\|_2^2 + \frac{d}{2}\ln\left(2d(m\,|G|)^2 \ln(2dmn)\right) + \ln\frac{4\beta_h}{\delta}}{2mn}}, \tag{6.44}$$

where

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

I omit the proof, since it is almost identical to Example 3. The key difference worth noting is that, since the model does not use edge observations, the graph's maximum degree does not appear in β_h.

It is interesting to compare Example 4 to the posterior decoding risk bound in Example 1. We could do so by simply multiplying the righthand side of Equation 6.44 by 2, since the expected posterior decoding Hamming loss, $\overline{L}_0(h)$, is at most twice $\overline{L}_1(h)$.

However, since the weights are uniformly bounded in Example 1, it would be an unfair comparison. To make a fair comparison, we can adapt Example 4 for the class $\{h \in \mathcal{H}_{\text{M3N}} : \|\mathbf{w}\|_2 \leq 1\}$ using the analysis from Example 2. One can then show that

$$\overline{L}_0(h) \leq 2\overline{L}_1(h) \leq 2\hat{L}_{\text{sm}}(h, \hat{\mathbf{Z}}) + \frac{8}{mn} + 20 \left\|\boldsymbol{\Gamma}^{\boldsymbol{\pi}}\right\|_\infty \sqrt{\frac{d \ln(2m\,|G|) + \ln \frac{2}{\delta}}{2mn}}.$$

There may be situations in which this bound is tighter than the one from Example 1, such as when ρ or κ are very small.

6.5.3 Possibly Unbounded Domains

Until now, I have assumed that the observations are uniformly bounded in the unit ball. This assumption is common in the literature, but it does not quite match what happens in practice. Typically, one will rescale each dimension of the input space using the minimum and maximum values found in the training data. While this procedure guarantees a bound on the observations at training time, the bound may not hold at test time when one rescales by the limits estimated from the training set. This outcome would violate the preconditions of Lemmas 13 and 14, thereby invalidating the stability guarantees used to prove the previous examples.

Now, suppose we knew that the observations were bounded with high probability. In the following example, I construct a hypothetical data distribution under which this assumption holds. I combine this with Theorem 3 to derive a variant of Example 3.

Example 5. Fix any $m \geq 1$, $n \geq 1$, $\boldsymbol{\pi} \in \Pi(n)$, $\delta \in (0, 1)$ and $G \triangleq (\mathcal{V}, \mathcal{E})$. Suppose the data generating process, $\mathbb{D}$, is defined as follows. For each $y \in \mathcal{Y}$, assume there

is an associated isotropic Gaussian over $\mathcal{X} \subseteq \mathbb{R}^k$, with mean $\mu_y \in \mathcal{X} : \|\mu_y\|_2 \leq 1$ and variance $\sigma_y^2 \leq (2k \ln(2kn^2))^{-1}$. First, $\mathbf{Y}$ is sampled according to some arbitrary distribution, conditioned on G. Then, for each $i \in [n]$, conditioned on $Y_i = y_i$, a vector of observations, $x_i \in \mathcal{X}$, is sampled according to $(\mu_{y_i}, \sigma_{y_i}^2)$.

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

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

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

where

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

The proof is provided in Appendix D.7.

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

6.6 Discussion

In this chapter, I proposed new PAC-Bayes bounds for structured prediction that use the local definitions of stability from Section 3.1. Like the covering number-based bound in the previous chapter, the PAC-Bayes bounds can decrease with both the number of examples, m, and the size of each example, n. The stability conditions used in this chapter generalize collective stability, thereby accommodating a broader range of structured loss functions, including max-margin and soft-max learning.

The examples in Section 6.5 identify several take-aways for practitioners. Primarily, they indicate the importance of templating (or, parameter-tying). Observe that all of the bounds depend on d, the number of parameters[2], via a term that is $\tilde{\mathrm{O}}(d/n)$. Clearly, if d scales linearly with n, the number of nodes, then this term is bounded away from zero as $n \to \infty$. Consequently, one cannot hope to generalize from one example. Though I do not prove this formally, the intuition is fairly simple: if there is a different $\mathbf{w}_i$ for each node i, and $\mathbf{w}_{ij}$ for each edge $\{i, j\}$, then one example provides exactly one "micro example" from which one can estimate $\{\mathbf{w}_i\}_{i \in \mathcal{V}}$ and $\{\mathbf{w}_{ij}\}_{\{i,j\} \in \mathcal{E}}$. In this setting, my bounds become $\tilde{\mathrm{O}}(1/\sqrt{m})$, which is no better (and no worse) than previous bounds. Thus, templating is crucial to achieving the fast generalization rate.[3]

Another observation is that Examples 3 to 5 depend on the norm of the weight vector, $\mathbf{w}$. Specifically, I used the 2-norm, for its relationship to Gaussian priors; though,

[2] I believe that this dependence is unavoidable when derandomizing PAC-Bayes bounds for structured prediction. Evidence to support this conjecture is given by McAllester's (2007) bound, which depends on the number templates, and the number of parameters is roughly linear in the number of templates.

[3] It may be possible to achieve a fast rate without templating if one imposes a sparsity assumption on the optimal weight vector, but it seems likely that the sparsity would depend on n.

one could substitute any norm, due to the equivalence of norms in finite dimension. Dependence on the norm of the weights is a standard feature of most generalization bounds. This term is commonly interpreted as a measure of hypothesis complexity. Weight regularization during training controls the norm of the weights, thereby effectively limiting the complexity of the learned model.

The structure of the the model influences the bounds via Δ_G, the maximum degree of the graph, and $|G|$, the total number of nodes and edges. (Since the bounds are sub-logarithmic in G, and $\frac{1}{n} \ln |G| \leq \frac{2}{n} \ln n$, one could reasonably argue that Δ_G is the only important structural term.) It is important to note that the edges in the model need not necessarily correspond to concrete relationships in the data. For example, there are many ways to define the "influential" neighbors of a user in a social network, though the user may be connected to nearly everyone in the network; the adjacencies one models may be a subset of the true adjacencies. Therefore, Δ_G and $|G|$ are quantities that one can control; they become part of the trade-off between representational power and overfitting. In light of this trade-off, recall that the stability term, β_h, partially depends on whether one conditions on the observations in the edge features; as shown in Examples 4 and 5, β_h can be reduced to $\mathrm{O}(\|\mathbf{w}\|_2)$ if one does not. On the other hand, if observations are modeled in the edge features, and $\Delta_G = \mathrm{O}(\sqrt{n})$, then the bounds become $\tilde{\mathrm{O}}(1/\sqrt{m})$. Thus, under this modeling assumption, controlling the maximum degree is critical.

There are several ways in which my analysis can be refined and extended. In Lemma 7, which I use to establish the stability of the ramp loss, I used a rather course application of Hölder's inequality to isolate the influence of the weights. This technique ignores the relative magnitudes of the node and edge weights. Indeed, it may be the case

that the edge weights are significantly lower than the node weights. A finer analysis of the weights could improve Equation 6.32 and might yield new insights for weight regularization. One could also abstract the desirable properties of the potential functions to accommodate a broader class than the linear potentials used in our examples. Finally, I conjecture that the bounds could be tightened by adapting Germain et al.'s (2009) analysis to bound $\phi^2(h, \hat{\mathbf{Z}}) \triangleq \left(\overline{L}(h) - \hat{L}(h, \hat{\mathbf{Z}})\right)^2$ instead of $\phi(h, \hat{\mathbf{Z}}) \triangleq \overline{L}(\mathbb{Q}) - \hat{L}(\mathbb{Q}, \hat{\mathbf{Z}})$. The primary challenge would be bounding the moment-generating function, $\mathbb{E}_{\hat{\mathbf{Z}} \sim \mathbb{D}^m}\left[e^{u\,\phi^2(h, \hat{\mathbf{Z}})}\right]$, since my martingale-based method would not work. If successful, this analysis could yield bounds that tighten when the empirical loss is small.

Chapter 7: Learning Marginals with Strongly Convex Variational Inference

In this chapter, I show that learning with a strongly convex free energy results in more accurate marginal probabilities. I begin with a theoretical motivation for using strongly convex free energies (Section 7.1), concluding that one should prefer those whose modulus of convexity is constant with respect to the size of the graph. I then demonstrate when this condition holds for two popular variational techniques (Section 7.2). These insights suggest a framework for optimizing the strength of convexity in a family of variational methods (Section 7.2.2). I conclude with an experimental evaluation (Section 7.3), which verifies that strongly convex free energies can indeed result in more accurate marginals.

The results of this chapter apply to the class of pairwise MRFs introduced in Section 2.3. I do not any specific form of the node and edge potentials, nor do I assume that the model is templated. For simplicity of exposition, I do not consider conditioning on evidence; therefore, the notation omits $\mathbf{X}$.

7.1 A Case for Strong Convexity

Recall the form of the (variational) free energy, $E(\boldsymbol{\mu}; \boldsymbol{\theta}) = -\boldsymbol{\theta} \cdot \boldsymbol{\mu} + \Phi^*(\boldsymbol{\mu})$, where Φ^* (or $\tilde{\Phi}^*$ for approximations) is the convex conjugate of the log-partition function, Φ (or

$\tilde{\Phi}$). Because the dot product is linear, the convexity of the free energy is determined by the convexity of Φ^* or $\tilde{\Phi}^*$. Some approximate conjugates are known to be convex, yet few studies discuss the *strength* of convexity, which is defined in Section 2.4. (For this chapter, I use the definition associated with differentiable functions, Definition 2.)

The true conjugate function, Φ^*, is a strongly convex function of the full probability table. Since the marginals are a linear function of the probability table, Φ^* is also a strongly convex function of $\mathcal{M}$—albeit with an unknown modulus. Approximations of Φ^* that are simply convex ignore this fact, and may result in less accurate marginals.

The purpose of this section is to motivate the use of strongly convex free energies. I start by connecting strong convexity to stability, showing that strong convexity is both sufficient (Section 7.1.1) and necessary (Section 7.1.2) for uniform stability, which can be used to derive bounds on the quality of learned marginals (Section 7.1.3). More importantly, the theory suggests that the modulus of convexity is crucial, and that one should prefer moduli that are independent of the size of the graph (Section 7.1.4). Proofs from this section are deferred to Section E.2.

7.1.1 Strong Convexity Guarantees Stability

There is a well-known duality between strong convexity and the *Lipschitz continuity* of the gradient.

Definition 18. A differentiable function, $\varphi : \mathcal{S} \to \mathbb{R}$, has a λ-*Lipschitz continuous gradi-*

ent if and only if, for all $s, s' \in \mathcal{S}$,

$$\|\nabla\varphi(s) - \nabla\varphi(s')\|_2 \leq \lambda \|s - s'\|_2 . \tag{7.1}$$

Lemma 10 (Hiriart-Urruty and Lemaréchal, 2001, Theorem 4.2.1)**.** *Let $\varphi : \mathcal{S} \to \mathbb{R}$ denote a differentiable function, and $\varphi^\star : \mathcal{S}^\star \to \mathbb{R}$ its convex conjugate. If $\varphi^\star$ is κ-strongly convex, then φ has a $(1/\kappa)$-Lipschitz continuous gradient.*

Since the gradient of $\tilde{\Phi}$ corresponds to the pseudomarginals of the distribution, a strongly convex conjugate function lets us bound the stability of approximate marginal inference. This is summarized in the following lemma.[1]

Lemma 11. *Assume that $\tilde{E}$ uses a κ-strongly convex conjugate function, $\tilde{\Phi}^*$. Then, for any $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$,*

$$\frac{1}{\sqrt{|G|}} \|\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}) - \tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}')\|_2 \leq \frac{1}{\kappa\sqrt{|G|}} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 . \tag{7.2}$$

Lemma 11 upper-bounds the root-mean-squared difference between the respective pseudomarginals of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. Observe that one can trivially upper-bound this quantity by $\sqrt{2}$ by assuming that the marginals are completely different. In contrast, the right-hand side of Equation 7.2 shrinks as a function of the size of the graph, $|G|$, and the L^2 distance between the potentials, $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$, provided κ is lower-bounded by a function that is independent of these terms. Of course, since the potentials have length $\mathrm{O}(|G|)$, their L^2 distance could be $\mathrm{O}(\sqrt{|G|})$; but there are some cases in which the distance could be small. In Section 7.1.3, I discuss one such scenario and use it to derive a bound on the

[1] Wainwright derived a similar result (2006, Lemma 6). My lemma is more explicit about the role of the modulus of convexity.

root-mean-squared error (RMSE) of learned pseudomarginals.

7.1.2 Convexity Alone Does Not Guarantee Stability

Strong convexity is central to Lemma 11. In fact, there is good reason to believe that strong convexity is a *necessary* condition for *uniform* stability. To understand why, I return to the relationship between strong convexity and Lipschitz gradients. Lemma 10 states that the former property implies the latter; however, the converse is also true.

Lemma 12 (Hiriart-Urruty and Lemaréchal, 2001, Theorem 4.2.2)**.** *Let* $\varphi : \mathcal{S} \to \mathbb{R}$ *denote a differentiable function, and* $\varphi^\star : \mathcal{S}^\star \to \mathbb{R}$ *its convex conjugate. If* φ *has a* λ*-Lipschitz continuous gradient, then* $\varphi^\star$ *is* $(1/\lambda)$*-strongly convex.*

This establishes an equivalence between strong convexity and Lipshitz gradients: φ has a $(1/\kappa)$-Lipschitz continuous gradient *if and only if* $\varphi^\star$ is κ-strongly convex. In the context of variational inference, this means that Equation 7.2 holds if and only if $\tilde{\Phi}^*$ is strongly convex. Mere convexity (i.e., $\kappa = 0$) is insufficient for guaranteeing stability. In fact, for any simply convex $\tilde{\Phi}^*$, it may be possible to construct an example in which marginal inference is not stable.

For instance, consider the extreme case in which $\tilde{\Phi}^*$ is linear in $\tilde{\mathcal{M}}$. This means that $\tilde{E}$ is also linear. Mangasarian and Shiau (1987) prove by counterexample that solutions to linear programs are not *Lipschitz continuous* (a form of stability) with respect to perturbations in the objective coefficients (in this case, the potentials). Therefore, inference with a linear conjugate function cannot have non-trivial uniform stability. Note that this insight implies that MAP inference cannot be uniformly stable.

7.1.3 Stability Yields Learning Guarantees

Equation 7.2 is especially meaningful in the context of learning. Suppose we are trying to learn a distribution, $p(\mathbf{Y}; \boldsymbol{\theta}^\star)$, parameterized by some potentials, $\boldsymbol{\theta}^\star$. We assume that the class of models to which $\boldsymbol{\theta}^\star$ belongs is known, and that the variable interactions, defined by a graph G, are fixed. Our goal is to estimate $\boldsymbol{\theta}^\star$ given m independent draws from the distribution, $(\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(m)})$. To do so, we minimize the negative log-likelihood (NLL) of the variational distribution, $\tilde{p}$, induced by an approximate log-partition, $\tilde{\Phi}$. The approximation is for efficiency, since we make repeated evaluations of the objective during learning. Assume that $\tilde{\Phi}^*$, the convex conjugate of $\tilde{\Phi}$, is κ-strongly convex. Let $\mathcal{L}(\mathbf{Y}; \boldsymbol{\theta}) \triangleq -\ln \tilde{p}(\mathbf{Y}; \boldsymbol{\theta})$ denote the NLL under $\tilde{p}$, and let

$$\mathcal{L}_m(\boldsymbol{\theta}) \triangleq \frac{1}{m} \sum_{j=1}^{m} \mathcal{L}(\mathbf{y}^{(j)}; \boldsymbol{\theta}). \tag{7.3}$$

Let

$$\bar{\boldsymbol{\theta}} \triangleq \arg\min_{\boldsymbol{\theta}} \mathbb{E}\left[\mathcal{L}(\mathbf{Y}; \boldsymbol{\theta})\right], \tag{7.4}$$

$$\text{and} \qquad \hat{\boldsymbol{\theta}}_m \triangleq \arg\min_{\boldsymbol{\theta}} \mathcal{L}_m(\boldsymbol{\theta}) + \Lambda_m \left\|\boldsymbol{\theta}\right\|_2^2. \tag{7.5}$$

If $\Lambda_m \to 0$ as $m \to \infty$, then $\bar{\boldsymbol{\theta}} = \lim_{m\to\infty} \hat{\boldsymbol{\theta}}_m$.

Because $\mathcal{L}_m$ uses the approximate log-partition, $\hat{\boldsymbol{\theta}}_m$ is not a *consistent* estimator. In other words, in the limit of infinite data, $\hat{\boldsymbol{\theta}}_m$ may be different from $\boldsymbol{\theta}^\star$. Nonetheless, we have that $\boldsymbol{\mu}(\boldsymbol{\theta}^\star) = \tilde{\boldsymbol{\mu}}(\bar{\boldsymbol{\theta}})$, as shown in Section E.2.2. In light of this, substituting $\hat{\boldsymbol{\theta}}_m$ and $\bar{\boldsymbol{\theta}}$

into Equation 7.2, we have that the RMSE of the learned marginals, $\tilde{\boldsymbol{\mu}}(\hat{\boldsymbol{\theta}}_m)$, with respect to the true marginals, $\boldsymbol{\mu}(\boldsymbol{\theta}^\star)$, is proportional to the distance between $\hat{\boldsymbol{\theta}}_m$ and $\bar{\boldsymbol{\theta}}$, divided by the modulus of convexity, κ. As $\hat{\boldsymbol{\theta}}_m$ converges to $\bar{\boldsymbol{\theta}}$, the RMSE decreases at a rate that is inversely proportional to κ.

Convergence of *M-estimators* has been studied extensively. Many of these works (e.g., Bickel et al., 2009; Kakade et al., 2010; Ravikumar et al., 2011; Negahban et al., 2012; Bradley and Guestrin, 2012; Meng et al., 2014) rely on a *restricted eigenvalue (RE)* assumption. Essentially, this assumes that the eigenvalues of $\nabla^2\mathcal{L}(\,\cdot\,;\boldsymbol{\theta})$—which is independent of $\mathbf{Y}$, and therefore the same as $\nabla^2\mathcal{L}_m(\boldsymbol{\theta})$—evaluated in the vicinity of $\bar{\boldsymbol{\theta}}$, are bounded away from zero; meaning, the NLL is strongly convex in a region around $\bar{\boldsymbol{\theta}}$. I will further assume that, with probability $\geq 1-\delta$ over draws of the training set, both $\bar{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_m$ (which is a random variable) are contained in a convex set within which $\nabla^2\mathcal{L}(\,\cdot\,;\boldsymbol{\theta})$ is positive definite, thereby implying that the NLL is strongly convex in this set. The minimum eigenvalue of the Hessian (hence, the modulus of convexity) may depend on δ, m and G, but should be bounded away from zero by a constant as $m \to \infty$. This requirement will always be met if $\nabla^2\mathcal{L}(\,\cdot\,;\bar{\boldsymbol{\theta}})$ is positive definite.

Assumption 1. Assume that there exists a constant, $\bar{\gamma} > 0$, such that the minimum eigenvalue of $\nabla^2\mathcal{L}(\,\cdot\,;\bar{\boldsymbol{\theta}})$ is at least $\bar{\gamma}$. Further, for any $\delta \in (0,1)$ and $m \geq 1$, there exists a convex set, $\mathcal{S} \subseteq \mathbb{R}^{|\boldsymbol{\theta}|}$, encompassing both $\bar{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_m$, and a function, $\gamma(\delta, m, G) = \Omega(1)$, such that, with probability $\geq 1-\delta$ over draws of m i.i.d. examples, the minimum eigenvalue of $\nabla^2\mathcal{L}(\,\cdot\,;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{S}$ is at least $\gamma(\delta, m, G)$.

Combining Assumption 1 and Lemma 11, one can prove a high-probability error

bound on the marginals of a model learned with strongly convex variational inference.

Proposition 6. *Let* $\Lambda_m \triangleq 1/\sqrt{m}$. *Assume that* $\tilde{\Phi}^*$ *is* κ*-strongly convex, that Assumption 1 holds, and that* $\left\|\bar{\boldsymbol{\theta}}\right\|_\infty \leq 1$. *Then, for any* $\delta \in (0,1)$, *with probability at least* $1-2\delta$ *over draws of* m *i.i.d. examples,*

$$\frac{1}{\sqrt{|G|}} \left\| \tilde{\boldsymbol{\mu}}(\hat{\boldsymbol{\theta}}_m) - \boldsymbol{\mu}(\boldsymbol{\theta}^\star) \right\|_2 \leq \frac{|\mathcal{Y}|}{\kappa\, \gamma(\delta, m, G)\sqrt{m}} \left(2 + \sqrt{\frac{1}{2} \ln \frac{2\,|\mathcal{Y}|^2\,|G|}{\delta}} \right). \tag{7.6}$$

Like most error bounds, Equation 7.6 has an inverse dependence on the square root of m, so the bound decreases as the training set grows. What is interesting about this bound is that it incorporates the modulus of convexity, κ, of the variational free energy. Because of the inverse dependence on κ, the bound tightens as κ grows. Note that the upper bound for $\left\|\bar{\boldsymbol{\theta}}\right\|_\infty$ can be replaced with any constant. Also note that Proposition 6 is easily adapted for the mean-absolute error (MAE), since the RMSE upper-bounds the MAE.

7.1.4 Prefer a Constant Modulus

Equations 7.2 and 7.6 have an inverse dependence on the modulus of convexity. One should therefore prefer higher values, leading to sharper bounds. However, stronger convexity might mean that the approximation is looser. For instance, one can trivially boost the modulus by scaling the conjugate function with a temperature parameter. This reduces the bounds, but creates a totally entropic distribution. One therefore wonders whether there is a "right" amount of convexity that trades off stability for marginal accuracy.

