Learning with large examples: A short survey

Sanyam Kapoor, Chhavi Yadav
{sk6876,cy1235}@nyu.edu

December 19, 2017

Abstract

Most data around us has an internal structure - implicit or explicit. The knowledge of the semantics of those internal structures can help us learn more about the data and subsequently generalize well. Such problems are classified under the umbrella of *Structured Prediction* where the typical output space tends to be exponential with respect to the inputs. The collective nature of the inference in such relational models brings new challenges. Many a time the number of examples used to train structured prediction models are very few in number, sometimes even one. Theory explaining generalization in these settings is scarce. In this survey, we present the theoretical analysis to address the following question - *Can we generalize from a single example with large internal structure?* via standard learning theoretic tools - *Rademacher complexity, collective stability and graph- dependence*. In particular, our contribution is the analysis and critique of an approach based on the framework of Markov Networks, to provide *PAC* generalization bounds for graph-based hypotheses. To contrast, we survey a combinatorial approach for generalization bounds in a similar setting. We also make comments on practical applications of these bounds.

Contents

1 Introduction

Structured Prediction is the set of problems in supervised machine learning where the output is a structure/object instead of a scalar or real value. Typically, it involves joint reasoning over the interdependent output variables. A canonical example is the task of image segmentation where we aim to label each pixel based on a proximity measure. In simpler cases, it could just be background/foreground separation and in more advanced scenarios it would be dense labelling each pixel with an object class. Another popular example would be the task of parse tree generation from natural language.

These examples exhibit the existence of an implicit structure which is realized via some form of dependency. In contrast, some datasets have explicit structure by nature. For instance connected users in a social network tend to have similar traits. Posing these problems as standard multi-class supervised classification tasks is troublesome because each output label can have an exponential number of possibilities. For instance, the image segmentation task could be seen as a multi-class classification task where the dense labellings for foreground/background segmentation would induce a total of $2^{m \times n}$ possible segmentations for an image of size $m \times n$. This is huge even for a run-of-the-mill digital camera nowadays.

Instead we turn to *Structured Prediction* to provide global collective inference which reduces the sample output space by exploiting structural relations within each sample. In this theoretical survey, we analyse two recent works which provide different approaches to explain generalization bounds for structure learning from samples which have a large internal structure. The rest of the survey is organized as follows - we first discuss generalization bounds for graph-based hypotheses and realizing some shortcomings of the work, we explore an alternative combinatorial approach under restrictions.

2 Graph-based Hypotheses

In this survey, we are interested in models that perform collective inferences and graph-based hypotheses are a natural way of modelling the relations and dependencies. The inferences are driven by the nature of topology of the graph. More formally, we put forth some definitions from the standard graphical models literature.

First let us introduce some standard notations. Let $\mathcal{G}$ be a general family of undirected graphs and $\mathcal{G}_n$ be the family of such graphs of order n. $\mathcal{N}(i)$ represents the neighbors of vertex i in any graph $G \in \mathcal{G}_n$ and $\mathcal{N}^d(i)$ represents all such neigh-

bors within an edge distance d in the graph. Let $\mathcal{X}$ represent the measurable input space and $\mathcal{Y}$ represent the measurable output space. We refer a complete instance of sample as $z \in \mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$.

Definition 1. *For an undirected graph $G \triangleq (V, E)$, the random field is given by the set of all random variables $\mathbf{Z} \triangleq \{Z_i : i \in V\}$. We say that $\mathbf{Z}$ is Markovian if $\forall i$ $Z_i \perp\!\!\!\perp \bar{\mathbf{Z}} | \mathbf{Z}_{\mathcal{N}^d(i)}$ where $\bar{\mathbf{Z}} \triangleq \mathbf{Z} \backslash \mathbf{Z}_{\mathcal{N}^d(i)}$.*

In other words, a random field is *Markovian* if every random variable is conditionally independent of the remainder of the graph given its neighbors. This is also alternatively referred to as the *Markov blanket* of Z_i [1].

We aim to learn a hypothesis from the hypothesis class $\mathcal{H} = \{h : \mathcal{X}^n \to \hat{\mathcal{Y}}^n\}$ where $\hat{\mathcal{Y}}$ is not necessarily same as $\mathcal{Y}$. We are also interested in performing collective inference since any change in the output of one sample might affect the output of many others.