One criterion stands out: *the modulus should not have an inverse dependence on*

$|G|$. This insight is the most important takeaway of this section. When learning large graphical models, it is usually the case that the number of examples is small relative to the size of the graph. In this setting, κ can have great impact. If $\kappa = \Omega(1/|G|)$, then the learning rate (Equation 7.6) is $\tilde{\mathrm{O}}\left(|G|/\sqrt{m}\right)$, which is vacuous for $|G| > \sqrt{m}$. In contrast, if $\kappa = \Omega(1)$, then the learning rate is $\tilde{\mathrm{O}}\left(1/\sqrt{m}\right)$. This observation motivates the study of $\Omega(1)$-strongly convex free energies in the next section.

7.2 Strongly Convex Free Energies

In light of Section 7.1.4, it is important to identify strongly convex free energies for which the modulus of convexity is lower-bounded by a function that does not decrease with $|G|$. In this section, I present new guarantees for two popular variational methods. First, I provide model-dependent conditions under which the tree-reweighted negative entropy is $\Omega(1)$-strongly convex (Section 7.2.1). To prove this result, I prove a similar claim for the negative entropy of a tree-structured model (given in Section E.3.1). I also analyze the class of counting number entropies (which subsumes tree-reweighting), proving an interesting relationship between the counting numbers and the modulus of convexity (Section 7.2.2). Using this insight, I then provide a counting number optimization that guarantees κ-strong convexity, for any $\kappa > 0$, independent of the model.

7.2.1 Tree-Reweighting

The tree-reweighted entropy (Wainwright et al., 2005) is a convex combination of tree entropies. In this section, I give conditions under which its modulus of convexity is

lower-bounded by a function of the parameters and structural properties, independent of graph size.

Fix a graph, G, and let $\mathcal{T}(G)$ denote its spanning trees. For a tree $T \triangleq (\mathcal{V}, \mathcal{E}_T) \in \mathcal{T}(G)$, its entropy is given by

$$H_T(\tilde{\boldsymbol{\mu}}) \triangleq \sum_{v \in \mathcal{V}} (1 - \deg(v)) H_v(\tilde{\mu}_v) + \sum_{e \in \mathcal{E}_T} H_e(\tilde{\mu}_e), \tag{7.7}$$

where $\deg(v)$ is the degree of node v, and

$$H_v(\tilde{\mu}_v) \triangleq -\sum_{j=1}^{|\mathcal{Y}|} \tilde{\mu}_v^j \log \tilde{\mu}_v^j$$
$$\text{and} \quad H_e(\tilde{\mu}_e) \triangleq -\sum_{i=1}^{|\mathcal{Y}|} \sum_{j=1}^{|\mathcal{Y}|} \tilde{\mu}_e^{ij} \log \tilde{\mu}_e^{ij}$$

are the node and edge local entropies. (Equation 7.7 is also the Bethe entropy.) For a distribution, ρ, over $\mathcal{T}(G)$, the tree-reweighted entropy is given by

$$\begin{aligned} H_{\text{TR}}(\tilde{\boldsymbol{\mu}}) &\triangleq \sum_{T \in \mathcal{T}(G)} \rho(T)\, H_T(\tilde{\boldsymbol{\mu}}) \\ &= \sum_{v \in \mathcal{V}} \Big(1 - \sum_{e : v \in e} \rho(e)\Big) H_v(\mu_v) + \sum_{e \in \mathcal{E}} \rho(e) H_e(\mu_e). \end{aligned} \tag{7.8}$$

Wainwright (2006) showed that if each edge, $e \in \mathcal{E}$, has positive marginal probability, $\rho(e) > 0$ (i.e., e appears in at least one tree, T, with $\rho(T) > 0$), then $-H_{\text{TR}}$ is at least $\Omega(1/|G|)$-strongly convex. Unfortunately, this modulus decreases as a function of the size of the graph. This is partly because Wainwright's analysis considers all models in the exponential family. Here, I prove a more optimistic lower bound for models that

exhibit good *contraction*.

Definition 19. Fix a graph, $G \triangleq (\mathcal{V}, \mathcal{E})$, and potentials, $\boldsymbol{\theta}$, which induce a probability density, p. For any $(u, v) : \{u, v\} \in \mathcal{E}$, define the *contraction coefficient* as

$$\vartheta_{\boldsymbol{\theta}}(u, v) \triangleq \sup_{y, y' \in \mathcal{Y}} \left\| p\left(Y_u \,|\, Y_v = y; \boldsymbol{\theta}\right) - p\left(Y_u \,|\, Y_v = y'; \boldsymbol{\theta}\right) \right\|_{\text{TV}}.$$

Denote the maximum of the contraction coefficients by

$$\vartheta^{\star}_{\boldsymbol{\theta}} \triangleq \sup_{(u,v):\{u,v\} \in \mathcal{E}} \vartheta_{\boldsymbol{\theta}}(u, v).$$

The contraction coefficients measure the dependence between adjacent variables in a graphical model. A contraction coefficient of 1 implies determinism, and 0 implies independence. In Section E.3.2, I describe an efficient procedure for computing the contraction coefficients in a tree-structured model.

Roughly speaking, the contraction coefficients are determined by the ratio of "local" signal to "relational" signal. If the local signal is strong, Y_v has little influence on Y_u. For models with a sufficiently high ratio of local-to-relational signal, dependence decays with graph distance at a geometric rate. In this case, one can show that $-H_T$ is $\Omega(1)$-strongly convex (see Section E.3.1). Using this result, one obtains the following.

Proposition 7. *Fix a graph, $G \triangleq (\mathcal{V}, \mathcal{E})$, with maximum degree independent of $|\mathcal{V}|$. Fix a distribution, ρ, over the spanning trees, $\mathcal{T}(G)$, such that there exists a constant, $C > 0 : \forall\, e \in \mathcal{E},\ \rho(e) \geq C$, that lower-bounds the edge probabilities. Let $\Theta \subseteq \mathbb{R}^{|\boldsymbol{\theta}|}$ denote the set of potentials such that each tree $T \in \mathcal{T}(G) : \rho(T) > 0$, with maximum degree Δ_T, has*

maximum contraction coefficient $\vartheta^\star_{\boldsymbol{\theta},T} \leq 1/\Delta_T$. *Let* $\tilde{\mathcal{M}}(\Theta) \triangleq \{\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ *denote the set of pseudomarginals realizable under any* $\boldsymbol{\theta} \in \Theta$. *Then,* $-H_{\text{TR}}$ *is* $\Omega(1)$*-strongly convex in* $\tilde{\mathcal{M}}(\Theta)$.

The proof is given in Section E.4.1. See Section E.4.2 for implications of Proposition 7 for a grid-graph model.

Proposition 7 guarantees $\Omega(1)$-strong convexity, but it still does not identify the modulus. Further, it is model-dependent, and may not hold for certain potentials. Therefore, applying Proposition 6 to tree-reweighted variational inference is only meaningful when learning in a constrained model space that admits good contraction. In the next section, I describe a technique to tune the modulus to any specified value, regardless of the model.

7.2.2 Counting Number Optimization

Counting number techniques decompose the entropy into a weighted sum of node and edge local entropies. For $\mathbf{c} \triangleq ((c_v)_{v\in\mathcal{V}}, (c_e)_{e\in\mathcal{E}})$, the counting number entropy is

$$H_{\text{C}}(\tilde{\boldsymbol{\mu}}) \triangleq \sum_{v\in\mathcal{V}} c_v H_v(\tilde{\mu}_v) + \sum_{e\in\mathcal{E}} c_e H_e(\tilde{\mu}_e). \tag{7.9}$$

Note that H_{C} generalizes the Bethe entropy (Equation 7.7), which is given by $c_v = 1 - \deg(v)$ and $c_e = 1$. One can also recreate the tree-reweighted entropy (Equation 7.8) with $c_v = 1 - \sum_{e:v\in e} \rho(e)$ and $c_e = \rho(e)$. In this section, I show how to find counting numbers that preserve strong convexity, with a modulus that is lower-bounded by a given value.

Since $-H_v$ and $-H_e$ are convex, it is clear from Equation 7.9 that $-H_{\text{C}}$ is convex

for nonnegative counting numbers. Heskes (2006) derived more sophisticated sufficient conditions for convexity by reparameterizing the counting numbers. Specifically, $-H_{\mathrm{c}}$ is convex if there exist nonnegative *auxiliary* counting numbers, $(\alpha_v \geq 0)_{v\in\mathcal{V}}$, $(\alpha_e \geq 0)_{e\in\mathcal{E}}$ and $(\alpha_{v,e} \geq 0)_{e\in\mathcal{E},v\in e}$, such that

$$\forall v \in \mathcal{V},\ \ c_v = \alpha_v - \sum_{e:v\in e} \alpha_{v,e}, \tag{7.10}$$

$$\text{and} \quad \forall e \in \mathcal{E},\ \ c_e = \alpha_e + \sum_{v:v\in e} \alpha_{v,e}. \tag{7.11}$$

The effect of the auxiliary counting numbers, in particular, $\alpha_{v,e}$, is to shift weight between the regular counting numbers, c_v and c_e. Heskes' conditions mean that c_v can be negative and still guarantee convexity. One can further show that $-H_{\mathrm{c}}$ is *strongly* convex whenever α_e is uniformly lower-bounded; α_v and $\alpha_{v,e}$, however, are only required to be nonnegative.

Proposition 8. *Fix a graph, $G \triangleq (\mathcal{V}, \mathcal{E})$, and assume that every node is in at least one edge. If* $\mathbf{c}$ *satisfies Equations 7.10 and 7.11 for some $\kappa > 0$, $(\alpha_v \geq 0)_{v\in\mathcal{V}}$, $(\alpha_e \geq \kappa)_{e\in\mathcal{E}}$ and $(\alpha_{v,e} \geq 0)_{e\in\mathcal{E},v\in e}$, then $-H_{\mathrm{c}}$, is $(\kappa/3)$-strongly convex.*

The proof is given in Section E.5.

Proposition 8 can be used to characterize the strong convexity of a range of algorithms that optimize counting numbers. For example, observing that the Bethe approximation often outperformed tree-reweighting in practice, Meshi et al. (2009) proposed a "convexified" Bethe approximation. Their algorithm finds a set of counting numbers that best approximates the Bethe counting numbers, $\mathbf{c}^{\mathrm{B}}$, while satisfying Heskes' convexity

conditions (Equations 7.10 and 7.11). They also proposed incorporating a constraint that, for all $v \in \mathcal{V}$, $c_v + \sum_{e:v\in e} c_e = 1$; this ensures that the counting numbers are *variable-valid* for a fully factored (i.e., edgeless) model. Via Proposition 8, adding a constraint that $\alpha_e \geq 3\kappa$ ensures that the resulting negative entropy is κ-strongly convex. This yields the following constrained quadratic program (QP), which I refer to as the *strongly convexified Bethe* approximation:

$$\begin{aligned} \min_{\mathbf{c},\boldsymbol{\alpha}\geq\mathbf{0}} \quad & \|\mathbf{c} - \mathbf{c}^{\text{B}}\|_2^2 \\ \text{s.t.} \quad & \forall v \in \mathcal{V},\ c_v + \sum_{e:v\in e} \alpha_{v,e} \geq 0\,; \\ & \forall e \in \mathcal{E},\ c_e - \sum_{v:v\in e} \alpha_{v,e} \geq 3\kappa\,; \\ & \forall v \in \mathcal{V},\ c_v + \sum_{e:v\in e} c_e = 1. \end{aligned} \tag{7.12}$$

Note that Equation 7.12 only depends on the graph structure; it is independent of the potentials. Thus, the QP only needs to be solved once, prior to learning, for each example in the training set. Moreover, examples that have the same structure can use the same counting numbers. Certain graphs, such as regular graphs, may admit an analytic solution to Equation 7.12, thereby avoiding numerical optimization.

One can strongly convexify any desired counting numbers. For instance, Hazan and Shashua (2008) proposed a convex counting number optimization that encourages $c_e = 1$ uniformly. With a small modification to Equation 7.12, one can make Hazan and Shashua's method strongly convex. One can also optimize the tree-reweighted entropy. Though $-H_{\text{TR}}$ is already $\Omega(1)$-strongly convex for certain models (per Proposition 7), it

may be difficult to identify the modulus. By substituting the tree-reweighted counting numbers for $\mathbf{c}^{\text{B}}$ in the objective, one can ensure that $-H_{\text{TR}}$ is at least κ-strongly convex, for any given κ, independent of the model.

For certain graphs and values of κ, the variable validity constraint may make the optimization infeasible. In these cases, I propose switching to a slackened QP:

$$\begin{aligned} \min_{\mathbf{c},\boldsymbol{\alpha}\geq\mathbf{0},\boldsymbol{\xi}} \quad & \|\mathbf{c}-\mathbf{c}^{\text{B}}\|_2^2 + C\,\|\boldsymbol{\xi}\|_2^2 \\ \text{s.t.} \quad & \forall v \in \mathcal{V},\ c_v + \sum_{e:v\in e} \alpha_{v,e} \geq 0\,; \\ & \forall e \in \mathcal{E},\ c_e - \sum_{v:v\in e} \alpha_{v,e} \geq 3\kappa\,; \\ & \forall v \in \mathcal{V},\ c_v + \sum_{e:v\in e} c_e = 1 + \xi_v\,. \end{aligned} \tag{7.13}$$

This introduces a free parameter, $C \geq 0$, that adjusts the trade-off between fitting the target counts (in the equation below, the Bethe counts) and variable validity. I explore this trade-off in Section 7.3.3.

7.3 Experiments

The following empirical evaluation tests the hypothesis that strongly convex free energies result in better learned marginals, as suggested by Proposition 6. Evaluations of approximate inference techniques typically use the *true* model to measure the discrepancy in the marginals. That is, given the model that generated the data, $\boldsymbol{\theta}^\star$, most studies measure $\|\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}^\star) - \boldsymbol{\mu}(\boldsymbol{\theta}^\star)\|$, where $\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}^\star)$ uses the true model with approximate inference. While this isolates the quality of the approximation, it ignores the fact that approximate infer-

ence is typically used both at train and test time. It is therefore valuable to test the quality of the approximation using a model that is *learned* with said approximation. Wainwright (2006) called this "learning the 'wrong' graphical model," since the learned model may not converge to the true model. I prefer to call it "learning the 'right' graphical model for the 'wrong' inference," since it finds the best parameters for the given variational method. I therefore report scores for both true and learned models.

7.3.1 Data Generator

The synthetic data generator is based on those used in prior work (e.g., Hazan and Shashua, 2008; Meshi et al., 2009) to evaluate approximate marginal inference. Data is generated from an (8×8) non-toroidal grid-structured model, in which each node, v, is associated with a binary variable, $Y_v \in \{\mathbf{e}_1, \mathbf{e}_2\}$. The model is defined by the following process, for either "attractive" or "mixed" potentials. First, fix $\omega_s > 0$ and $\omega_p > 0$. For each node, flip a fair coin, $c_v \in \{\pm 1\}$, and let $w_v \triangleq \omega_s \, c_v \left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right]$. If the model is "attractive," uniformly set $w_e \triangleq \omega_p \operatorname{vec}\left(\left[\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right]\right)$, where $\operatorname{vec}(\,\cdot\,)$ converts a matrix to a vector; if "mixed," flip another fair coin, $c_e \in \{\pm 1\}$, and set $w_e \triangleq \omega_p \, c_e \operatorname{vec}\left(\left[\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right]\right)$. To create local perturbations (i.e., evidence), draw a uniformly random $x_v \sim \mathbb{U}[0, 1]$ for each node, and let

$$\forall v,\ \theta_v \triangleq w_v \, x_v, \quad \text{and} \quad \forall e = \{u, v\},\ \theta_e \triangleq w_e \left(\frac{x_u + x_v}{2}\right),$$

Then, the data distribution is defined as

$$p\left(\mathbf{Y} = \mathbf{y}; \boldsymbol{\theta}\right) \triangleq \sum_{v \in \mathcal{V}} \theta_v \cdot y_v + \sum_{e \in \mathcal{E}} \theta_e \cdot (y_u \otimes y_v).$$

This is equivalent to an Ising model with field potentials $\theta_v \sim \mathbb{U}[-\omega_s, \omega_s]$, and interaction potentials $\theta_e \sim \mathbb{U}[0, \omega_p]$, for attractive, or $\theta_e \sim \mathbb{U}[-\omega_p, \omega_p]$, for mixed.

7.3.2 Experiment Design

I use four variational methods from the literature:

LBP: The Bethe approximation (i.e., "loopy" BP).

C-Bethe: Meshi et al.'s (2009) convexified Bethe, which is equivalent to Equation 7.12 with $\kappa = 0$.

TRBP: Wainwright et al.'s (2005) tree-reweighted BP, with the tree distribution described in Section E.4.2.

C-Unif: Hazan and Shashua's (2008) convex counting number optimization, which prefers $c_e = 1$ uniformly.

Of the four, only the last three are guaranteed to be convex; LBP is not convex on a grid. TRBP is in fact strongly convex, though the true modulus depends on the model, and may be difficult to identify. Hazan and Shashua's method actually enforces *strict* convexity, but since the modulus can be arbitrarily close to zero, I consider it effectively just convex. I also compare strongly convexified versions of C-Bethe, TRBP and C-Unif, using the counting number optimization. This results in counting numbers that are provably κ-strongly convex, for a given $\kappa > 0$. I denote these versions by **SC-Bethe**, **SC-TRBP** and **SC-Unif**, respectively, and indicate the value of κ whenever relevant.

For each value of $\omega_s \in \{0.05, 1\}$ and $\omega_p \in \{0.1, 0.2, 0.5, 1, 2, 5\}$, I generate 20 models using the above synthetic generator. Each model acts as a learning trial. For each model, I compute the true marginal probabilities using exact (junction tree) inference and sample 100 joint assignments to $\mathbf{Y}$. I use these samples to train a model for each variational method (and value of κ), using L-BFGS to minimize the regularized

NLL (Equation 7.5). The regularization parameter, Λ_m, is set to $1/\sqrt{m}$, per Proposition 6. I then compute the node marginals using variational inference with the true (i.e., generating) and learned models. For each set of approximate marginals, I compute the root-mean-squared error (RMSE) with respect to the true, exact marginals. I report the average RMSE over 20 trials.

Note that size of each example is $|G| = 176$, which is greater than the number of examples, $m = 100$. We thus have a limited number of large examples. According to Proposition 6, this is a setting in which the modulus of convexity—which determines the stability of inference—is important.

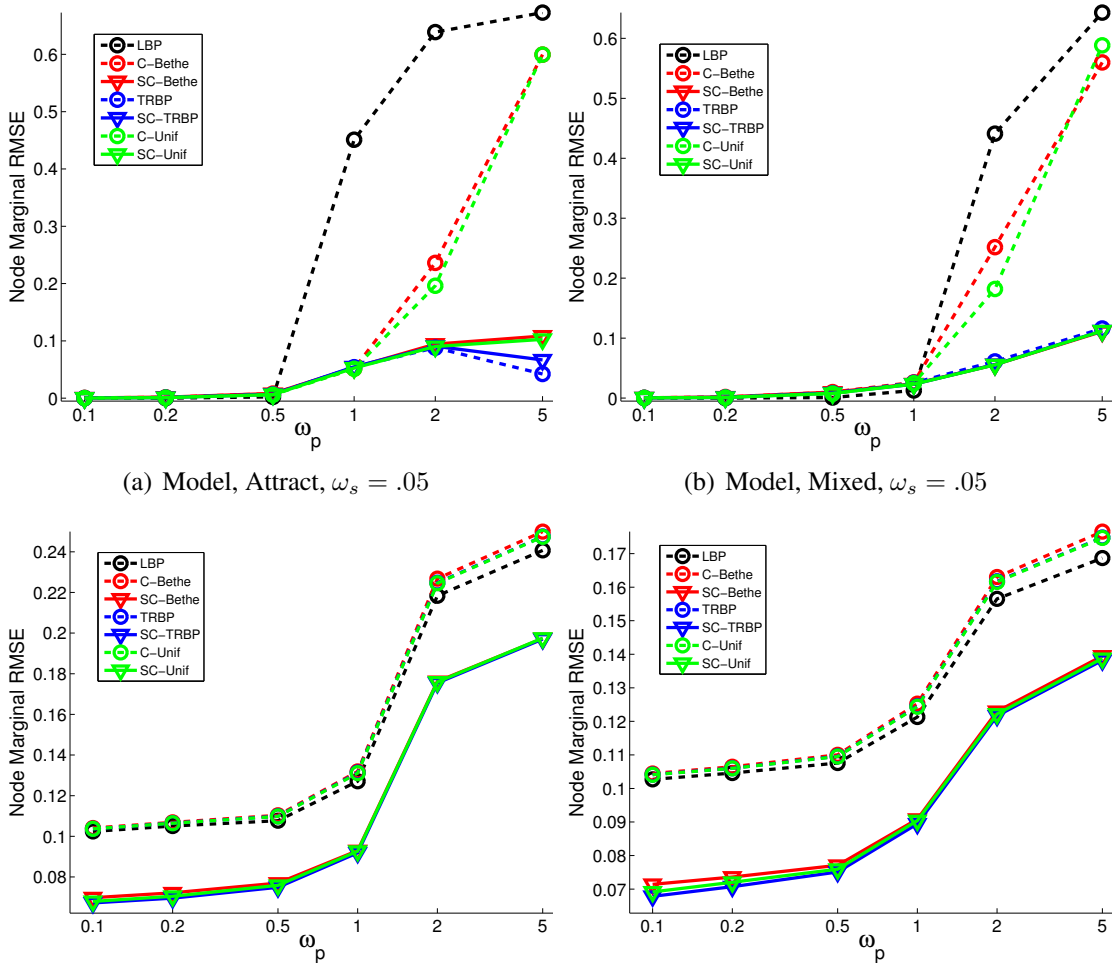
These experiments are implemented in MATLAB, using data structures from Mark Schmidt's *Undirected Graphical Models* (UGM) toolkit (2013b). To optimize the learning objective, I use Schmidt's implementation of L-BFGS with Wolfe line search (2013a). For exact inference and sampling, I use UGM's junction tree implementation. For all variational inference algorithms, I use a custom implementation of counting number belief propagation (CBP), based on Schwing et al.'s (2011) message updates; this can optimize any variational method whose entropy can be expressed with counting numbers. To optimize the counting number QP (Equation 7.12 or 7.13), I use MATLAB's `quadprog`, with the interior point method. To measure statistical significance, I use a paired t-test, with rejection threshold $.05$.

7.3.3 Results

Due to space restrictions, I defer the full catalog of figures to Appendix F. Figure 7.1 highlights select plots.

Strong Convexity Improves Marginal Inference. Figures F.1a-d plot the RMSE of the node marginals as a function of the interaction parameter, ω_p. Inference is performed with the true model. The SC methods use the post hoc optimal value of κ (and C) in the counting number optimization. All methods perform about the same for $\omega_s = 1$ and $\omega_p \leq 2$. LBP has a slight advantage for mixed potentials with $\omega_p \leq 1$, which concurs with previous conclusions (e.g., Meshi et al., 2009) that LBP performs well when there is strong local signal. Focusing on $\omega_s = .05$, the convex methods offer significant improvement over LBP for $\omega_p \geq 1$ with attractive and $\omega_p \geq 2$ with mixed potentials. This shows that convexity helps when there is low local-to-relational signal. In particular, note that the strongly convex methods (TRBP and all SC variants) exhibit dramatically lower error in this setting (see Figures 7.1a-b), with over 10x improvement over LBP.

Strong Convexity Improves Learned Marginals. Figures F.1e-h also plot RMSE as a function of ω_p, but using the learned model to compute the marginals. The SC methods yield statistically significant improvements in almost all data models. Figures 7.1c-d highlight the improvement, which is most prominent when $\omega_s = 1$. In certain cases, SC reduces the error of the convex baselines by over 40%. These results support the hypothesis of Proposition 6, that using a variational free energy that is provably $\Omega(1)$-strongly convex can significantly improve the quality of learned marginals. Moreover, the



(a) Model, Attract, $\omega_s = .05$

(b) Model, Mixed, $\omega_s = .05$

(c) Learned, Attract, $\omega_s = 1$

(d) Learned, Mixed, $\omega_s = 1$

Figure 7.1: Select plots of RMSE (averaged over 20 trials) of the approximate node marginals w.r.t. the true marginals, as a function of the interaction parameter, ω_p. Data is generated with either "attractive" or "mixed" potentials. Figs. (a)-(b) use the true model for inference, and (c)-(d) use the learned model. The black dotted line is LBP; color dotted lines are the convex baselines, and solid lines are their SC counterparts, using the post hoc optimal value of κ (and C for $\kappa \geq .1$). See Section 7.3.3 for discussion and Appendix F for all figures.

SC counting number optimization can even improve TRBP—which is already strongly convex, though the modulus is model-dependent.

Tuning κ in the SC Methods. The value of κ used in the SC counting number optimization can have great impact on the quality of the marginals. The theory in Section 7.1 suggests that increasing the modulus of convexity improves stability and marginal accuracy; however, altering κ affects the quality of the entropy approximation, hence, the marginals. Thus, there is a trade-off that needs to be explored. Figures F.2 and F.3 plot the RMSE of the marginals as a function of κ, using the true and learned models respectively, for select values of ω_s and ω_p. Since values of $\kappa \geq .1$ result in non-variable-valid counting numbers for this grid, I use the slackened QP and report the score for the post hoc optimal C. These plots yield the following insights. When the true potentials are given, and the model has low local-to-relational signal ($\omega_s = .05$, $\omega_p \geq 2$), any modulus of convexity above a certain threshold yields significant improvement. When using variational inference for training, if there is low local signal ($\omega_s = .05$), use the highest value of κ that supports variable validity. Since the local signal is weak, it is even more important to be variable-valid. If local signal is strong ($\omega_p = 1$), one can relax variable validity and push κ further.

Slackened Variable Validity. When using a value of κ that requires slackening variable validity, this requires selecting a value for the slack parameter, C. The quality of the slackened solution can vary with C, since this parameter controls the trade-off between variable validity and fitting the target counts. Figures F.4 and F.5 show select plots of

RMSE as a function of C, focusing on the Bethe and tree-reweighted approximations. Data is generated using mixed potentials. In general, we find that the optimal value of C depends on κ, with lower values of κ favoring lower values of C. This is likely because lower C makes it easier for the QP solver to reduce the slack variables. When training with $\kappa \geq .1$, a good rule of thumb is to set C fairly high; I found that $C = 100$ works well overall.

7.4 Discussion

In this chapter, I have shown, both theoretically and empirically, that variational inference with a strongly convex free energy can improve the accuracy of marginal probabilities. I proved sufficient conditions under which two popular variational methods are strongly convex, and proposed a novel counting number optimization that guarantees κ-strong convexity, for any κ. The experimental results indicate that using this approach to specify a modulus can dramatically reduce the error of approximate marginal inference, suggesting substantial, tangible benefit to applications of graphical models.

The fact that the counting number QP only depends on the graph structure raises several interesting questions for future research. First, are there graphs for which the QP can be simplified, so as to scale to very large graphs? On a related note, are there graphs that admit an analytic solution? Finally, are there general structural conditions under which one can solve the QP for some range of κ?

Chapter 8: Conclusion

I have demonstrated that the stability of inference is a critical factor in the design and analysis of structured prediction. I derived generalization bounds that decrease with both the number of examples and the size of each example, meaning it is possible to learn from a few large examples, or even just one giant example. The bounds highlight the importance of inference stability as a sufficient condition for improved generalization. I also investigated the benefits of learning with strongly convex variational inference, which resulted in a new variational technique that can strongly convexify a variety of free energies. This technique yields better learned marginals due to the duality between strong convexity and stability, which once more underlines the significance of inference stability. The overriding theme to this work is that stable predictors and loss functions improve learning, both in theory in practice.

The effects of stability can be viewed a number of ways. Stability can be seen as promoting robustness to noise, which is essentially random, local perturbations. From the perspective of hypothesis complexity, stability can be viewed as regulating expressivity. In the bias-variance view of learning, stability controls the variance. These properties all contribute to better learning guarantees.

Yet, what this means for practitioners is more than just some elegant theory; it

means that we can do more with fewer examples. Though we are living in the so-called age of "big data," much of it is unlabeled. To train successful structured predictors, we can either develop better unsupervised or semi-supervised methods, or we can develop supervised methods that can effectively learn from limited data. I have focused on supervised learning and shown that inference stability is a criterion one can optimize to gain more accuracy from fewer examples.

8.1 Future Directions

One of the foremost challenges in modern artificial intelligence is *automated knowledge base construction*. This task typically involves extracting and reasoning about knowledge from public data sources, such as the World Wide Web and social media. Many leading approaches (e.g., Jiang et al., 2012; Pujara et al., 2013) use some form of collective reasoning to infer unknown facts and resolve mentions to their canonical entities. I posit that inference stability is critically important in this setting, due to the enormous scale of the knowledge representation and the fact that the data source is dynamic.

When constructing a knowledge base, the evidence grows and evolves as over time. Each newly integrated document prompts the system to update its inferences over the remaining unknowns.[1] Incorporating new evidence into collective reasoning is challenging; since a new observation can affect multiple related unknowns, updating inference can require recomputing all predictions. For example, suppose the knowledge base is represented by a large probalistic graphical model, and that inference involves comput-

[1] For simplicity of exposition, I assume that the model has already been learned and deployed. Though, in practice, the system may also update the model parameters.

ing the MAP assignment. Though there exists an algorithm for exactly[2] updating MAP assignments (Sümer et al., 2011), in the worst case, it is linear in the number of changed assignments. Thus, for applications with very large structure, such as knowledge base construction, updating inference can be very expensive.

However, suppose one could guarantee that the post-update predictions would not differ substantially from the pre-update predictions. Then one could guarantee that updating MAP inference would be efficient. Another option (proposed by Pujara et al., 2015) is to approximate the full inference update by conditioning on on a set of previous predictions, meaning only the active (i.e., unfixed) variables need to be updated. The complexity of this operation is linear in the size of the active set. If one could guarantee that the approximate update would not significantly differ from the full inference update, one could potentially achieve a massive speedup in update time with minimal approximation error. The dominant idea is that stable inference algorithms enable faster inference updates.

We should therefore design inference algorithms with collective stability guarantees. Of particular interest are algorithms that compute joint assignments, such as MAP inference. Though the theoretical insights of Section 7.1.2 suggest that MAP inference cannot attain *uniform* collective stability, perhaps there are conditions under which it is locally stable. Moreover, there may be approximate inference algorithms that compromise full collective reasoning so as to improve collective stability. For example, one could decompose the global inference optimization into independent local optimizations, similar to the technique of decomposed learning (Samdani and Roth, 2012). The study of collectively stable (approximate) MAP inference is a rich area of future research that

[2]To my knowledge, this algorithm does not apply to approximate MAP inference.

could have great impact on automated knowledge base construction and other large-scale applications of online structured prediction.

Appendix A: Technical Lemmas from Section 2.3.4

This appendix contains two technical lemmas deferred from Section 2.3.4. For the following, the potential functions are defined using the linear node features (Equation 2.7) and edge features (Equation 2.8).

Lemma 13. *Fix a graph, $G \triangleq (\mathcal{V}, \mathcal{E})$, with maximum degree Δ_G. Suppose $\mathcal{X}$ is uniformly bounded by the p-norm ball with radius R; i.e., $\sup_{x \in \mathcal{X}} \|x\|_p \leq R$. Then, for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}^n$ and $\mathbf{y} \in \mathcal{Y}^n$,*

$$\|\mathbf{f}(\mathbf{x}, \mathbf{y}) - \mathbf{f}(\mathbf{x}', \mathbf{y})\|_p \leq (\Delta_G + 2) R \, D_{\text{H}}(\mathbf{x}, \mathbf{x}'). \tag{A.1}$$

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

$$\|\mathbf{f}(\mathbf{x}, \mathbf{y}) - \mathbf{f}(\mathbf{x}', \mathbf{y})\|_p \leq 2R \, D_{\text{H}}(\mathbf{x}, \mathbf{x}'). \tag{A.2}$$

Proof We start by considering a pair, $\mathbf{x}, \mathbf{x}' \in \mathcal{X}^n : D_{\text{H}}(\mathbf{x}, \mathbf{x}') = 1$, that differ at a single coordinate, corresponding to a node i. This means that the aggregate features differ at one

local feature, and any edge involving i. Thus, using the triangle inequality, we have that

$$\begin{aligned}\|\mathbf{f}(\mathbf{x},\mathbf{y}) - \mathbf{f}(\mathbf{x}',\mathbf{y})\|_p &= \left\| \begin{bmatrix} f_i(\mathbf{x},\mathbf{y}) - f_i(\mathbf{x}',\mathbf{y}) \\ \sum_{j:\{i,j\}\in\mathcal{E}} f_{ij}(\mathbf{x},\mathbf{y}) - f_{ij}(\mathbf{x}',\mathbf{y}) \end{bmatrix} \right\|_p \\ &\leq \|f_i(\mathbf{x},\mathbf{y}) - f_i(\mathbf{x}',\mathbf{y})\|_p \\ &\quad + \sum_{j:\{i,j\}\in\mathcal{E}} \|f_{ij}(\mathbf{x},\mathbf{y}) - f_{ij}(\mathbf{x}',\mathbf{y})\|_p . \end{aligned} \tag{A.3}$$

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

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

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

and

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

Combining these inequalities with Equation A.3, and using the fact that i participates in at most Δ_G edges, we have that

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

For no edge observations, the righthand side is simply $2R$. Thus, since the bounds hold for any single coordinate perturbation, Equations A.1 and A.2 follow from the triangle inequality. ■

Lemma 14. *Fix a graph, $G \triangleq (\mathcal{V},\mathcal{E})$, and recall that $|G| \triangleq |\mathcal{V}| + |\mathcal{E}|$. Suppose $\mathcal{X}$ is uniformly bounded by the p-norm ball with radius R; i.e., $\sup_{x\in\mathcal{X}} \|x\|_p \leq R$. Then, for all $\mathbf{x} \in \mathcal{X}^n$ and $\mathbf{y} \in \mathcal{Y}^n$,*

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

Proof Invoking the triangle inequality, we have that

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

which completes the proof. ■

Appendix B: Proofs from Chapter 4

This appendix contains the deferred proofs from Section 6.5. We begin with some supplemental background in measure concentration. We then prove Proposition 1, and derive a concentration inequality implied by the result. We conclude with the proofs of Propositions 2 and 3.

B.1 The Method of Bounded Differences

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

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

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

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

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

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

Note that the order in which we condition on variables is arbitrary, and does not necessarily need to correspond to any spatio-temporal process. The important property is that the sequence of σ-algebras generated by the conditioned variables are *nested* (McDiarmid (1998) called this a *filter*), which is guaranteed by the construction of $(V_1, \ldots, V_n)$.

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

Lemma 15. *If ξ is a random variable, such that $\mathbb{E}[\xi] = 0$ and $a \leq \xi \leq b$ almost surely, then for any $\tau \in \mathbb{R}$,*

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

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

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

then

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

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

B.2 Coupling

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

Using a construction due to Fiebig (1993), one can create a coupling of two sequences of random variables, such that the probability that any two corresponding variables are different is upper-bounded by the ϑ-mixing coefficients in Definition 9. The following is an adaptation of this result (due to Samson, 2000) for continuous domains.

Lemma 16. *Let $\mathbf{Z}^{(1)} \triangleq (Z_i^{(1)})_{i=1}^n$ and $\mathbf{Z}^{(2)} \triangleq (Z_i^{(2)})_{i=1}^n$ be random variables with respective distributions $\mathbb{D}_1$ and $\mathbb{D}_2$ over a sample space $\mathcal{Z}^n$. Then there exists a coupling $\hat{\mathbb{D}}$, with marginal distributions $\hat{\mathbb{D}}(\mathbf{Z}^{(1)}) = \mathbb{D}_1(\mathbf{Z}^{(1)})$ and $\hat{\mathbb{D}}(\mathbf{Z}^{(2)}) = \mathbb{D}_2(\mathbf{Z}^{(2)})$, such that, for any $i \in [n]$,*

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

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

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

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

B.3 Proof of Moment-Generating Function Bound (Proposition 1)

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

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

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

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

We therefore have, via Equation B.1, that

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

where the supremum over $\mathbf{z} \in \mathcal{Z}_{\overline{\mathcal{B}}}^{i-1}$ ensures that $\mathbf{Z}_{\pi_i(1:i-1)} = \mathbf{z}$ does not contradict $\overline{\mathcal{B}}$. Recall that each permutation in $\boldsymbol{\pi}$ has the same prefix, thus preserving the order of

conditioned variables, and ensuring that the sequence of σ-algebras is nested.

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

$$\sup_{\mathbf{z} \in \mathcal{Z}_{\overline{\mathcal{B}}}^{i-1}} \sup_{z \in \mathcal{Z}_{\overline{\mathcal{B}}}} V_i^{\boldsymbol{\pi}} - \inf_{z' \in \mathcal{Z}_{\overline{\mathcal{B}}}} V_i^{\boldsymbol{\pi}}$$

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

is bounded, so as to apply Lemma 15. (Again, the supremum over $z, z' \in \mathcal{Z}_{\overline{\mathcal{B}}}$ ensures consistency between $\mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z)$, $\mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z')$ and $\overline{\mathcal{B}}$.) To do so, we will use the coupling construction from Lemma 16. Fix any $\mathbf{z} \in \mathcal{Z}_{\overline{\mathcal{B}}}^{i-1}$ and $z, z' \in \mathcal{Z}_{\overline{\mathcal{B}}}$, and let $N \triangleq n - i$. Define random variables $\boldsymbol{\xi}^{(1)} \triangleq (\xi_j^{(1)})_{j=1}^N$ and $\boldsymbol{\xi}^{(2)} \triangleq (\xi_j^{(2)})_{j=1}^N$, with coupling distribution $\hat{\mathbb{D}}$ such that

$$\hat{\mathbb{D}}\big(\boldsymbol{\xi}^{(1)}\big) \triangleq \mathbb{D}\left(\mathbf{Z}_{\pi_i(i+1:n)} \,|\, \overline{\mathcal{B}}, \mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z)\right)$$

$$\text{and} \quad \hat{\mathbb{D}}\big(\boldsymbol{\xi}^{(2)}\big) \triangleq \mathbb{D}\left(\mathbf{Z}_{\pi_i(i+1:n)} \,|\, \overline{\mathcal{B}}, \mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z')\right). \tag{B.3}$$

In other words, the marginal distributions of $\boldsymbol{\xi}^{(1)}$ and $\boldsymbol{\xi}^{(2)}$ are equal to the conditional distributions of $\mathbf{Z}_{\pi_i(i+1:n)}$ given $\overline{\mathcal{B}}$ and, respectively, $\mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z)$ or $\mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z')$. Note that we have renumbered the coupled variables according to π_i. This does not affect the distribution, but it does affect how we later apply Lemma 16. Denote by π_i^{-1} the inverse of π_i (i.e., $\pi_i^{-1}(\pi_i(1:n)) = [n]$), and let

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

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

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

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

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

Because the expectations are conditioned on $\overline{\mathcal{B}}$, both realizations, $(\mathbf{z}, z, \boldsymbol{\xi}^{(1)})$ and $(\mathbf{z}, z', \boldsymbol{\xi}^{(2)})$, are "good," in the sense that Equation 3.1 holds. We therefore have that

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

In the second inequality, we assumed that $z \neq z'$. Recall from Lemma 16 and Definition 9

that

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

This holds uniformly for all valid $\mathbf{z} \in \mathcal{Z}_{\overline{\mathcal{B}}}^{i-1}$ and $z, z' \in \mathcal{Z}_{\overline{\mathcal{B}}}$; thus,

$$
\sup_{\mathbf{z}\in\mathcal{Z}^{i-1}} \sup_{z\in\mathcal{Z}} V_i^{\boldsymbol{\pi}} - \inf_{z'\in\mathcal{Z}} V_i^{\boldsymbol{\pi}} = \sup_{\substack{\mathbf{z}\in\mathcal{Z}^{i-1} \\ z,z'\in\mathcal{Z}}} \mathbb{E}\left[\psi\big(\mathbf{z}, z, \boldsymbol{\xi}^{(1)}\big) - \psi\big(\mathbf{z}, z', \boldsymbol{\xi}^{(2)}\big) \right] \leq \beta \sum_{j=i}^{n} \gamma_{ij}^{\boldsymbol{\pi}}.
$$

Then, since we have identified a uniform upper bound for Equation B.2, we apply Lemma 15 and obtain

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

which completes the proof.

B.4 Proof of Concentration Inequality (Corollary 1)

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

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

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

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

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

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

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

B.5 Proof of Proposition 2

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

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

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

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

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

which completes the proof.

B.6 Proof of Proposition 3

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

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

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

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

by definition.

Appendix C: Proofs from Chapter 5

This appendix contains the deferred proofs from Chapter 5.

C.1 Proof of Lemma 2

For any assignments $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}^n$, let $\mathcal{I} \triangleq \{i \in [n] : z_i \neq z'_i\}$ denote the set of coordinates at which their values differ. By definition, for any $h \in \mathcal{H}$,

$$\begin{aligned}\|(c \circ h)(\mathbf{z}) - (c \circ h)(\mathbf{z}')\|_1 &= \sum_{j=1}^{n} \left|c(y_j, h_j(\mathbf{x})) - c(y'_j, h_j(\mathbf{x}'))\right| \\ &= \sum_{i \in \mathcal{I}} |c(y_i, h_i(\mathbf{x})) - c(y'_i, h_i(\mathbf{x}'))| \\ &\quad + \sum_{j \notin \mathcal{I}} |c(y_j, h_j(\mathbf{x})) - c(y_j, h_j(\mathbf{x}'))| \,. \end{aligned} \tag{C.1}$$

Focusing on the first sum, for any $i \in \mathcal{I}$, we have via the first admissibility condition that

$$\begin{aligned}|c(y_i, h_i(\mathbf{x})) - c(y'_i, h_i(\mathbf{x}'))| &\leq |c(y_i, h_i(\mathbf{x})) - c(y_i, h_i(\mathbf{x}'))| + |c(y_i, h_i(\mathbf{x}')) - c(y'_i, h_i(\mathbf{x}'))| \\ &\leq |c(y_i, h_i(\mathbf{x})) - c(y_i, h_i(\mathbf{x}'))| + M.\end{aligned}$$

Therefore,

$$\sum_{i\in\mathcal{I}} |c(y_i, h_i(\mathbf{x})) - c(y_i', h_i(\mathbf{x}'))| \leq M\,|\mathcal{I}| + \sum_{i\in\mathcal{I}} |c(y_i, h_i(\mathbf{x})) - c(y_i, h_i(\mathbf{x}'))|$$
$$= M\,D_{\text{H}}(\mathbf{z}, \mathbf{z}') + \sum_{i\in\mathcal{I}} |c(y_i, h_i(\mathbf{x})) - c(y_i, h_i(\mathbf{x}'))|\,.$$

Combining this with Equation C.1, we have that

$$\|(c\circ h)(\mathbf{z}) - (c\circ h)(\mathbf{z}')\|_1 \leq M\,D_{\text{H}}(\mathbf{z}, \mathbf{z}') + \sum_{j=1}^{n} |c(y_j, h_j(\mathbf{x})) - c(y_j, h_j(\mathbf{x}'))|$$
$$\leq M\,D_{\text{H}}(\mathbf{z}, \mathbf{z}') + \lambda\,\|h(\mathbf{x}) - h(\mathbf{x}')\|_1$$
$$\leq M\,D_{\text{H}}(\mathbf{z}, \mathbf{z}') + \lambda\,\beta\,D_{\text{H}}(\mathbf{z}, \mathbf{z}'),$$

where we have used the second admissibility condition, then uniform collective stability.

C.2 Proof of Lemma 3

By definition, c_ρ is bounded by [0,1], which trivially establishes that it is 1-uniformly range-bounded, thereby satisfying the first admissibility condition. Fix any μ and μ', and let

$$u \triangleq \underset{y'\in\mathcal{Y}:y\neq y'}{\arg\max}\,\langle y', \mu\rangle \quad \text{and} \quad u' \triangleq \underset{y'\in\mathcal{Y}:y\neq y'}{\arg\max}\,\langle y', \mu'\rangle\,.$$

Without loss of generality, assume that

$$\langle y, \mu\rangle - \langle u, \mu\rangle \geq \langle y, \mu'\rangle - \langle u', \mu'\rangle\,.$$

Then, for any $y \in \mathcal{Y}$, we have that

$$\begin{aligned}
|(\langle y, \mu \rangle - \langle u, \mu \rangle) - (\langle y, \mu' \rangle - \langle u', \mu' \rangle)| &= \langle y, \mu - \mu' \rangle + \langle u', \mu' \rangle - \langle u, \mu \rangle \\
&\leq \langle y, \mu - \mu' \rangle + \langle u', \mu' \rangle - \langle u', \mu \rangle \\
&= \langle y - u', \mu - \mu' \rangle \\
&\leq \|y - u'\|_\infty \|\mu - \mu'\|_1 \\
&\leq \|\mu - \mu'\|_1 .
\end{aligned}$$

Further, for any $\delta, \delta' \in \mathbb{R}$,

$$|r_\rho(\delta) - r_\rho(\delta')| \leq \left| \frac{1-\delta}{\rho} - \frac{1-\delta'}{\rho} \right| = \frac{1}{\rho} |\delta - \delta'| .$$

Combining these inequalities, we have that

$$|c_\rho(y, \mu) - c_\rho(y, \mu')| \leq \frac{1}{\rho} \|\mu - \mu'\|_1 ,$$

which establishes the second admissibility condition.

C.3 Proof of Lemma 4

Fix any hypothesis, $h \in \mathcal{H}^1_{\text{SC}}$, with weights $\mathbf{w}$. Fix any two inputs, $\mathbf{x}, \mathbf{x}' \in \mathcal{X}^n$, and let

$$\tilde{\boldsymbol{\mu}}(\mathbf{x}) \triangleq \min_{\tilde{\boldsymbol{\mu}} \in \tilde{\mathcal{M}}} \tilde{E}(\tilde{\boldsymbol{\mu}} \,|\, \mathbf{x}; \mathbf{w}) \quad \text{and} \quad \tilde{\boldsymbol{\mu}}(\mathbf{x}') \triangleq \min_{\tilde{\boldsymbol{\mu}} \in \tilde{\mathcal{M}}} \tilde{E}(\tilde{\boldsymbol{\mu}} \,|\, \mathbf{x}'; \mathbf{w})$$

denote their respective conditional pseudomarginals. By assumption, the variational free energy,

$$\tilde{E}(\tilde{\boldsymbol{\mu}} \,|\, \mathbf{x}; \mathbf{w}) = -\mathbf{w} \cdot \mathbf{f}(\mathbf{x}, \tilde{\boldsymbol{\mu}}) + \tilde{\Phi}^*(\tilde{\boldsymbol{\mu}}),$$

uses a conjugate function, $\tilde{\Phi}^*$, that is κ-strongly convex with respect to the 1-norm. Therefore, applying Lemma 1, we have that

$$\begin{aligned}
\|\tilde{\boldsymbol{\mu}}(\mathbf{x}) - \tilde{\boldsymbol{\mu}}(\mathbf{x}')\|_1^2 &= \frac{1}{2} \|\tilde{\boldsymbol{\mu}}(\mathbf{x}') - \tilde{\boldsymbol{\mu}}(\mathbf{x})\|_1^2 + \frac{1}{2} \|\tilde{\boldsymbol{\mu}}(\mathbf{x}) - \tilde{\boldsymbol{\mu}}(\mathbf{x}')\|_1^2 \\
&\leq \frac{1}{\kappa} \left(\tilde{E}(\tilde{\boldsymbol{\mu}}(\mathbf{x}') \,|\, \mathbf{x}; \mathbf{w}) - \tilde{E}(\tilde{\boldsymbol{\mu}}(\mathbf{x}) \,|\, \mathbf{x}; \mathbf{w}) \right) \\
&\quad + \frac{1}{\kappa} \left(\tilde{E}(\tilde{\boldsymbol{\mu}}(\mathbf{x}) \,|\, \mathbf{x}'; \mathbf{w}) - \tilde{E}(\tilde{\boldsymbol{\mu}}(\mathbf{x}') \,|\, \mathbf{x}'; \mathbf{w}) \right) \\
&= \frac{1}{\kappa} \left(-\mathbf{w} \cdot (\mathbf{f}(\mathbf{x}, \tilde{\boldsymbol{\mu}}(\mathbf{x}')) - \mathbf{f}(\mathbf{x}, \tilde{\boldsymbol{\mu}}(\mathbf{x}))) + \tilde{\Phi}^*(\tilde{\boldsymbol{\mu}}(\mathbf{x}')) - \tilde{\Phi}^*(\tilde{\boldsymbol{\mu}}(\mathbf{x})) \right) \\
&\quad + \frac{1}{\kappa} \left(-\mathbf{w} \cdot (\mathbf{f}(\mathbf{x}', \tilde{\boldsymbol{\mu}}(\mathbf{x})) - \mathbf{f}(\mathbf{x}', \tilde{\boldsymbol{\mu}}(\mathbf{x}'))) + \tilde{\Phi}^*(\tilde{\boldsymbol{\mu}}(\mathbf{x})) - \tilde{\Phi}^*(\tilde{\boldsymbol{\mu}}(\mathbf{x}')) \right) \\
&= \frac{1}{\kappa} \mathbf{w} \cdot (\mathbf{f}(\mathbf{x}, \tilde{\boldsymbol{\mu}}(\mathbf{x})) - \mathbf{f}(\mathbf{x}', \tilde{\boldsymbol{\mu}}(\mathbf{x}))) \\
&\quad + \frac{1}{\kappa} \mathbf{w} \cdot (\mathbf{f}(\mathbf{x}', \tilde{\boldsymbol{\mu}}(\mathbf{x}')) - \mathbf{f}(\mathbf{x}, \tilde{\boldsymbol{\mu}}(\mathbf{x}'))) \\
&\leq \frac{1}{\kappa} \|\mathbf{w}\|_2 \|\mathbf{f}(\mathbf{x}, \tilde{\boldsymbol{\mu}}(\mathbf{x})) - \mathbf{f}(\mathbf{x}', \tilde{\boldsymbol{\mu}}(\mathbf{x}))\|_2 \\
&\quad + \frac{1}{\kappa} \|\mathbf{w}\|_2 \|\mathbf{f}(\mathbf{x}', \tilde{\boldsymbol{\mu}}(\mathbf{x}')) - \mathbf{f}(\mathbf{x}, \tilde{\boldsymbol{\mu}}(\mathbf{x}'))\|_2 .
\end{aligned}$$

The last inequality uses Cauchy-Schwarz. Via Lemma 13 (Equation A.1) since we assume that $\sup_{x \in \mathcal{X}} \|x\|_2 \leq 1$, we have that

$$\|\mathbf{f}(\mathbf{x}, \tilde{\boldsymbol{\mu}}(\mathbf{x})) - \mathbf{f}(\mathbf{x}', \tilde{\boldsymbol{\mu}}(\mathbf{x}))\|_2 \leq (\Delta_G + 2)\, D_{\text{H}}(\mathbf{x}, \mathbf{x}')$$

$$\text{and} \quad \|\mathbf{f}(\mathbf{x}', \tilde{\boldsymbol{\mu}}(\mathbf{x}')) - \mathbf{f}(\mathbf{x}, \tilde{\boldsymbol{\mu}}(\mathbf{x}'))\|_2 \leq (\Delta_G + 2)\, D_{\text{H}}(\mathbf{x}, \mathbf{x}').$$

(The fact that we use the pseudomarginals instead of the basis vectors does not matter; what matters is that both output types obey the simplex constraint.) Then, leveraging the fact that $\|\mathbf{w}\|_2 \leq 1$, we have that

$$\|\tilde{\boldsymbol{\mu}}(\mathbf{x}) - \tilde{\boldsymbol{\mu}}(\mathbf{x}')\|_1^2 \leq \frac{2}{\kappa}\left(\Delta_G + 2\right) D_{\text{H}}(\mathbf{x}, \mathbf{x}').$$

Taking the square root of both sides, and noting that $\sqrt{D_{\text{H}}(\mathbf{x}, \mathbf{x}')} \leq D_{\text{H}}(\mathbf{x}, \mathbf{x}')$, completes the proof.

C.4 Proof of Lemma 5

The proof requires the following technical lemma.

Lemma 17. *The d-dimensional hypercube, $[0, \Lambda]^d$, admits an ϵ-cover, under the 1-norm, of cardinality $\left\lceil \left(\frac{\Lambda d}{2\epsilon}\right)^d \right\rceil$.*

Proof Partition the hypercube into k smaller hypercube *cells*, each with edge length $(2\epsilon/d)$. It is straightforward to show that any point within each cell is within L^1 distance ϵ from the center of the cell. Therefore, the hypercube is ϵ-covered by the centers of the cells. To find the minimum number of cells needed, we let $k(2\epsilon/d)^d \geq \Lambda^d$ and solve for k, then round up to the nearest integer. ■

Note that $\mathcal{H}^1_{\text{sc}}$ is essentially just the d-dimensional unit ball of weight vectors. We will construct a cover of this space, effectively discretizing the hypothesis space. Then, using the strong convexity of the variational free energy—similar to the same way we did

in the proof of Lemma 4—we will show that this discretization covers the output space.

The unit ball has diameter 2, and is thus contained in a hypercube with side length 2. Therefore, to ϵ-cover the unit ball, it suffices to construct an ϵ-covering of the hypercube $[0, 2]^d$, then translate the points by -1 in all dimensions and take the intersection with the unit ball. For some ϵ' to be defined later, let $\mathcal{C} \subseteq \{\mathbf{w}' \in \mathbb{R}^d : \|\mathbf{w}'\|_2 \leq 1\}$ denote this covering. By definition, every $\mathbf{w} \in \mathcal{H}^1_{\text{sc}}$ is at most ϵ' L^1 distance from some $\mathbf{w}' \in \mathcal{C}$. Further, by Lemma 17, $|\mathcal{C}| \leq \left\lceil \left(\frac{d}{\epsilon'}\right)^d \right\rceil$.

Fix any $\mathbf{w} \in \mathcal{H}^1_{\text{sc}}$, and let $\mathbf{w}' \in \mathcal{C}$ denote its closest vector. For any input, $\mathbf{x} \in \mathcal{X}^n$, let

$$\tilde{\boldsymbol{\mu}}(\mathbf{w}) \triangleq \min_{\tilde{\boldsymbol{\mu}} \in \tilde{\mathcal{M}}} \tilde{E}(\tilde{\boldsymbol{\mu}} \,|\, \mathbf{x}; \mathbf{w}) \quad \text{and} \quad \tilde{\boldsymbol{\mu}}(\mathbf{w}') \triangleq \min_{\tilde{\boldsymbol{\mu}} \in \tilde{\mathcal{M}}} \tilde{E}(\tilde{\boldsymbol{\mu}} \,|\, \mathbf{x}; \mathbf{w}')$$

denote the respective conditional pseudomarginals. Since the variational free energy, $\tilde{E}$, is κ-strongly convex, we can apply Lemma 1, like we did in the proof of Lemma 4, and obtain

$$\begin{aligned}
\|\tilde{\boldsymbol{\mu}}(\mathbf{w}) - \tilde{\boldsymbol{\mu}}(\mathbf{w}')\|_1^2 &= \frac{1}{2} \|\tilde{\boldsymbol{\mu}}(\mathbf{w}') - \tilde{\boldsymbol{\mu}}(\mathbf{w})\|_1^2 + \frac{1}{2} \|\tilde{\boldsymbol{\mu}}(\mathbf{w}) - \tilde{\boldsymbol{\mu}}(\mathbf{w}')\|_1^2 \\
&\leq \frac{1}{\kappa} \Big(\tilde{E}(\tilde{\boldsymbol{\mu}}(\mathbf{w}') \,|\, \mathbf{x}; \mathbf{w}) - \tilde{E}(\tilde{\boldsymbol{\mu}}(\mathbf{w}) \,|\, \mathbf{x}; \mathbf{w}) \Big) \\
&\quad + \frac{1}{\kappa} \Big(\tilde{E}(\tilde{\boldsymbol{\mu}}(\mathbf{w}) \,|\, \mathbf{x}; \mathbf{w}') - \tilde{E}(\tilde{\boldsymbol{\mu}}(\mathbf{w}') \,|\, \mathbf{x}; \mathbf{w}') \Big) \\
&= \frac{1}{\kappa} \Big(-\mathbf{w} \cdot (\mathbf{f}(\mathbf{x}, \tilde{\boldsymbol{\mu}}(\mathbf{w}')) - \mathbf{f}(\mathbf{x}, \tilde{\boldsymbol{\mu}}(\mathbf{w}))) + \tilde{\Phi}^*(\tilde{\boldsymbol{\mu}}(\mathbf{w}')) - \tilde{\Phi}^*(\tilde{\boldsymbol{\mu}}(\mathbf{w})) \Big) \\
&\quad + \frac{1}{\kappa} \Big(-\mathbf{w}' \cdot (\mathbf{f}(\mathbf{x}, \tilde{\boldsymbol{\mu}}(\mathbf{w})) - \mathbf{f}(\mathbf{x}, \tilde{\boldsymbol{\mu}}(\mathbf{w}'))) + \tilde{\Phi}^*(\tilde{\boldsymbol{\mu}}(\mathbf{w})) - \tilde{\Phi}^*(\tilde{\boldsymbol{\mu}}(\mathbf{w}')) \Big) \\
&= \frac{1}{\kappa} (\mathbf{w} - \mathbf{w}') \cdot \mathbf{f}(\mathbf{x}, \tilde{\boldsymbol{\mu}}(\mathbf{w})) + \frac{1}{\kappa} (\mathbf{w}' - \mathbf{w}) \cdot \mathbf{f}(\mathbf{x}, \tilde{\boldsymbol{\mu}}(\mathbf{w}')) \\
&\leq \frac{1}{\kappa} \|\mathbf{w} - \mathbf{w}'\|_1 \|\mathbf{f}(\mathbf{x}, \tilde{\boldsymbol{\mu}}(\mathbf{w}))\|_\infty + \frac{1}{\kappa} \|\mathbf{w}' - \mathbf{w}\|_1 \|\mathbf{f}(\mathbf{x}, \tilde{\boldsymbol{\mu}}(\mathbf{w}'))\|_\infty
\end{aligned}$$

The last line uses Hölder's inequality. We have constructed $\mathcal{C}$ such that $\|\mathbf{w} - \mathbf{w}'\|_1 \leq \epsilon'$. To bound $\|\mathbf{f}(\mathbf{x}, \tilde{\boldsymbol{\mu}}(\mathbf{w}))\|_\infty$ and $\|\mathbf{f}(\mathbf{x}, \tilde{\boldsymbol{\mu}}(\mathbf{w}'))\|_\infty$, we could use Lemma 14, but this yields a loose upper bound for the ∞-norm. We will therefore use a special analysis.

Observe that

$$\begin{aligned}\|\mathbf{f}(\mathbf{x}, \tilde{\boldsymbol{\mu}})\|_\infty &= \left\| \begin{bmatrix} \sum_{i \in \mathcal{V}} f_i(\mathbf{x}, \tilde{\boldsymbol{\mu}}) \\ \sum_{\{i,j\} \in \mathcal{E}} f_{ij}(\mathbf{x}, \tilde{\boldsymbol{\mu}}) \end{bmatrix} \right\|_\infty \\ &= \max \left\{ \left\| \sum_{i \in \mathcal{V}} f_i(\mathbf{x}, \tilde{\boldsymbol{\mu}}) \right\|_\infty, \left\| \sum_{\{i,j\} \in \mathcal{E}} f_{ij}(\mathbf{x}, \tilde{\boldsymbol{\mu}}) \right\|_\infty \right\}\end{aligned}$$

Since $\|x\|_2 \leq 1$, and $\|\tilde{\mu}\|_1 = 1$, it can be shown, using an analysis similar to Lemma 14, that

$$\left\| \sum_{i \in \mathcal{V}} f_i(\mathbf{x}, \tilde{\boldsymbol{\mu}}) \right\|_\infty \leq |\mathcal{V}| \quad \text{and} \quad \left\| \sum_{\{i,j\} \in \mathcal{E}} f_{ij}(\mathbf{x}, \tilde{\boldsymbol{\mu}}) \right\|_\infty \leq |\mathcal{E}| \,.$$

Since G has maximum degree Δ_G, $|\mathcal{E}| \leq n\Delta_G$. For a graph with at least one edge, $\Delta_G \geq 1$, meaning $|\mathcal{V}| = n \leq n\Delta_G$. We can therefore use

$$\|\mathbf{f}(\mathbf{x}, \tilde{\boldsymbol{\mu}})\|_\infty \leq n\Delta_G.$$

Putting the pieces together, we have that

$$\|\tilde{\boldsymbol{\mu}}(\mathbf{w}) - \tilde{\boldsymbol{\mu}}(\mathbf{w}')\|_1 \leq \sqrt{\frac{2n\Delta_G\epsilon'}{\kappa}}.$$

To satisfy Definition 12, we need to show that, for every $\mathbf{w} \in \mathcal{H}^1_{\text{SC}}$, there exists a $\mathbf{w}' \in \mathcal{C}$

such that

$$\sup_{\mathbf{x}\in\mathcal{X}^n} \frac{1}{n\left|\mathcal{Y}\right|} \left\|\tilde{\boldsymbol{\mu}}_{\mathrm{u}}(\mathbf{w}) - \tilde{\boldsymbol{\mu}}_{\mathrm{u}}(\mathbf{w}')\right\|_1 \leq \epsilon,$$

where $\tilde{\boldsymbol{\mu}}_{\mathrm{u}}$ selects only the unary clique (i.e., node) pseudomarginals. (The $n\left|\mathcal{Y}\right|$ in the denominator is the length of the resulting pseudomarginal vector.) If we set

$$\epsilon' \triangleq \frac{\kappa n \epsilon^2 \left|\mathcal{Y}\right|^2}{2\Delta_G},$$

then

$$\begin{aligned}\sup_{\mathbf{x}\in\mathcal{X}^n} \frac{1}{n\left|\mathcal{Y}\right|} \left\|\tilde{\boldsymbol{\mu}}_{\mathrm{u}}(\mathbf{w}) - \tilde{\boldsymbol{\mu}}_{\mathrm{u}}(\mathbf{w}')\right\|_1 &\leq \sup_{\mathbf{x}\in\mathcal{X}^n} \frac{1}{n\left|\mathcal{Y}\right|} \left\|\tilde{\boldsymbol{\mu}}(\mathbf{w}) - \tilde{\boldsymbol{\mu}}(\mathbf{w}')\right\|_1 \\ &\leq \frac{1}{n\left|\mathcal{Y}\right|} \sqrt{\frac{2n\Delta_G}{\kappa} \cdot \frac{\kappa n \epsilon^2 \left|\mathcal{Y}\right|^2}{2\Delta_G}} = \epsilon.\end{aligned}$$

Thus, there exists a set, $\mathcal{C}$, that ϵ-covers $\mathcal{H}^1_{\mathrm{SC}}$, with size $|\mathcal{C}| \leq \left\lceil \left(\frac{2d\Delta_G}{\kappa n \epsilon^2 |\mathcal{Y}|^2}\right)^d \right\rceil$.

Appendix D: Proofs from Chapter 6

This appendix contains the deferred proofs from Chapter 6.

D.1 Proof of Theorem 3

For $i = 0, 1, 2, \ldots$, let $\beta_i \triangleq 2^{i+1}$. Since Equation 6.1 fails with probability $\delta + m\nu$, one could simply invoke Theorem 2 for each β_i with $\delta_i \triangleq \beta_i^{-1}(\delta + m\nu)$. This approach would introduce an additional $\mathrm{O}\left(\ln(m\nu)^{-1}\right)$ term in the numerator of Equation 6.22. I therefore choose instead to cover β and u simultaneously. Accordingly, for $j = 0, 1, 2, \ldots$, let

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

Each β_i defines a set of "bad" hypotheses, $\mathcal{B}_{\mathcal{H}}^i$, which we use in Equation 6.4 to define a function $\tilde{\phi}_i$. Let $\delta_{ij} \triangleq \delta \beta_i^{-1} 2^{-(j+1)}$, and define an event

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

Note that none of the above depend on $(\beta, \eta, \mathbb{Q})$. Using the event B defined in Equation 6.11, we have, via Proposition 1, that

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

Then, using the same reasoning as Equation 6.12, with $E \triangleq \bigcup_{i=0}^{\infty} \bigcup_{j=0}^{\infty} E_{ij}$,

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

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

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

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

$(\beta_{i^\star}/n, \mathcal{B}_{\mathcal{Z}}, \eta)$-local stability. Then, letting

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

we have that

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

Moreover,

$$\begin{aligned} D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{1}{\delta_{i^\star j^\star}} &\le D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2\beta_{i^\star}}{\delta} + \frac{1}{2}\ln\left(\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})}{\ln(2\beta_{i^\star}/\delta)} + 1\right) \\ &\le D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2\beta_{i^\star}}{\delta} + \frac{1}{2}\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2\beta_{i^\star}}{\delta}\right). \end{aligned} \tag{D.3}$$

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

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

The first inequality uses Equation 6.8; the second uses Equation D.1; the third and fourth use Equations D.2 and D.3. Noting that $\beta_{i^\star} \le 2\beta$ completes the proof.

D.2 Proof of Proposition 5

Fix any $h \in \mathcal{H}$ and $\mathbf{z} \notin \mathcal{B}_{\mathcal{Z}}$. By Definition 17, there exists a set $\mathcal{B}_{\mathcal{H}}(h)$ with measure $\mathbb{Q}_h(\mathcal{B}_{\mathcal{H}}(h)) \leq \eta$. For any $\mathbf{z} \notin \mathcal{B}_{\mathcal{Z}}$, let $\mathcal{B}_{\mathcal{H}}(h, \mathbf{z}) \triangleq \mathcal{B}_{\mathcal{H}}(h)$, and note that $\mathbb{Q}_h(\mathcal{B}_{\mathcal{H}}(h, \mathbf{z})) \leq \eta$ as well. Further, for any $h' \notin \mathcal{B}_{\mathcal{H}}(h, \mathbf{z})$, $\|h - h'\| \leq \beta$. Thus, by Definition 16,

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

which completes the proof.

D.3 Proof of Lemma 7

To simplify notation, let:

$$\mathbf{y}_1 \triangleq \arg\max_{\mathbf{u} \in \mathcal{Y}^n} D_{\text{H}}(\mathbf{y}, \mathbf{u}) + h(\mathbf{x}, \mathbf{u}); \qquad \mathbf{y}_2 \triangleq \arg\max_{\mathbf{u} \in \mathcal{Y}^n} h(\mathbf{x}, \mathbf{u});$$

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

Using this notation, we have that

$$\begin{aligned} & n\,|L_{\text{r}}(h, \mathbf{z}) - L_{\text{r}}(h, \mathbf{z}')| \\ & \quad = |(D_{\text{H}}(\mathbf{y}, \mathbf{y}_1) + h(\mathbf{x}, \mathbf{y}_1) - h(\mathbf{x}, \mathbf{y}_2)) - (D_{\text{H}}(\mathbf{y}', \mathbf{y}_1') + h(\mathbf{x}', \mathbf{y}_1') - h(\mathbf{x}', \mathbf{y}_2'))| \\ & \quad \leq |(D_{\text{H}}(\mathbf{y}, \mathbf{y}_1) + h(\mathbf{x}, \mathbf{y}_1)) - (D_{\text{H}}(\mathbf{y}', \mathbf{y}_1') + h(\mathbf{x}', \mathbf{y}_1'))| + |h(\mathbf{x}, \mathbf{y}_2) - h(\mathbf{x}', \mathbf{y}_2')|\,, \end{aligned} \tag{D.4}$$

using the triangle inequality.

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

$$\begin{aligned}
|h(\mathbf{x}, \mathbf{y}_2) - h(\mathbf{x}', \mathbf{y}'_2)| &= h(\mathbf{x}, \mathbf{y}_2) - h(\mathbf{x}', \mathbf{y}'_2) \\
&\leq h(\mathbf{x}, \mathbf{y}_2) - h(\mathbf{x}', \mathbf{y}_2) \\
&= \mathbf{w} \cdot (\mathbf{f}(\mathbf{x}, \mathbf{y}_2) - \mathbf{f}(\mathbf{x}', \mathbf{y}_2)) \\
&\leq \|\mathbf{w}\|_q \, \|\mathbf{f}(\mathbf{x}, \mathbf{y}_2) - \mathbf{f}(\mathbf{x}', \mathbf{y}_2)\|_p \\
&\leq \|\mathbf{w}\|_q \, (\Delta_G + 2) R \, D_{\text{H}}(\mathbf{x}, \mathbf{x}'). \qquad \text{(D.5)}
\end{aligned}$$

The first inequality uses the optimality of $\mathbf{y}'_2$, implying $-h(\mathbf{x}', \mathbf{y}'_2) \leq -h(\mathbf{x}', \mathbf{y}_2)$; the second inequality uses Hölder's inequality; the third inequality uses Lemma 13 (Equation A.1). Note that we obtain the same upper bound if we assume that $h(\mathbf{x}, \mathbf{y}_2) \leq h(\mathbf{x}', \mathbf{y}'_2)$, since we can reverse the terms inside the absolute value and proceed with $\mathbf{y}'_2$ instead of $\mathbf{y}_2$.

We now return to the first absolute difference. To reduce clutter, it will help to use the loss-augmented potentials, $\tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w})$, from Equation 6.30. Recall that $\delta(\mathbf{y})$ denotes the loss augmentation vector for $\mathbf{y}$. We then have that

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

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

$$\begin{aligned}
\left|\tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}) \cdot \hat{\mathbf{y}}_1 - \tilde{\boldsymbol{\theta}}(\mathbf{x}', \mathbf{y}'; \mathbf{w}) \cdot \hat{\mathbf{y}}'_1\right| &= \tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}) \cdot \hat{\mathbf{y}}_1 - \tilde{\boldsymbol{\theta}}(\mathbf{x}', \mathbf{y}'; \mathbf{w}) \cdot \hat{\mathbf{y}}'_1 \\
&\leq \tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}) \cdot \hat{\mathbf{y}}_1 - \tilde{\boldsymbol{\theta}}(\mathbf{x}', \mathbf{y}'; \mathbf{w}) \cdot \hat{\mathbf{y}}_1 \\
&= (\boldsymbol{\theta}(\mathbf{x}; \mathbf{w}) + \delta(\mathbf{y}) - \boldsymbol{\theta}(\mathbf{x}'; \mathbf{w}) - \delta(\mathbf{y}')) \cdot \hat{\mathbf{y}}_1 \\
&= \mathbf{w} \cdot (\mathbf{f}(\mathbf{x}, \mathbf{y}_1) - \mathbf{f}(\mathbf{x}', \mathbf{y}_1)) + (\delta(\mathbf{y}) - \delta(\mathbf{y}')) \cdot \hat{\mathbf{y}}_1 \\
&\leq \|\mathbf{w}\|_q (\Delta_G + 2) R \, D_{\text{H}}(\mathbf{x}, \mathbf{x}') + (\delta(\mathbf{y}) - \delta(\mathbf{y}')) \cdot \hat{\mathbf{y}}_1 \\
&\leq \|\mathbf{w}\|_q (\Delta_G + 2) R \, D_{\text{H}}(\mathbf{x}, \mathbf{x}') + D_{\text{H}}(\mathbf{y}, \mathbf{y}'). \quad \text{(D.6)}
\end{aligned}$$

The first inequality uses the optimality of $\mathbf{y}'_1$; the second inequality uses Hölder's inequality and Lemma 13 again; the last inequality uses the fact that

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

The upper bound in Equation D.6 also holds when $\tilde{\boldsymbol{\theta}}(\mathbf{x}', \mathbf{y}'; \mathbf{w}) \cdot \hat{\mathbf{y}}'_1 \geq \tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}) \cdot \hat{\mathbf{y}}_1$.

Combining Equations D.5 to D.7, we then have that

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

Dividing both sides by n yields Equation 6.32. To obtain Equation 6.33, we use Lemma 13's Equation A.2 in Equations D.5 and D.6, which reduces the term $(\Delta_G + 2)$ to just 2.

D.4 Proof of Lemma 8

The proof proceeds similarly to that of Lemma 7. Let

$$\mathbf{y}_1 \triangleq \arg\max_{\mathbf{u}\in\mathcal{Y}^n} D_{\text{H}}(\mathbf{y},\mathbf{u}) + h(\mathbf{x},\mathbf{u}); \qquad \mathbf{y}_2 \triangleq \arg\max_{\mathbf{u}\in\mathcal{Y}^n} h(\mathbf{x},\mathbf{u});$$

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

Using this notation, we have that

$$\begin{aligned} & n\left|L_{\text{r}}(h,\mathbf{z}) - L_{\text{r}}(h',\mathbf{z})\right| \\ & \quad \leq \left|(D_{\text{H}}(\mathbf{y},\mathbf{y}_1) + h(\mathbf{x},\mathbf{y}_1)) - (D_{\text{H}}(\mathbf{y},\mathbf{y}_1') + h'(\mathbf{x},\mathbf{y}_1'))\right| + \left|h(\mathbf{x},\mathbf{y}_2) - h'(\mathbf{x},\mathbf{y}_2')\right|, \end{aligned} \tag{D.7}$$

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

$$\begin{aligned} \left|h(\mathbf{x},\mathbf{y}_2) - h'(\mathbf{x},\mathbf{y}_2')\right| &= h(\mathbf{x},\mathbf{y}_2) - h'(\mathbf{x},\mathbf{y}_2') \\ &\leq h(\mathbf{x},\mathbf{y}_2) - h'(\mathbf{x},\mathbf{y}_2) \\ &= (\mathbf{w}-\mathbf{w}')\cdot\mathbf{f}(\mathbf{x},\mathbf{y}_2) \\ &\leq \left\|\mathbf{w}-\mathbf{w}'\right\|_q \left\|\mathbf{f}(\mathbf{x},\mathbf{y}_2)\right\|_p \\ &\leq \left\|\mathbf{w}-\mathbf{w}'\right\|_q |G|\, R, \end{aligned} \tag{D.8}$$

via Lemma 14. Further, using the loss-augmented potentials, and assuming $\tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}) \cdot \hat{\mathbf{y}}_1 \geq \tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}') \cdot \hat{\mathbf{y}}'_1$, we have that

$$\begin{aligned}
|(D_{\text{H}}(\mathbf{y}, \mathbf{y}_1) + h(\mathbf{x}, \mathbf{y}_1)) - (D_{\text{H}}(\mathbf{y}, \mathbf{y}'_1) + h'(\mathbf{x}, \mathbf{y}'_1))| &= \tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}) \cdot \hat{\mathbf{y}}_1 - \tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}') \cdot \hat{\mathbf{y}}'_1 \\
&\leq \tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}) \cdot \hat{\mathbf{y}}_1 - \tilde{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}; \mathbf{w}') \cdot \hat{\mathbf{y}}_1 \\
&= (\boldsymbol{\theta}(\mathbf{x}; \mathbf{w}) + \delta(\mathbf{y}) - \boldsymbol{\theta}(\mathbf{x}; \mathbf{w}') - \delta(\mathbf{y})) \cdot \hat{\mathbf{y}}_1 \\
&= (\mathbf{w} - \mathbf{w}') \cdot \mathbf{f}(\mathbf{x}, \mathbf{y}_1) \\
&\leq \|\mathbf{w} - \mathbf{w}'\|_q \, \|\mathbf{f}(\mathbf{x}, \mathbf{y}_1)\|_p \\
&\leq \|\mathbf{w} - \mathbf{w}'\|_q \, |G| \, R. \qquad \text{(D.9)}
\end{aligned}$$

Combining the inequalities and dividing by n completes the proof.

D.5 Proof of Example 2

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

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

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

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

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

By assumption, every allowable hypothesis has a weight vector $\mathbf{w}$ with $\left\|\mathbf{w}\right\|_2 \leq 1$. We also assume that $\sup_{x \in \mathcal{X}} \left\|x\right\|_2 \leq 1$. Therefore, with $R = 1$ and $\beta \triangleq (2\Delta_G + 4) + 1$, Lemma 7 immediately proves that $L_{\text{r}} \circ \{h \in \mathcal{H}_{\text{M3N}} : \left\|\mathbf{w}\right\|_2 \leq 1\}$ is (β/n)-uniformly stable. Invoking Corollary 2, we then have that, with probability at least $1 - \delta$, every $\mathbb{Q}_h : \left\|\mathbf{w}\right\|_2 \leq 1$ satisfies

$$\overline{L}_{\text{r}}(\mathbb{Q}_h) \leq \hat{L}_{\text{r}}(\mathbb{Q}_h, \hat{\mathbf{Z}}) + 2\left((2\Delta_G + 4) + 1\right) \left\|\mathbf{\Gamma}^{\boldsymbol{\pi}}\right\|_\infty \sqrt{\frac{d \ln(2m \left|G\right|) + \ln \frac{2}{\delta}}{2mn}}. \qquad \text{(D.10)}$$

By construction, every $h' \sim \mathbb{Q}_h$ satisfies $\left\|\mathbf{w}' - \mathbf{w}\right\|_2 \leq (m \left|G\right|)^{-1}$, so $\mathbb{Q}$ has $(1/(m \left|G\right|), 0)$-local hypothesis stability. As demonstrated in Equation 6.36, L_{r} has $(2 \left|G\right| / n, \emptyset)$-

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

local hypothesis stability. Thus, via Proposition 5, $(L_{\mathrm{r}}, \mathbb{Q})$ has $(2/(mn), \emptyset, 0)$-local stability. Then, via Proposition 4 and Equation 6.31, we have that

$$\overline{L}_{\mathrm{H}}(h) \leq \overline{L}_{\mathrm{r}}(h) \leq \overline{L}_{\mathrm{r}}(\mathbb{Q}_h) + \frac{2}{mn}, \tag{D.11}$$

and

$$\hat{L}_{\mathrm{r}}(\mathbb{Q}_h, \hat{\mathbf{Z}}) \leq \hat{L}_{\mathrm{r}}(h, \hat{\mathbf{Z}}) + \frac{2}{mn} \leq \hat{L}_{\mathrm{h}}(h, \hat{\mathbf{Z}}) + \frac{2}{mn}. \tag{D.12}$$

Combining Equations D.10 to D.12 completes the proof.

D.6 Proof of Lemma 9

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

Fact 3. *If X is a Gaussian random variable, with mean μ and variance σ^2, then, for any $\epsilon > 0$,*

$$\Pr\left\{|X - \mu| \geq \epsilon\right\} \leq 2\exp\left(-\frac{\epsilon^2}{2\sigma^2}\right). \tag{D.13}$$

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

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

We therefore have that

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

The second inequality uses the union bound; the last uses Fact 3. Summing over $i = 1, \ldots, d$ completes the proof.

D.7 Proof of Example 5

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

Observe that, for any x and μ_y,

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

So, if $\|x\|_2 \geq 2$, then $\|x - \mu_y\|_2 \geq 1$. Therefore, using the union bound, and Lemma 9,

we can upper-bound the measure of $\mathcal{B}_{\mathcal{Z}}$ as follows:

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

Conditioned on $\overline{\mathcal{B}}$, we have that Lemmas 13 and 14 hold for $R = 2$; hence, so do Lemmas 7 and 8. With $\mathbb{P}$, $\mathbb{Q}_h$ and $\mathcal{B}_{\mathcal{H}_{\text{M3N}}}(h)$ constructed identically to Example 3, this means that $\mathbb{Q}_h$ is $(\beta_h/n, \mathcal{B}_{\mathcal{Z}}, 1/(mn))$-locally stable. Further, L_{r} has $(4\,|G|\,/n, \mathcal{B}_{\mathcal{Z}})$-local hypothesis stability, and $\mathbb{Q}$ has $(1/(m\,|G|), 1/(mn))$-local hypothesis stability; by Proposition 5, this means that $(L_{\text{r}}, \mathbb{Q})$ has $(4/(mn), \mathcal{B}_{\mathcal{Z}}, 1/(mn))$-local stability. Thus, invoking Theorem 3 and Proposition 4, with $\nu = 1/n$, we have that, with probability at least $1 - \delta - m/n$, all $l \in [m]$ satisfy $\mathbf{Z}^{(l)} \notin \mathcal{B}_{\mathcal{Z}}$, and all $h \in \mathcal{H}_{\text{M3N}}$ satisfy

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

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

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

Combining these inequalities completes the proof.

Appendix E: Proofs from Chapter 7

E.1 Properties of Strong Convexity

Strong convexity can be characterized in a number of ways. The following facts provide some conditions that are equivalent to Definition 1.

Fact 4. *A differentiable function, $\varphi : \mathcal{S} \to \mathbb{R}$, of a convex set, $\mathcal{S}$, is κ-strongly convex w.r.t. a norm, $\|\cdot\|$, if and only if, for all $s, s' \in \mathcal{S}$,*

$$\kappa \|s - s'\|^2 \leq \langle \nabla\varphi(s) - \nabla\varphi(s'),\, s - s' \rangle .$$

Fact 5. *A twice-differentiable function, $\varphi : \mathcal{S} \to \mathbb{R}$, of a convex set, $\mathcal{S}$, is κ-strongly convex w.r.t. a norm, $\|\cdot\|$, if and only if, for all $s, s' \in \mathcal{S}$,*

$$\kappa \|s\|^2 \leq \left\langle s, \nabla^2\varphi(s')\, s \right\rangle .$$

For the 2-norm, Fact 5 means that the minimum eigenvalue of the Hessian is lower-bounded by κ.

E.2 Proofs from Section 7.1

This section contains all deferred proofs from Section 7.1.

E.2.1 Proof of Stability Lemma (Lemma 11)

Recall that $\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}')$ are the gradients of $\tilde{\Phi}(\boldsymbol{\theta})$ and $\tilde{\Phi}(\boldsymbol{\theta}')$, respectively. Since the conjugate function, $\tilde{\Phi}^*$, is assumed to be κ-strongly convex, we have via Lemma 10 and Definition 18 that

$$\|\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}) - \tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}')\|_2 = \left\|\nabla\tilde{\Phi}(\boldsymbol{\theta}) - \nabla\tilde{\Phi}(\boldsymbol{\theta}')\right\|_2 \leq \frac{1}{\kappa}\,\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2\,. \tag{E.1}$$

Dividing both sides by $\sqrt{|G|}$ completes the proof.

E.2.2 The Expected NLL Minimizer Produces the True Marginals

Observe that $\hat{\boldsymbol{\theta}}_m$ effectively fits the empirical marginals of the dataset, $\frac{1}{m}\sum_{j=1}^{m}\hat{\mathbf{y}}^{(j)}$. Thus, as $m \to \infty$, the marginals induced by $\hat{\boldsymbol{\theta}}_m$ and $\boldsymbol{\theta}^\star$ converge. This is formalized in the following lemma.

Lemma 18. *Let $\boldsymbol{\mu}(\boldsymbol{\theta}^\star)$ denote the true marginals of a distribution. Let $\bar{\boldsymbol{\theta}}$ denote the minimizer of the expected NLL, per Equation 7.4. Then,*

$$\boldsymbol{\mu}(\boldsymbol{\theta}^\star) = \tilde{\boldsymbol{\mu}}(\bar{\boldsymbol{\theta}}).$$

Proof Expanding the expected NLL, we have

$$\mathbb{E}\left[-\ln \tilde{p}(\mathbf{Y};\boldsymbol{\theta})\right] = \tilde{\Phi}(\boldsymbol{\theta}) - \mathbb{E}[\boldsymbol{\theta}\cdot\hat{\mathbf{Y}}].$$

The gradient of this is

$$\nabla\,\mathbb{E}\left[-\ln \tilde{p}(\mathbf{Y};\boldsymbol{\theta})\right] = \tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}) - \mathbb{E}[\hat{\mathbf{Y}}] = \tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}) - \boldsymbol{\mu}(\boldsymbol{\theta}^\star).$$

Since the NLL is differentiable, the gradient at the minimum is zero. Thus, when $\nabla\,\mathbb{E}\left[-\ln \tilde{p}(\mathbf{Y};\bar{\boldsymbol{\theta}})\right] = 0$, we have $\tilde{\boldsymbol{\mu}}(\bar{\boldsymbol{\theta}}) = \boldsymbol{\mu}(\boldsymbol{\theta}^\star)$. ■

E.2.3 Proof of Marginals Error Bound (Proposition 6)

By Lemma 18, $\boldsymbol{\mu}(\boldsymbol{\theta}^\star) = \tilde{\boldsymbol{\mu}}(\bar{\boldsymbol{\theta}})$. Further, because $\tilde{\Phi}^*$ is assumed to be κ-strongly convex, using Lemma 11, we have that

$$\begin{aligned}\frac{1}{\sqrt{|G|}}\left\|\tilde{\boldsymbol{\mu}}(\hat{\boldsymbol{\theta}}_m) - \boldsymbol{\mu}(\boldsymbol{\theta}^\star)\right\|_2 &= \frac{1}{\sqrt{|G|}}\left\|\tilde{\boldsymbol{\mu}}(\hat{\boldsymbol{\theta}}_m) - \tilde{\boldsymbol{\mu}}(\bar{\boldsymbol{\theta}})\right\|_2 \\ &\leq \frac{1}{\kappa\sqrt{|G|}}\left\|\hat{\boldsymbol{\theta}}_m - \bar{\boldsymbol{\theta}}\right\|_2. \end{aligned} \tag{E.2}$$

The rest of the proof involves upper-bounding $\left\|\hat{\boldsymbol{\theta}}_m - \bar{\boldsymbol{\theta}}\right\|_2$.

Assumption 1 states that, with probability at least $1-\delta$, there exists a convex set, $\mathcal{S}$, encompassing $\bar{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_m$, such that the minimum eigenvalue of $\nabla^2\mathcal{L}(\,\cdot\,;\boldsymbol{\theta}) : \boldsymbol{\theta}\in\mathcal{S}$ is lower-bounded by $\gamma(\delta,m,G)$. By Fact 5, this event implies that the NLL is $\gamma(\delta,m,G)$-

strongly convex in $\mathcal{S}$. Since $\nabla^2\mathcal{L}(\,\cdot\,;\boldsymbol{\theta}) = \nabla^2\mathcal{L}_m(\boldsymbol{\theta})$, the same can be said for $\mathcal{L}_m$, so the regularized NLL,

$$\mathcal{L}_m^{\mathrm{R}}(\boldsymbol{\theta}) \triangleq \mathcal{L}_m(\boldsymbol{\theta}) + \Lambda_m \left\|\boldsymbol{\theta}\right\|_2^2,$$

is also $\gamma(\delta, m, G)$-strongly convex in $\mathcal{S}$. Therefore, with probability at least $1-\delta$ over draws of m examples,

$$\begin{aligned}
\left\|\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_m\right\|_2^2 &\leq \frac{1}{\gamma(\delta, m, G)} \left\langle \nabla\mathcal{L}_m^{\mathrm{R}}(\bar{\boldsymbol{\theta}}) - \nabla\mathcal{L}_m^{\mathrm{R}}(\hat{\boldsymbol{\theta}}_m),\, \bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_m \right\rangle \\
&= \frac{1}{\gamma(\delta, m, G)} \left\langle \nabla\mathcal{L}_m^{\mathrm{R}}(\bar{\boldsymbol{\theta}}),\, \bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_m \right\rangle \\
&\leq \frac{1}{\gamma(\delta, m, G)} \left\|\nabla\mathcal{L}_m^{\mathrm{R}}(\bar{\boldsymbol{\theta}})\right\|_2 \left\|\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_m\right\|_2.
\end{aligned}$$

The second line follows from the fact that $\hat{\boldsymbol{\theta}}_m$ is the minimizer of $\mathcal{L}_m^{\mathrm{R}}$, which is differentiable, so $\nabla\mathcal{L}_m^{\mathrm{R}}(\hat{\boldsymbol{\theta}}_m) = \mathbf{0}$. The last line uses Cauchy-Schwarz. Dividing both sides by $\left\|\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_m\right\|_2$, and combining with Equation E.2, we have that, with probability at least $1-\delta$,

$$\frac{1}{\sqrt{|G|}} \left\|\tilde{\boldsymbol{\mu}}(\hat{\boldsymbol{\theta}}_m) - \boldsymbol{\mu}(\boldsymbol{\theta}^\star)\right\|_2 \leq \frac{\left\|\nabla\mathcal{L}_m^{\mathrm{R}}(\bar{\boldsymbol{\theta}})\right\|_2}{\kappa\,\gamma(\delta, m, G)\sqrt{|G|}}. \tag{E.3}$$

Using the triangle inequality, the norm of the gradient decomposes as

$$\begin{aligned}
\left\|\nabla\mathcal{L}_m^{\mathrm{R}}(\bar{\boldsymbol{\theta}})\right\|_2 &= \left\|\nabla\mathcal{L}_m(\bar{\boldsymbol{\theta}}) + 2\Lambda_m\bar{\boldsymbol{\theta}}\right\|_2 \\
&\leq \left\|\nabla\mathcal{L}_m(\bar{\boldsymbol{\theta}})\right\|_2 + 2\Lambda_m \left\|\bar{\boldsymbol{\theta}}\right\|_2.
\end{aligned} \tag{E.4}$$

Let $N = \left|\bar{\boldsymbol{\theta}}\right|$, and note that $N = |\mathcal{Y}|\,|\mathcal{V}| + |\mathcal{Y}|^2\,|\mathcal{E}| \leq |\mathcal{Y}|^2\,|G|$. Therefore, using the

definition of Λ_m, and leveraging the assumption that $\left\|\bar{\boldsymbol{\theta}}\right\|_\infty \leq 1$, we have that

$$2\Lambda_m \left\|\bar{\boldsymbol{\theta}}\right\|_2 \leq 2\sqrt{\frac{N}{m}} \left\|\bar{\boldsymbol{\theta}}\right\|_\infty \leq 2\left|\mathcal{Y}\right| \sqrt{\frac{|G|}{m}}. \tag{E.5}$$

Turning now to the gradient of $\mathcal{L}_m$, we can expand Equation 7.3 as

$$\mathcal{L}_m(\boldsymbol{\theta}) = \frac{1}{m}\sum_{j=1}^{m} \tilde{\Phi}(\boldsymbol{\theta}) - \boldsymbol{\theta} \cdot \hat{\mathbf{y}}^{(j)}.$$

Since $\tilde{\boldsymbol{\mu}}(\bar{\boldsymbol{\theta}})$ is the gradient of $\tilde{\Phi}(\boldsymbol{\theta})$, and is in fact equal to the true marginals, $\boldsymbol{\mu}(\boldsymbol{\theta}^\star)$, we have that the gradient of $\mathcal{L}_m$ is

$$\nabla\mathcal{L}_m(\bar{\boldsymbol{\theta}}) = \frac{1}{m}\sum_{j=1}^{m} \tilde{\boldsymbol{\mu}}(\bar{\boldsymbol{\theta}}) - \hat{\mathbf{y}}^{(j)} = \boldsymbol{\mu}(\boldsymbol{\theta}^\star) - \frac{1}{m}\sum_{j=1}^{m} \hat{\mathbf{y}}^{(j)}.$$

Note that the gradient is a zero-mean random vector; random because it depends on the draw of the training set. We will bound this quantity with high probability, using a technique borrowed from London et al. (2014).

It helps to denote the gradient by a vector, $\nabla\mathcal{L}_m(\bar{\boldsymbol{\theta}}) \triangleq \mathbf{g} \in \mathbb{R}^N$. Fix some value $\epsilon > 0$. For $\mathbf{g}$ to be greater than ϵ, at least one of its coordinates must have magnitude at least $\epsilon/\sqrt{N}$; otherwise, we would have

$$\|\mathbf{g}\|_2 = \sqrt{\sum_{i=1}^{N} |g_i|^2} < \sqrt{\sum_{i=1}^{N} \frac{\epsilon^2}{N}} = \epsilon.$$

Thus, using the union bound, we have that

$$\begin{aligned}\Pr\left\{\|\mathbf{g}\|_2 \geq \epsilon\right\} &\leq \Pr\left\{\exists\, i:\ |g_i| \geq \frac{\epsilon}{\sqrt{N}}\right\} \\ &\leq \sum_{i=1}^{N} \Pr\left\{|g_i| \geq \frac{\epsilon}{\sqrt{N}}\right\}.\end{aligned}$$

Each g_i is the difference of the mean and sample average of a sufficient statistic for some node variable Y_v (or edge variable Y_e) having label y_v (or y_e). The sufficient statistics are bounded in the interval $[0, 1]$, so $|g_i| \leq 1$. Moreover, the sample average is taken from m i.i.d. draws from the target distribution. Therefore, applying Hoeffding's inequality to each i, we have that

$$\Pr\left\{|g_i| \geq \frac{\epsilon}{\sqrt{N}}\right\} \leq 2\exp\left(\frac{-2m\epsilon^2}{N}\right).$$

Summing over $i = 1, \ldots, N$, we have

$$\Pr\left\{\|\mathbf{g}\|_2 \geq \epsilon\right\} \leq 2N\exp\left(\frac{-2m\epsilon^2}{N}\right).$$

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

$$\left\|\nabla\mathcal{L}_m(\bar{\boldsymbol{\theta}})\right\|_2 \leq \sqrt{\frac{N\ln\frac{2N}{\delta}}{2m}} \leq |\mathcal{Y}|\sqrt{\frac{|G|\ln\frac{2|\mathcal{Y}|^2|G|}{\delta}}{2m}}. \tag{E.6}$$

The last inequality uses the fact that $N \leq |\mathcal{Y}|^2 |G|$.

Substituting Equations E.5 and E.6 into Equation E.4, and rearranging the terms,

we have that with probability at least $1 - \delta$,

$$\left\|\nabla \mathcal{L}_m^{\mathrm{R}}(\bar{\boldsymbol{\theta}})\right\|_2 \leq |\mathcal{Y}| \sqrt{\frac{|G|}{m}} \left(\sqrt{\frac{1}{2} \ln \frac{2 |\mathcal{Y}|^2 |G|}{\delta}} + 2 \right).$$

Then, combining the above with Equation E.3, we have that with probability at least $1-2\delta$ over draws of the training set,

$$\frac{1}{\sqrt{|G|}} \left\| \tilde{\boldsymbol{\mu}}(\hat{\boldsymbol{\theta}}_m) - \boldsymbol{\mu}(\boldsymbol{\theta}^\star) \right\|_2 \leq \frac{|\mathcal{Y}|}{\kappa \, \gamma(\delta, m, G) \sqrt{m}} \left(\sqrt{\frac{1}{2} \ln \frac{2 |\mathcal{Y}|^2 |G|}{\delta}} + 2 \right),$$

which completes the proof.

E.3 Tree-Structured Models

In this section, I analyze tree-structured models. I show that the negative entropy of a tree-structured model is strongly convex, with a modulus that depends on the contraction coefficients induced by the model. This result is used in the proof of Proposition 7. I also show how the contraction coefficients of a tree-structured model can be measured efficiently.

E.3.1 Strong Convexity of the Tree Negative Entropy

When the model is structured according to a tree, T, the marginal polytope, $\mathcal{M}$, is exactly equivalent to the local marginal polytope, $\tilde{\mathcal{M}}$. Further, its entropy function, H_T, can be expressed succinctly as a function of the marginals, using the Bethe entropy formula (see Equation 7.7). Wainwright (2006) showed that $-H_T$ is $\Omega(1/\,|G|)$-strongly convex. This is

a pessimistic lower bound, since it considers all models in the exponential family. Indeed, one can show that tree-structured models with good contraction (see Definition 19) and bounded degree induce a negative entropy that is $\Omega(1)$-strongly convex.

Proposition 9. *Fix a tree, T, with maximum degree $\Delta_T = \mathrm{O}(1)$, independent of $|\mathcal{V}|$. Let $\Theta \subseteq \mathbb{R}^{|\boldsymbol{\theta}|}$ denote the set of potentials with maximum contraction coefficient $\vartheta^\star_{\boldsymbol{\theta}} \leq 1/\Delta_T$, and let $\mathcal{M}(\Theta) \triangleq \{\boldsymbol{\mu}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ denote the corresponding set of realizable marginals. Then, the negative entropy, $-H_T$, is $\Omega(1)$-strongly convex in $\mathcal{M}(\Theta)$.*

Proof The Hessian of the log-partition, $\Phi(\boldsymbol{\theta})$, is the *covariance matrix*,

$$\Sigma(\mathbf{Y}; \boldsymbol{\theta}) \triangleq \mathbb{E}\left[\hat{\mathbf{y}}\hat{\mathbf{y}}^\top; \boldsymbol{\theta}\right] - \mathbb{E}\left[\hat{\mathbf{y}}; \boldsymbol{\theta}\right] \mathbb{E}\left[\hat{\mathbf{y}}^\top; \boldsymbol{\theta}\right],$$

where $\mathbb{E}[\,\cdot\,; \boldsymbol{\theta}]$ denotes an expectation over the distribution parameterized by $\boldsymbol{\theta}$. (For a derivation of this fact, see Wainwright and Jordan (2008).) Let $\Sigma^{-1}(\mathbf{Y}; \boldsymbol{\theta})$ denote the inverse covariance (i.e., *precision*) matrix. Since Φ is the convex conjugate of the negative entropy, $-H$, the Hessian of one is the inverse Hessian of the other. This insight yields the following lemma.

Lemma 19. *The negative entropy, $-H$, is $(1/\lambda_{max})$-strongly convex in $\mathcal{M}(\Theta)$, where $\lambda_{max} \triangleq \max_{\boldsymbol{\theta} \in \Theta} \|\Sigma(\mathbf{Y}; \boldsymbol{\theta})\|_2$ is the maximum eigenvalue of the covariance matrix, over all potentials in Θ.*

Proof Via Fact 5, $-H$ is κ-strongly convex in $\mathcal{M}(\Theta)$ if the eigenvalues of $\nabla^2\left(-H(\boldsymbol{\mu}(\boldsymbol{\theta}))\right)$, for every $\boldsymbol{\mu}(\boldsymbol{\theta}) \in \mathcal{M}(\Theta)$ (i.e., every $\boldsymbol{\theta} \in \Theta$), are bounded away from zero by κ. Via con-

vex conjugacy,

$$\nabla^2 \left(-H(\boldsymbol{\mu}(\boldsymbol{\theta}))\right) = \left(\nabla^2 \Phi(\boldsymbol{\theta})\right)^{-1} = \Sigma^{-1}(\mathbf{Y}; \boldsymbol{\theta}).$$

Therefore, the minimum eigenvalue of $-H(\boldsymbol{\mu}(\boldsymbol{\theta}))$ is equal to the maximum eigenvalue of $\Sigma(\mathbf{Y}; \boldsymbol{\theta})$. ■

Thus, to lower-bound the convexity of $-H$, it suffices to uniformly upper-bound the spectral norm of $\Sigma(\mathbf{Y}; \boldsymbol{\theta})$, over all $\boldsymbol{\theta} \in \Theta$. A simple way to do this (used by Wainwright, 2006) is to analyze the trace norm (i.e., sum of the diagonal), which upper-bounds the spectral norm. The diagonal elements of the covariance matrix are uniformly upper-bounded by $1/4$, since the sufficient statistics are bounded in $[0, 1]$. This yields an upper bound of $\mathrm{O}(|G|)$. This bound is too loose, since it grows with the size of the graph.

A better approach is to analyze the induced 1-norm (i.e., maximum column sum) or ∞-norm (i.e., maximum row sum), which, for symmetric matrices, are equivalent, and conveniently upper-bound the spectral norm. (This is because $\|\mathbf{A}\|_2 \leq \sqrt{\|\mathbf{A}\|_1 \|\mathbf{A}\|_\infty} = \sqrt{\|\mathbf{A}\|_1 \|\mathbf{A}\|_1} = \|\mathbf{A}\|_1$.) Intuitively, the 1-norm of the covariance matrix captures the maximum dependence as a function of graph distance. To bound the 1-norm, we will relate each covariance coefficient to a product of contraction coefficients. For contraction less than 1—i.e., without determinism—this product will decrease geometrically with graph distance. This geometric series converges, provided the structure has bounded degree and sufficiently small contraction.

The proof uses a technical lemma that is often credited to Dobrushin. I use a version of this given by Kontorovich (2012).

Lemma 20 (Kontorovich, 2012, Lemma 2.1)**.** *Let* $\nu : \Omega \to \mathbb{R}$ *be a signed,* balanced *measure, such that* $\sum_{\omega \in \Omega} \nu(\omega) = 0$*. Let* $K : \Omega \times \Omega \to \mathbb{R}$ *be a* Markov kernel*, where* $K(\omega \,|\, \omega') \geq 0$*,* $\sum_{\omega} K(\omega \,|\, \omega') = 1$*, and*

$$(K\nu)(\omega) \triangleq \sum_{\omega' \in \Omega} K(\omega \,|\, \omega')\, \nu(\omega').$$

Then

$$\|K\nu\|_{\text{TV}} = \sum_{\omega} \left| \sum_{\omega'} K(\omega \,|\, \omega')\, \nu(\omega') \right| \leq \vartheta \sum_{\omega'} |\nu(\omega')| = \vartheta \, \|\nu\|_{\text{TV}} \,,$$

where

$$\vartheta \triangleq \sup_{\omega, \omega' \in \Omega} \|K(\cdot \,|\, \omega) - K(\cdot \,|\, \omega')\|_{\text{TV}} \,.$$

is the contraction coefficient of K*.*

Fix any $\boldsymbol{\theta} \in \Theta$. For the following, I use the shorthand $p_{\boldsymbol{\theta}}(y)$ to denote $p(Y = y; \boldsymbol{\theta})$, and similar probabilities. I also let $\sigma_{\boldsymbol{\theta}}(y_u, y_v)$ denote the entry of the covariance matrix corresponding to $Y_u = y_u$ and $Y_v = y_v$.

Let $\pi(1), \ldots, \pi(l)$ denote the sequence of nodes along a path. Note that π is the unique path connecting its end points, since the model is tree-structured. The covariance

entries corresponding to $Y_{\pi(1)} = y_{\pi(1)}$ and $Y_{\pi(l)} = y_{\pi(l)}$ can be written recursively as

$$\begin{aligned}
\sigma_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(l)}) &= p_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(l)}) - p_{\boldsymbol{\theta}}(y_{\pi(1)}) p_{\boldsymbol{\theta}}(y_{\pi(l)}) \\
&= \sum_{y_{\pi(l-1)}} p_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(l-1)}, y_{\pi(l)}) - p_{\boldsymbol{\theta}}(y_{\pi(1)}) p_{\boldsymbol{\theta}}(y_{\pi(l-1)}, y_{\pi(l)}) \\
&= \sum_{y_{\pi(l-1)}} p_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(l-1)}) p_{\boldsymbol{\theta}}(y_{\pi(l)} \,|\, y_{\pi(l-1)}) \\
&\qquad - p_{\boldsymbol{\theta}}(y_{\pi(1)}) p_{\boldsymbol{\theta}}(y_{\pi(l-1)}) p_{\boldsymbol{\theta}}(y_{\pi(l)} \,|\, y_{\pi(l-1)}) \\
&= \sum_{y_{\pi(l-1)}} p_{\boldsymbol{\theta}}(y_{\pi(l)} \,|\, y_{\pi(l-1)}) \left(p_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(l-1)}) - p_{\boldsymbol{\theta}}(y_{\pi(1)}) p_{\boldsymbol{\theta}}(y_{\pi(l-1)}) \right) \\
&= \sum_{y_{\pi(l-1)}} p_{\boldsymbol{\theta}}(y_{\pi(l)} \,|\, y_{\pi(l-1)}) \, \sigma_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(l-1)}).
\end{aligned}$$

Note that the second equality follows from the Markov property; since $Y_{\pi(l)}$ is conditionally independent of $Y_{\pi(1)}$ given $Y_{\pi(l-1)}$, we have that $p_{\boldsymbol{\theta}}(y_{\pi(l)} \,|\, y_{\pi(l-1)}, y_{\pi(l)}) = p_{\boldsymbol{\theta}}(y_{\pi(l)} \,|\, y_{\pi(l-1)})$.

In the righthand expression, the conditional probability under $p_{\boldsymbol{\theta}}$ defines a Markov kernel. Moreover, the covariance with $y_{\pi(1)}$ defines a signed measure,

$$\nu(Y; y_{\pi(1)}) \triangleq \sigma_{\boldsymbol{\theta}}(y_{\pi(1)}, Y),$$

which is balanced, since

$$\begin{aligned}
\sum_y \nu(y; y_{\pi(1)}) &= \sum_y \sigma_{\boldsymbol{\theta}}(y_{\pi(1)}, y) \\
&= \sum_y p_{\boldsymbol{\theta}}(y_{\pi(1)}, y) - p_{\boldsymbol{\theta}}(y_{\pi(1)}) p_{\boldsymbol{\theta}}(y) \\
&= p_{\boldsymbol{\theta}}(y_{\pi(1)}) - p_{\boldsymbol{\theta}}(y_{\pi(1)}) = 0.
\end{aligned}$$

Therefore, via Lemma 20, we have that

$$\sum_{y_{\pi(l)}} \left|\sigma_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(l)})\right| = \sum_{y_{\pi(l)}} \left| \sum_{y_{\pi(l-1)}} p_{\boldsymbol{\theta}}(y_{\pi(l)} \mid y_{\pi(l-1)})\, \sigma_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(l-1)}) \right|$$
$$\leq \vartheta^{\star}_{\boldsymbol{\theta}} \sum_{y_{\pi(l-1)}} \left|\sigma_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(l-1)})\right|$$

Applying this identity recursively, we have that

$$\begin{aligned}
\sum_{y_{\pi(l)}} \left|\sigma_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(l)})\right| &\leq \vartheta^{\star}_{\boldsymbol{\theta}} \sum_{y_{\pi(l-1)}} \left|\sigma_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(l-1)})\right| \\
&\vdots \\
&\leq (\vartheta^{\star}_{\boldsymbol{\theta}})^{l-2} \sum_{y_{\pi(2)}} \left|\sigma_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(2)})\right| \\
&\leq (\vartheta^{\star}_{\boldsymbol{\theta}})^{l-1} \sum_{y'_{\pi(1)}} \left|\sigma_{\boldsymbol{\theta}}(y_{\pi(1)}, y'_{\pi(1)})\right| \\
&\leq \frac{|\mathcal{Y}|}{4} (\vartheta^{\star}_{\boldsymbol{\theta}})^{l-1}.
\end{aligned}$$

The last inequality follows from the fact that the covariance of any variable assignment is at most $1/4$ in magnitude, and the covariance between any two assignments to the same variable is also at most $1/4$.

Given an upper bound on the covariances of node assignments, one can bound the covariance of edge assignments. Consider edges $\{a, b\}, \{c, d\} \in \mathcal{E}$. Due to the tree structure, the edges lie at opposite ends of a unique path connecting their constituent nodes. Without loss of generality, assume that this path has the order $a, b, \ldots, c, d$, and that the length of the path from b to c is l. By the Markov property, Y_a and Y_d are condition-

ally independent given Y_b and Y_c. Thus, for any configuration $(Y_a, Y_b) = (y_a, y_b)$ and $(Y_c, Y_d) = (y_c, y_d)$, we have that

$$\begin{aligned}
\sum_{y_c, y_d} |\sigma_{\boldsymbol{\theta}}((y_a, y_b), (y_c, y_d))| &= \sum_{y_c, y_d} |p_{\boldsymbol{\theta}}(y_a, y_b, y_c, y_d) - p_{\boldsymbol{\theta}}(y_a, y_b) p_{\boldsymbol{\theta}}(y_c, y_d)| \\
&= \sum_{y_c, y_d} \Big| p_{\boldsymbol{\theta}}(y_a, y_d \,|\, y_b, y_c) p_{\boldsymbol{\theta}}(y_b, y_c) \\
&\qquad - p_{\boldsymbol{\theta}}(y_a \,|\, y_b) p_{\boldsymbol{\theta}}(y_b) p_{\boldsymbol{\theta}}(y_d \,|\, y_c) p_{\boldsymbol{\theta}}(y_c) \Big| \\
&= \sum_{y_c, y_d} \Big| p_{\boldsymbol{\theta}}(y_a \,|\, y_b) p_{\boldsymbol{\theta}}(y_d \,|\, y_c) p_{\boldsymbol{\theta}}(y_b, y_c) \\
&\qquad - p_{\boldsymbol{\theta}}(y_a \,|\, y_b) p_{\boldsymbol{\theta}}(y_d \,|\, y_c) p_{\boldsymbol{\theta}}(y_b) p_{\boldsymbol{\theta}}(y_c) \Big| \\
&= \sum_{y_c, y_d} p_{\boldsymbol{\theta}}(y_a \,|\, y_b) p_{\boldsymbol{\theta}}(y_d \,|\, y_c) \, |\sigma_{\boldsymbol{\theta}}(y_b, y_c)| \\
&= p_{\boldsymbol{\theta}}(y_a \,|\, y_b) \sum_{y_c} |\sigma_{\boldsymbol{\theta}}(y_b, y_c)| \sum_{y_d} p_{\boldsymbol{\theta}}(y_d \,|\, y_c) \\
&= p_{\boldsymbol{\theta}}(y_a \,|\, y_b) \sum_{y_c} |\sigma_{\boldsymbol{\theta}}(y_b, y_c)| \\
&\leq \frac{|\mathcal{Y}|}{4} \, (\vartheta^{\star}_{\boldsymbol{\theta}})^{l-1}.
\end{aligned}$$

The same argument can be used to bound the covariance between node and edge variables, where the relevant path length l becomes the length from the node to the closest endpoint of the edge. The base case of covariance between a node or edge state indicator and another state is also at most $1/4$.

Thus far, I have derived upper bounds on the entries of the covariance matrix, which correspond to covariances between three types of pairs: node variables and node variables; node variables and edge variables; and edge variables and edge variables. For a

distribution induced by a tree-structured model, with maximum degree Δ_T, the 1-norm of a column corresponding to a node assignment $Y_u = y_u$ is

$$\begin{aligned}
\sigma_{\boldsymbol{\theta}}(Y_u = y_u) &= \sum_{y'_u} |\sigma_{\boldsymbol{\theta}}(y_u, y'_u)| + \sum_{v \in \mathcal{V} \setminus u} \sum_{y_v} |\sigma_{\boldsymbol{\theta}}(y_u, y_v)| \\
&\quad + \sum_{\{v,v'\} \in \mathcal{E}} \sum_{y_v, y_{v'}} |\sigma_{\boldsymbol{\theta}}(y_u, (y_v, y_{v'}))| \\
&\leq \frac{|\mathcal{Y}|}{4} + \frac{|\mathcal{Y}|}{4} \sum_{v \in \mathcal{V} \setminus u} (\vartheta^\star_{\boldsymbol{\theta}})^{l(u,v)-1} \\
&\quad + \frac{|\mathcal{Y}|}{4} \sum_{\{v,v'\} \in \mathcal{E}} (\vartheta^\star_{\boldsymbol{\theta}})^{\max\{0, \min\{l(u,v), l(u,v')\}-1\}} \\
&\leq \frac{|\mathcal{Y}|}{4} + \frac{|\mathcal{Y}|}{4} \sum_{d=1}^{\infty} \Delta_T^d (\vartheta^\star_{\boldsymbol{\theta}})^{d-1} \\
&\quad + \frac{|\mathcal{Y}|\,\Delta_T}{4} + \frac{|\mathcal{Y}|}{4} \sum_{d=1}^{\infty} \Delta_T^{d+1} (\vartheta^\star_{\boldsymbol{\theta}})^{d-1} \\
&= \frac{|\mathcal{Y}|}{4} + \frac{|\mathcal{Y}|\,\Delta_T}{4} \sum_{d=1}^{\infty} (\Delta_T\, \vartheta^\star_{\boldsymbol{\theta}})^{d-1} \\
&\quad + \frac{|\mathcal{Y}|\,\Delta_T}{4} + \frac{|\mathcal{Y}|\,\Delta_T^2}{4} \sum_{d=1}^{\infty} (\Delta_T\, \vartheta^\star_{\boldsymbol{\theta}})^{d-1} \\
&= \frac{|\mathcal{Y}|}{4} + \frac{|\mathcal{Y}|\,\Delta_T}{4(1 - \Delta_T\, \vartheta^\star_{\boldsymbol{\theta}})} + \frac{|\mathcal{Y}|\,\Delta_T}{4} + \frac{|\mathcal{Y}|\,\Delta_T^2}{4(1 - \Delta_T\, \vartheta^\star_{\boldsymbol{\theta}})}.
\end{aligned}$$

where $l(u, v)$ is the length of the path from node u to v. The second inequality holds because the number of nodes at distance d is at most Δ_T^d, and the maximum number of edges with endpoints at distance d is at most Δ_T^{d+1}, where we adjust for node and edge variables at distance zero. The last line applies the geometric series identity, since $\Delta_T\, \vartheta^\star_{\boldsymbol{\theta}} < \Delta_T / \Delta_T = 1$. An analogous argument bounds the 1-norm of any column corresponding to an edge assignment.

Since the 1-norm of every column of the covariance matrix is upper-bounded independently of $|G|$, it follows that the induced 1-norm of $\Sigma(\mathbf{Y}; \boldsymbol{\theta})$ is bounded independently of $|G|$; that is, $\|\Sigma(\mathbf{Y}; \boldsymbol{\theta})\|_1 = \mathrm{O}(1)$. This holds for every $\boldsymbol{\theta} \in \Theta$, though the constant may differ, depending on $\vartheta^\star_{\boldsymbol{\theta}}$. Recall that the 1-norm of the covariance matrix upper-bounds the spectral norm, since the covariance matrix is symmetric. Thus, the minimum eigenvalue of $\nabla^2 \left(-H(\boldsymbol{\mu}(\boldsymbol{\theta}))\right)$, for every $\boldsymbol{\mu}(\boldsymbol{\theta}) \in \mathcal{M}(\Theta)$, is lower-bounded by a constant, which means that the negative entropy is $\Omega(1)$-strongly convex in $\mathcal{M}(\Theta)$. ■

E.3.2 Measuring Contraction

In the previous section, I relate the convexity of $-H_T$ to the model's maximum contraction coefficient. For general graphical models, measuring the contraction coefficients may be intractable. However, when the model is tree-structured, there is an efficient algorithm.

For a tree-structured model, exact inference can be computed efficiently using message passing. Given the node and edge marginals, one can compute the conditional probabilities via

$$p\left(Y_u = y_u \,|\, Y_v = y_v \,; \boldsymbol{\theta}\right) = \frac{p\left(Y_u = y_u, Y_v = y_v \,; \boldsymbol{\theta}\right)}{p\left(Y_v = y_v \,; \boldsymbol{\theta}\right)}.$$

One can then compute the total variation distance; hence, the contraction coefficient. For variables with small domains (e.g., binary), this is efficient. Given the contraction coefficient for each $(u, v) : \{u, v\} \in \mathcal{E}$, computing the maximum contraction coefficient is trivial.

Note that marginal inference only needs to be computed once in this procedure. The

time complexity of inference in a tree-structured model, with $|\mathcal{Y}|$ labels and $|\mathcal{E}|$ edges is $\mathrm{O}(|\mathcal{Y}|^2 |\mathcal{E}|)$. For each undirected edge, there are two contraction coefficients (one per direction), each of which involves $|\mathcal{Y}|^2$ operations ($|\mathcal{Y}|$ additions to compute the total variation distance conditioned on Y_v; and $|\mathcal{Y}|$ values of Y_v to condition on to compute the supremum). Since there are $|\mathcal{E}|$ edges, the overall time complexity of computing the contraction coefficients is $\mathrm{O}(|\mathcal{Y}|^2 |\mathcal{E}|)$.

E.4 Tree-Reweighting

In this section, we prove Proposition 7, which gives a model-dependent lower bound on the modulus of convexity for the tree-reweighted negative entropy. We also explore the ramifications of Proposition 7 for a grid-structured model.

E.4.1 Proof of $-H_{\text{TR}}$ Strong Convexity (Proposition 7)

The following lemma relates the convexity of $-H_{\text{TR}}$ to the convexity of its constituent tree entropies, as well as the tree distribution.

Lemma 21. *(Wainwright, 2006, Appendix C) Fix a graph, $G \triangleq (\mathcal{V}, \mathcal{E})$, and a distribution, ρ, over the spanning trees, $\mathcal{T}(G)$, such that $\rho(e) > 0$ for all $e \in \mathcal{E}$. Let $\rho_e^\star \triangleq \min_{e \in \mathcal{E}} \rho(e)$ denote the minimum edge probability. Let $\kappa_T^\star$ denote the minimum convexity of $-H_T$ for any tree $T \in \mathcal{T}(G)$ with positive probability under ρ. Then the tree-reweighted negative entropy, $-H_{\text{TR}}$, is $(\rho_e^\star \kappa_T^\star)$-strongly convex.*

Thus, to prove $\Omega(1)$-strong convexity, one must show that the minimum edge probability, $\rho_e^\star$, and the minimum tree convexity, $\kappa_T^\star$ are both lower-bounded by values that are

independent of $|G|$.

In Proposition 7, it is assume that $\rho_e^\star$ is lower-bounded by a positive constant, $C > 0$. Since H_{TR} can be defined using *any* distribution over spanning trees, it is usually possible to construct an edge distribution for which this holds. (An example for a grid is given in Section E.4.2.) Therefore, the real challenge is to show that $\kappa_T^\star = \Omega(1)$. For each $T \in \mathcal{T}(G)$, denote the set of admissible potentials by $\Theta_T \subseteq \mathbb{R}^{|\boldsymbol{\theta}|}$, where dimensions corresponding to edges that don't exist in T have unbounded range. Note that

$$\Theta = \bigcap_{T \in \mathcal{T}(G): \rho(T) > 0} \Theta_T,$$

so

$$\tilde{\mathcal{M}}(\Theta) = \bigcap_{T \in \mathcal{T}(G): \rho(T) > 0} \tilde{\mathcal{M}}(\Theta_T).$$

Let $\tilde{\mathcal{M}}_T(\Theta)$ denote the projection of $\tilde{\mathcal{M}}(\Theta)$ onto the subspace defined by the nodes and edges in T, and note that $\tilde{\mathcal{M}}_T(\Theta) \subseteq \tilde{\mathcal{M}}_T(\Theta_T)$. Proposition 9 showed that, under suitable structural and contraction conditions, $-H_T$ is $\Omega(1)$-strongly convex in $\tilde{\mathcal{M}}_T(\Theta_T)$; hence, in $\tilde{\mathcal{M}}_T(\Theta)$ as well. When combined with Lemma 21, with $\rho_e^\star > C$, this proves that $-H_{\text{TR}}$ is $\Omega(1)$-strongly convex in $\tilde{\mathcal{M}}(\Theta)$.

E.4.2 Example Tree-Reweighting for a Grid Graph

Suppose the model is structured according to an $m \times n$ grid. This graph can be covered using a set of 4 chains, using the "snake-like" pattern illustrated in Figure E.1. Observe that each internal edge is covered by 2 chains, and each boundary edge is covered by

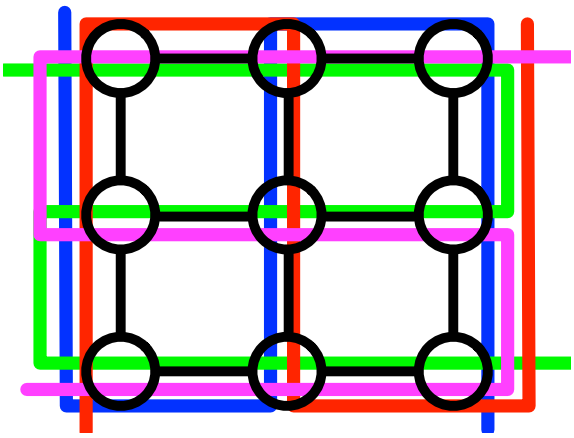


Figure E.1: Covering the edges of a grid graph with 4 chains.

3 chains. Therefore, using a uniform distribution over the chains, we have that each internal edge, e, has probability $\rho(e) = 1/2$, and each boundary edge, e', has probability $\rho(e') = 3/4$.

To apply Proposition 7 to this spanning tree distribution, take $C = 1/2$ as the minimum edge probability. The maximum degree of a chain is 2, so the maximum contraction coefficient, $\vartheta^\star_{\boldsymbol{\theta},T}$, must be at most $1/2$. It may be possible to upper-bound $\vartheta^\star_{\boldsymbol{\theta},T}$ analytically for all $\boldsymbol{\theta}$ in some space. Alternately, one could map out the space of feasible potentials by measuring $\vartheta^\star_{\boldsymbol{\theta},T}$, using the procedure from Section E.3.2.

E.5 Proof of $-H_{\mathrm{C}}$ Strong Convexity (Proposition 8)

In this section, I prove Proposition 8, which characterizes the modulus of convexity for counting number entropies. The proof of Proposition 8 requires two technical lemmas.

Lemma 22 (Shalev-Schwartz, 2007, Lemma 16)**.** *The function* $\varphi(\mathbf{z}) \triangleq \sum_i^d z_i \log z_i$ *is 1-strongly convex in the probability simplex,* $\{\mathbf{z} \in [0,1]^d : \|\mathbf{z}\|_1 = 1\}$*, w.r.t. the 1-norm.*

Lemma 23 (Heskes, 2006, Lemma A.1)**.** *The difference of entropies, equivalent to the negative conditional entropy,* $H_v(\tilde{\mu}_v) - H_e(\tilde{\mu}_e) = -H_{e|v}(\tilde{\mu}_e)$*, for* $v \in e$*, is a convex function of* $\tilde{\mu}_e$*.*

Every edge, e, is composed of exactly two nodes, $\{u, v\}$. By assumption, $\alpha_e \geq \kappa > 0$. Therefore, one can shift $(2\kappa/3)$ weight from α_e to α_u and α_v without affecting the counting numbers or Heskes's convexity conditions. Let:

$$\begin{aligned} \forall e \in \mathcal{E},\ \ \tilde{\alpha}_e &\triangleq \alpha_e - \frac{2\kappa}{3}\,; \\ \forall (v,e) : v \in e,\ \ \tilde{\alpha}_{v,e} &\triangleq \alpha_{v,e} + \frac{\kappa}{3}\,; \\ \forall v \in \mathcal{V},\ \ \tilde{\alpha}_v &\triangleq \alpha_v + \sum_{e:v\in e} \frac{\kappa}{3}. \end{aligned}$$

Observe that the new auxiliary counts satisfy Equations 7.10 and 7.11:

$$\forall v \in \mathcal{V},\ c_v = \alpha_v - \sum_{e:v\in e} \left(\alpha_{v,e} + \frac{\kappa}{3} - \frac{\kappa}{3}\right) = \tilde{\alpha}_v - \sum_{e:v\in e} \tilde{\alpha}_{v,e}\,; \tag{E.7}$$

$$\forall e \in \mathcal{E},\ c_e = \alpha_e + \sum_{v:v\in e} \left(\alpha_{v,e} + \frac{\kappa}{3} - \frac{\kappa}{3}\right) = \tilde{\alpha}_e + \sum_{v:v\in e} \tilde{\alpha}_{v,e}. \tag{E.8}$$

Now, every e has $\tilde{\alpha}_e \geq \kappa/3$. Further, because it is assumed that every node is involved in at least one edge, every v has $\tilde{\alpha}_v \geq \kappa/3$. (One could extend Proposition 8 to arbitrary graphs by assuming that every isolated node has $c_v \geq \kappa/3$.)

Substituting Equations E.7 and E.8 into Equation 7.9 and rearranging the terms, we

obtain

$$-H_{\mathrm{C}}(\tilde{\boldsymbol{\mu}}) = -\sum_{v\in\mathcal{V}} \tilde{\alpha}_v H_v(\tilde{\mu}_v) - \sum_{e\in\mathcal{E}} \tilde{\alpha}_e H_e(\tilde{\mu}_e) + \sum_{e\in\mathcal{E}} \sum_{v:v\in e} \tilde{\alpha}_{v,e}(H_v(\tilde{\mu}_v) - H_e(\tilde{\mu}_e))$$
$$= -\sum_{v\in\mathcal{V}} \tilde{\alpha}_v H_v(\tilde{\mu}_v) - \sum_{e\in\mathcal{E}} \tilde{\alpha}_e H_e(\tilde{\mu}_e) - \sum_{e\in\mathcal{E}} \sum_{v:v\in e} \tilde{\alpha}_{v,e} H_{e|v}(\tilde{\mu}_e). \tag{E.9}$$

We will analyze the entropy terms individually, using the gradient property of (strong) convexity (Fact 4).

Fix any two vectors $\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\mu}}' \in \tilde{\mathcal{M}}$, and let $\boldsymbol{\delta} \triangleq \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}'$. Recall that $\forall v,\ \|\tilde{\mu}_v\|_1 = \|\tilde{\mu}'_v\|_1 = 1$ and $\forall e,\ \|\tilde{\mu}_e\|_1 = \|\tilde{\mu}'_e\|_1 = 1$. Via Lemma 22, $-H_v$ and $-H_e$ are 1-strongly convex in the probability simplex with respect to the 1-norm. By Fact 4, this means that every node v satisfies,

$$\langle \nabla(-H_v(\tilde{\mu}_v)) - \nabla(-H_v(\tilde{\mu}'_v)),\ \delta_v \rangle \geq \|\delta_v\|_1^2 .$$

Therefore,

$$\begin{aligned}\tilde{\alpha}_v \langle \nabla(-H_v(\tilde{\mu}_v)) - \nabla(-H_v(\tilde{\mu}'_v)),\ \delta_v \rangle &\geq \tilde{\alpha}_v \|\delta_v\|_1^2 \\ &\geq \tilde{\alpha}_v \|\delta_v\|_2^2 \\ &\geq \frac{\kappa}{3} \|\delta_v\|_2^2 .\end{aligned}$$

The same holds for every edge e. Further, by Lemma 23, $H_{e|v}(\tilde{\mu}_e) = H_v(\tilde{\mu}_v) - H_e(\tilde{\mu}_e)$ is convex, meaning

$$\left\langle \nabla(-H_{e|v}(\tilde{\mu}_e)) - \nabla(-H_{e|v}(\tilde{\mu}'_e)),\ \delta_e \right\rangle \geq 0.$$

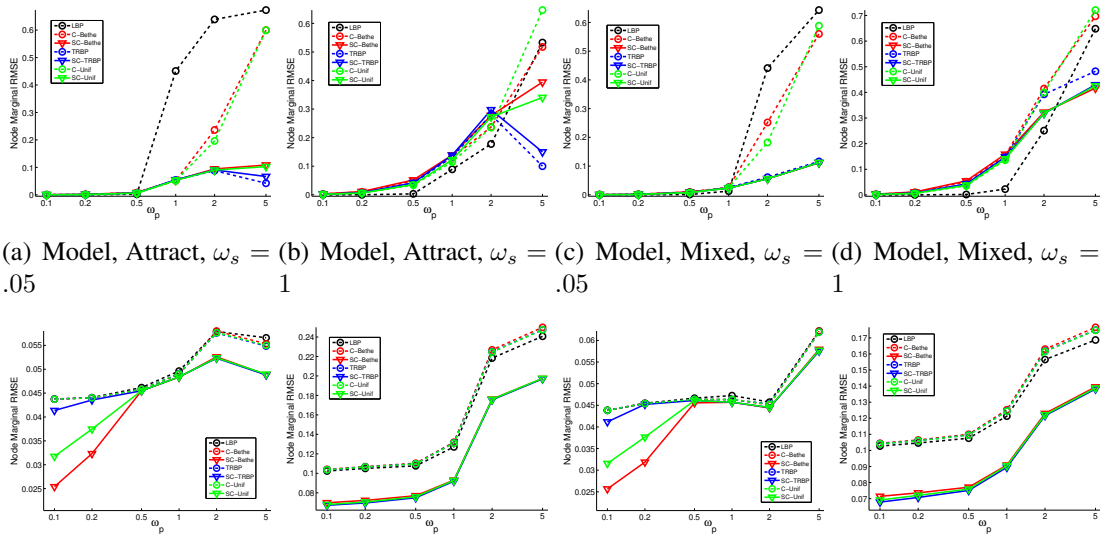
Thus, taking the gradient of Equation E.9, we have that

$$\begin{aligned}
\langle \nabla(-H_{\mathrm{C}}(\tilde{\boldsymbol{\mu}})) - \nabla(-H_{\mathrm{C}}(\tilde{\boldsymbol{\mu}}')),\, \boldsymbol{\delta} \rangle &= \sum_{v \in \mathcal{V}} \tilde{\alpha}_v \left\langle \nabla(-H_v(\tilde{\mu}_v)) - \nabla(-H_v(\tilde{\mu}'_v)),\, \delta_v \right\rangle \\
&\quad + \sum_{e \in \mathcal{E}} \tilde{\alpha}_e \left\langle \nabla(-H_e(\tilde{\mu}_e)) - \nabla(-H_e(\tilde{\mu}'_e)),\, \delta_e \right\rangle \\
&\quad + \sum_{e \in \mathcal{E}} \sum_{v : v \in e} \tilde{\alpha}_{v,e} \left\langle \nabla(-H_{e|v}(\tilde{\mu}_e)) - \nabla(-H_{e|v}(\tilde{\mu}'_e)),\, \delta_e \right\rangle \\
&\geq \frac{\kappa}{3} \sum_{v \in \mathcal{V}} \|\delta_v\|_2^2 \,+\, \frac{\kappa}{3} \sum_{e \in \mathcal{E}} \|\delta_e\|_2^2 \,+\, 0 \\
&= \frac{\kappa}{3} \|\boldsymbol{\delta}\|_2^2 \,,
\end{aligned}$$

which completes the proof, via Fact 4.

Appendix F: Figures from Chapter 7

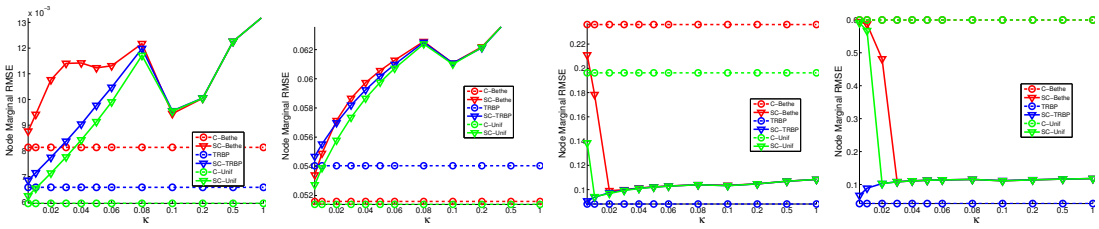
In all plots, results are averaged over 20 trials and the y-axis has been rescaled to fit the data. See Section 7.3.3 for discussion.



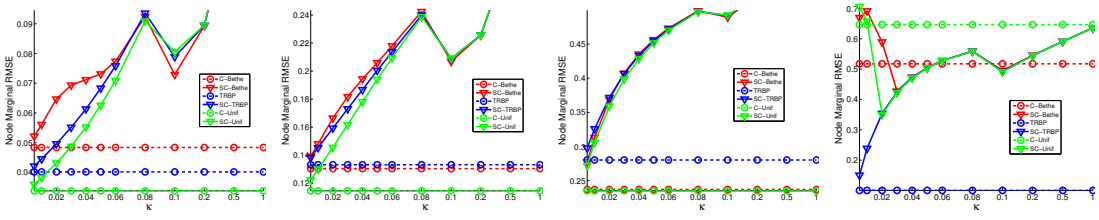
(a) Model, Attract, $\omega_s = .05$ (b) Model, Attract, $\omega_s = 1$ (c) Model, Mixed, $\omega_s = .05$ (d) Model, Mixed, $\omega_s = 1$

(e) Learned, Attract, $\omega_s = .05$ (f) Learned, Attract, $\omega_s = 1$ (g) Learned, Mixed, $\omega_s = .05$ (h) Learned, Mixed, $\omega_s = 1$

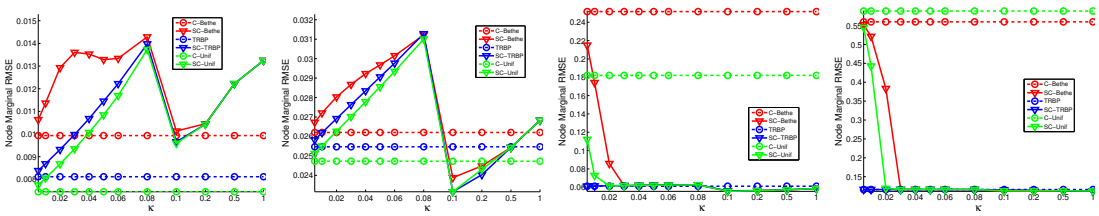
Figure F.1: Plots of RMSE of the node marginals as a function of the interaction parameter, ω_p. Inference is performed using the true model in (a)-(d), and the learned model in (e)-(h). The first two columns correspond to a model with attractive potentials; the third and fourth to a model with mixed potentials. The black dotted line is LBP; color dotted lines are the convex baselines, and solid lines are their SC counterparts. The SC methods use the post hoc optimal value of κ (and C) in the counting number optimization. For learned marginals, SC offers statistically significant error reduction—sometimes over 40%—for all data models and baselines, except C-Bethe at $\omega_p = .5$ in (g).



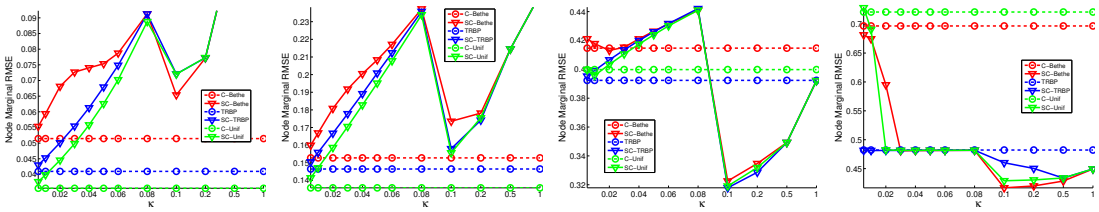
(a) Attract, $\omega_s = .05$, $\omega_p = .5$ (b) Attract, $\omega_s = .05$, $\omega_p = 1$ (c) Attract, $\omega_s = .05$, $\omega_p = 2$ (d) Attract, $\omega_s = .05$, $\omega_p = 5$



(e) Attract, $\omega_s = 1$, $\omega_p = .5$ (f) Attract, $\omega_s = 1, \omega_p = 1$ (g) Attract, $\omega_s = 1$, $\omega_p = 2$ (h) Attract, $\omega_s = 1$, $\omega_p = 5$

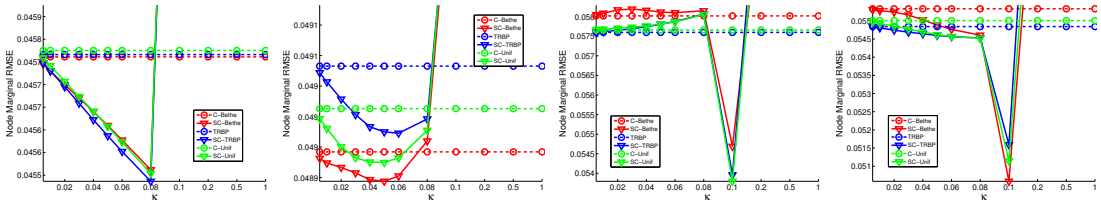


(i) Mixed, $\omega_s = .05$, $\omega_p = .5$ (j) Mixed, $\omega_s = .05$, $\omega_p = 1$ (k) Mixed, $\omega_s = .05$, $\omega_p = 2$ (l) Mixed, $\omega_s = .05$, $\omega_p = 5$

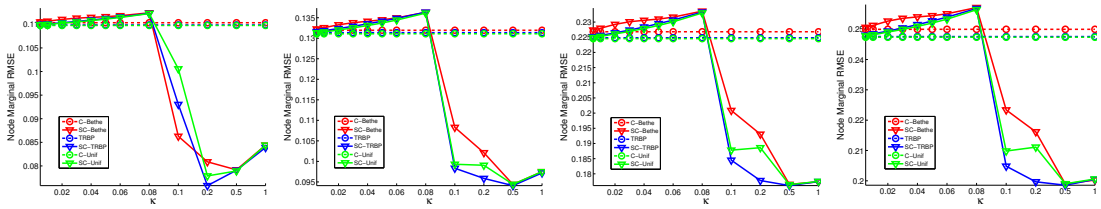


(m) Mixed, $\omega_s = 1$, $\omega_p = .5$ (n) Mixed, $\omega_s = 1, \omega_p = 1$ (o) Mixed, $\omega_s = 1, \omega_p = 2$ (p) Mixed, $\omega_s = 1, \omega_p = 5$

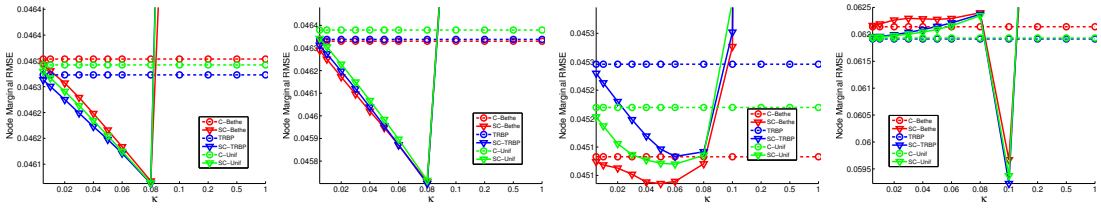
Figure F.2: Plots of RMSE of the node marginals as a function of the convexity parameter, κ, which determines the minimum modulus of convexity used in the counting number QP. For $\kappa < .1$, we use Equation 7.12; for $\kappa \geq .1$, we use Equation 7.13 and report the score for the post hoc optimal C. SC algorithms are plotted as solid lines, and their respective counterparts are overlaid as dashed lines. Inference is performed using the true model. The first two rows correspond to a model with attractive potentials; the third and fourth to a model with mixed potentials. In all plots, the x-axis scales logarithmically for $\kappa > .1$. Certain plots have been truncated vertically to better fit the data.



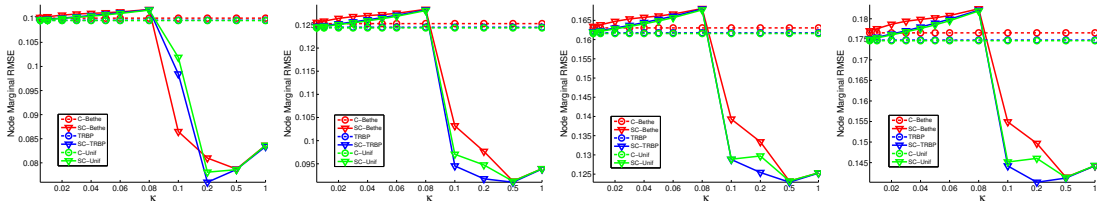
(a) Attract, $\omega_s = .05$, $\omega_p = .5$ (b) Attract, $\omega_s = .05$, $\omega_p = 1$ (c) Attract, $\omega_s = .05$, $\omega_p = 2$ (d) Attract, $\omega_s = .05$, $\omega_p = 5$



(e) Attract, $\omega_s = 1$, $\omega_p = .5$ (f) Attract, $\omega_s = 1, \omega_p = 1$ (g) Attract, $\omega_s = 1$, $\omega_p = 2$ (h) Attract, $\omega_s = 1$, $\omega_p = 5$

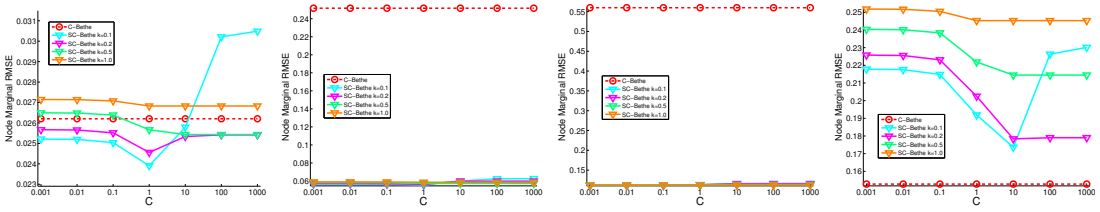


(i) Mixed, $\omega_s = .05$, $\omega_p = .5$ (j) Mixed, $\omega_s = .05$, $\omega_p = 1$ (k) Mixed, $\omega_s = .05$, $\omega_p = 2$ (l) Mixed, $\omega_s = .05$, $\omega_p = 5$

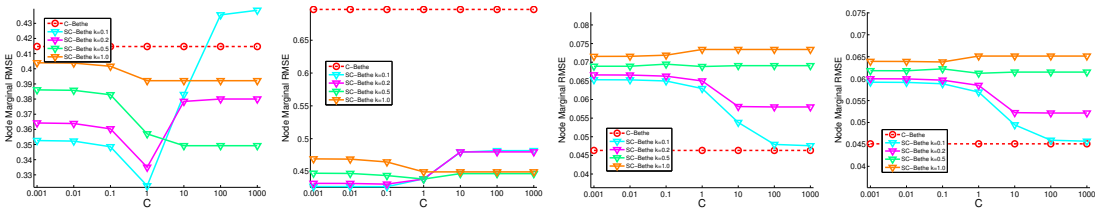


(m) Mixed, $\omega_s = 1$, $\omega_p = .5$ (n) Mixed, $\omega_s = 1, \omega_p = 1$ (o) Mixed, $\omega_s = 1, \omega_p = 2$ (p) Mixed, $\omega_s = 1, \omega_p = 5$

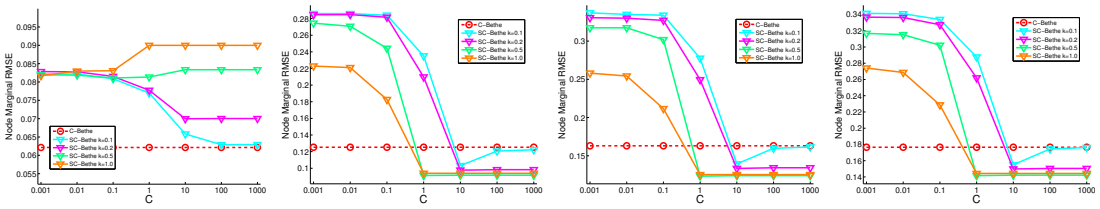
Figure F.3: Plots of RMSE of the node marginals as a function of the convexity parameter, κ, when using the learned model for inference.



(a) Model, $\omega_s = .05$, $\omega_p = 1$ (b) Model, $\omega_s = .05$, $\omega_p = 2$ (c) Model, $\omega_s = .05$, $\omega_p = 5$ (d) Model, $\omega_s = 1$, $\omega_p = 1$

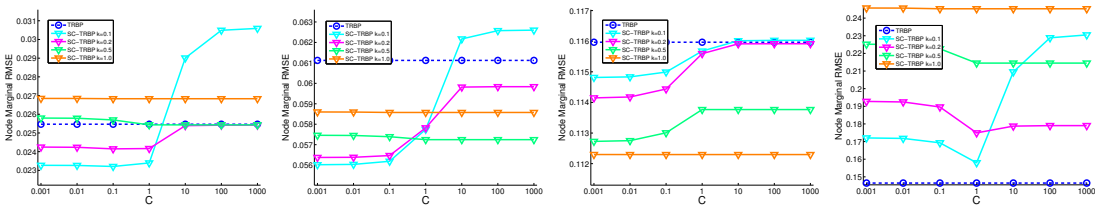


(e) Model, $\omega_s = 1$, $\omega_p = 2$ (f) Model, $\omega_s = 1$, $\omega_p = 5$ (g) Learned, $\omega_s = .05$, $\omega_p = 1$ (h) Learned, $\omega_s = .05$, $\omega_p = 2$

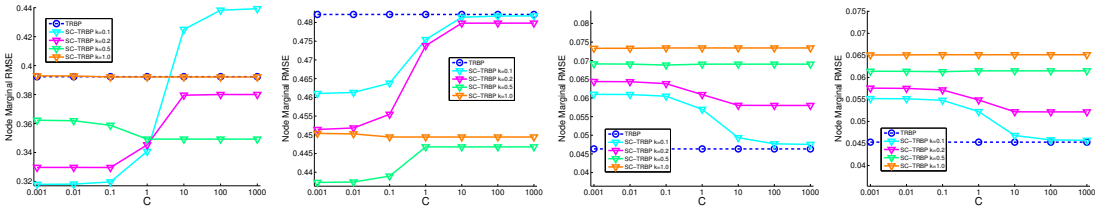


(i) Learned, $\omega_s = .05$, $\omega_p = 5$ (j) Learned, $\omega_s = 1$, $\omega_p = 1$ (k) Learned, $\omega_s = 1$, $\omega_p = 2$ (l) Learned, $\omega_s = 1$, $\omega_p = 5$

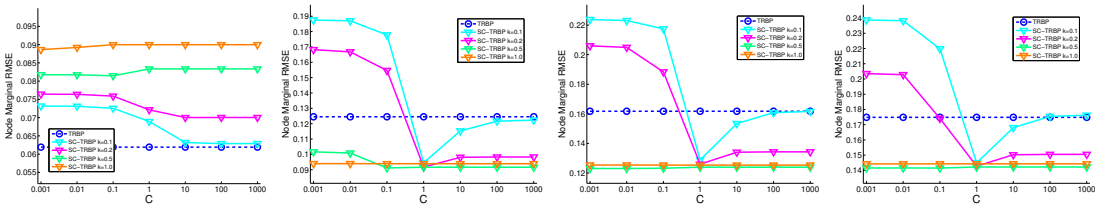
Figure F.4: Select plots of RMSE as a function of the slack parameter, C, in the slackened counting number QP (Equation 7.13), at higher values of κ. The slack parameter trades off between fitting the target counting numbers and satisfying variable validity. Data is generated using mixed potentials in all plots. These plots focus on the Bethe approximation. SC versions are solid color lines; C-Bethe is overlaid as a dashed red line.



(a) Model, $\omega_s = .05$, $\omega_p = 1$ (b) Model, $\omega_s = .05$, $\omega_p = 2$ (c) Model, $\omega_s = .05$, $\omega_p = 5$ (d) Model, $\omega_s = 1$, $\omega_p = 1$



(e) Model, $\omega_s = 1$, $\omega_p = 2$ (f) Model, $\omega_s = 1$, $\omega_p = 5$ (g) Learned, $\omega_s = .05$, $\omega_p = 1$ (h) Learned, $\omega_s = .05$, $\omega_p = 2$



(i) Learned, $\omega_s = .05$, $\omega_p = 5$ (j) Learned, $\omega_s = 1$, $\omega_p = 1$ (k) Learned, $\omega_s = 1$, $\omega_p = 2$ (l) Learned, $\omega_s = 1$, $\omega_p = 5$

Figure F.5: Select plots of RMSE as a function of the slack parameter, C, for the tree-reweighting approximation. SC versions are solid color lines; C-TRBP is overlaid as a dashed blue line.

Bibliography

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.

D. Anguelov, B. Taskar, V. Chatalbashev, D. Koller, D. Gupta, G. Heitz, and A. Ng. Discriminative learning of Markov random fields for segmentation of 3D scan data. In *Computer Vision and Pattern Recognition*, 2005.

P. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998.

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.

P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4), 2009.

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

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

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

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

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

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

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

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

C. Cortes, M. Mohri, D. Pechyony, and A. Rastogi. Stability of transductive regression algorithms. In *International Conference on Machine Learning*, 2008.

H. Daumé, III, J. Langford, and D. Marcu. Search-based structured prediction. *Machine Learning*, 75(3):297–325, 2009.

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.

A. Elisseeff, T. Evgeniou, and M. Pontil. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6:55–79, 2005.

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.

L. Getoor and B. Taskar. *Introduction to Statistical Relational Learning*. The MIT Press, 2007.

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

K. Gimpel and N. Smith. Softmax-margin CRFs: Training log-linear models with cost

functions. In *Conference of the North American Chapter of the Association of Computational Linguistics*, 2010.

A. Globerson and T. Jaakkola. Approximate inference using conditional entropy decompositions. In *Artificial Intelligence and Statistics*, pages 130–138, 2007.

Arnulf B. A. Graf, Alexander J. Smola, and Silvio Borer. Classification in a normalized feature space using support vector machines. *IEEE Transactions on Neural Networks*, 14(3), 2003.

S. Gross, O. Russakovsky, C. Do, and S. Batzoglou. Training conditional random fields for maximum labelwise accuracy. In *Neural Information Processing Systems*, 2006.

T. Hazan and A. Shashua. Convergent message-passing algorithms for inference over general graphs with convex free energies. In *Uncertainty in Artificial Intelligence*, 2008.

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

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

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

T. Heskes. Convexity arguments for efficient minimization of the Bethe and Kikuchi free energies. *Journal of Artificial Intelligence Research*, 26:153–190, 2006.

J. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Grundlehren Text Editions. Springer Berlin Heidelberg, 2001.

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

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

S. Jiang, D. Lowd, and D. Dou. Learning to refine an automatically extracted knowledge base using Markov logic. In *International Conference on Data Mining*, 2012.

P. Juszczak, D. Tax, and R. Duin. Feature scaling in support vector data description. In *Conference of the Advanced School for Computing and Imaging*, 2002.

S. Kakade, O. Shamir, K. Sindharan, and A. Tewari. Learning exponential families in high-dimensions: Strong convexity and sparsity. In *Artificial Intelligence and Statistics*, 2010.

S. Khamis, V. Morariu, and L. Davis. Combining per-frame and per-track cues for multi-person action recognition. In *European Conference on Computer Vision*, 2012.

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

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

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

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

S. Kutin and P. Niyogi. Almost-everywhere algorithmic stability and generalization error. In *Uncertainty in Artificial Intelligence*, 2002.

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

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

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

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

B. London, S. Khamis, S. Bach, B. Huang, L. Getoor, and L. Davis. Collective activity

detection using hinge-loss Markov random fields. In *CVPR Workshop on Structured Prediction: Tractability, Learning and Inference*, 2013b.

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

B. London, B. Huang, and L. Getoor. The benefits of learning with strongly convex approximate inference. In *International Conference on Machine Learning*, 2015a.

B. London, B. Huang, and L. Getoor. Stability and generalization in structured prediction. Preprint, 2015b.

O. Mangasarian and T. Shiau. Lipschitz continuity of solutions of linear inequalities, programs and complementarity problems. *SIAM Journal on Control and Optimization*, 25(3):583–595, 1987.

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

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

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

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

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

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

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

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

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

T. Meltzer, A. Globerson, and Y. Weiss. Convergent message passing algorithms – a unifying view. In *Uncertainty in Artificial Intelligence*, 2009.

Z. Meng, B. Eriksson, and A. Hero III. Learning latent variable Gaussian graphical models. In *International Conference on Machine Learning*, pages 1269–1277, 2014.

O. Meshi, A. Jaimovich, A. Globerson, and N. Friedman. Convexifying the Bethe free energy. In *Uncertainty in Artificial Intelligence*, 2009.

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.

S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1-3):161–193, 2006.

S. Negahban, P. Ravikumar, M. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

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

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

D. Pollard. *Convergence of stochastic procresses*. Springer-Verlag, 1984.

J. Pujara, H. Miao, L. Getoor, and W. Cohen. Knowledge graph identification. In *International Semantic Web Conference*, 2013.

J. Pujara, B. London, and L. Getoor. Budgeted online collective inference. In *Uncertainty in Artificial Intelligence*, 2015.

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.

P. Ravikumar, M. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

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

D. Roth and W. Yih. Integer linear programming inference for conditional random fields. In *International Conference on Machine Learning*, 2005.

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

R. Samdani and D. Roth. Efficient decomposed learning for structured prediction. In *International Conference on Machine Learning*, 2012.

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

M. Schmidt. minFunc. `http://www.di.ens.fr/~mschmidt/Software/minFunc`, 2013a.

M. Schmidt. UGM: Matlab code for undirected graphical models. `http://www.di.ens.fr/~mschmidt/Software/UGM`, 2013b.

A. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Distributed message passing for large scale graphical models. In *Computer Vision and Pattern Recognition*, 2011.

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

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

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

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11:2635–2670, 2010.

Ö. Sümer, U. Acar, A. Ihler, and R. Mettu. Adaptive exact inference in graphical models. *JMLR*, 12:3147–3186, 2011.

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

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

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

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

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

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

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

M. Wainwright, T. Jaakkola, and A. Willsky. A new class of upper bounds on the log partition function. *IEEE Trans. on Information Theory*, 51(7):2313–2335, 2005.

Y. Weiss, C. Yanover, and T. Meltzer. MAP estimation, linear programming and belief propagation with convex free energies. In *Uncertainty in Artificial Intelligence*, 2007.

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

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