Definition 2. *For any hypothesis h, a graph $G \in \mathcal{G}_n$ and input $\mathbf{X}$, let the random variables $\mathbf{Y} \triangleq h_G(\mathbf{X})$ corresponding to the predicted output vectors. Let the joint distribution of $(\mathbf{X}, \mathbf{Y})$ be represented by $\mathbb{P}_{G,h}$. We say that the hypothesis class $\mathcal{H}$ is d^{th}-order graph-based if the following condition holds*

$$\mathbb{P}_{G,h}(Y_i | \mathbf{X}, \mathbf{Y} \backslash Y_i) = \mathbb{P}_{G,h}(Y_i | \mathbf{X}_{\mathcal{N}^d(i)}, \mathbf{Y}_{\mathcal{N}^d(i)}) \tag{1}$$

Put simply, the predictions are only dependent of neighbors of a given random variable until a graph distance of d and don't affect or get affected by other random variables in the random field. Many familiar graphical models fall in this category like Markov Random Fields (MRFs) and Conditional Random Fields (CRFs) [5]. This property is going to be crucial in our further analysis by helping us limit the complexity of the hypothesis class. A subtle difference should be noted between $\mathbb{P}_{G,h}$ and the "actual" underlying data distribution.

Let $\ell : \mathcal{Y} \times \hat{\mathcal{Y}}^n \to \mathbb{R}$ be a loss function. We define the empirical loss of the hypothesis over the given realization of $\mathbf{Z}$ as $\hat{R}_G(h, \mathbf{Z}) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, h(\mathbf{X})_i)$. The "true" risk over all realizations of $\mathbf{Z}$ is then given by $R_G(h) \triangleq \mathbb{E}_{\mathbf{Z}}[\hat{L}(h, \mathbf{Z})]$. We call this the "Generalization Error" [1] [7] which quantitatively tells us the quality of our selected hypothesis over new unseen examples. In all generalization analyses, this is the quantity we remain interested in, with our current survey being no exception.

We are generally given m independent draws from $\mathbb{P}_G(\mathbf{Z})$. These can be considered as a single realization of mn random variables whose distribution factorizes

[1] Some works refer to "generalization error" as $|R_G(h) - \hat{R}_G(h, \mathbf{Z})|$, but that will not affect our analysis.

over each realization of m subsets of size n. This allows us to consider just one large example for our generalization setting. We are able to make this assumption because of the following theorem.

Theorem 1. *Let* $\mathbf{Z}$ *be a random field on a graph* G *and* $\mathbf{Z}'$ *be a random field on* $G' \triangleq \bigcup_{i=1}^{m} G_i$ *such that* $G_i \simeq G$*, representing* m *realizations of* $\mathbf{Z}$*. If* $\mathcal{H}$ *is graph-based, then for any* $m, n \geq 1$*, any* $G \in \mathcal{G}_n$ *and any* $h \in \mathcal{H}$ *we have that* $R_G(h) = R_{G'}(h)$

Since, our random variables are now dependent we need some extra measures for concentration of dependent random variables. We introduce the notion of *total variation* which incorporates the maximal variation among two probability distributions over the σ-algebra Σ as

$$\|\mathbb{P} - \mathbb{Q}\|_{TV} \triangleq \sup_{A \in \Sigma} |P(A) - Q(A)| \tag{2}$$

Definition 3. *Let* $\mathbf{Z} \triangleq \{Z_i\}_{i=1}^{n}$ $(Z_i \in \mathcal{Z})$ *be random variables with distribution* $\mathbb{P}$*,* $\mathbf{z} \in \mathcal{Z}^{i-1}$*,* $a, b \in \mathcal{Z}$*, we define* $\eta_{i,j}$ *the measure of dependence between two conditional probability distributions as*

$$\eta_{i,j} = \sup_{i} \|\mathbb{P}(\mathbf{Z}_{j:n}|\mathbf{Z}_{1:i-1}, z_i) - \mathbb{P}(\mathbf{Z}_{j:n}|\mathbf{Z}_{1:i-1}, z_i')\|_{TV} \tag{3}$$

We define a matrix $\mathbf{\Theta} \in \mathbb{R}^{n \times n}$ using $\eta_{i,j}$ and use the matrix infinity norm to represent the slowest decay of dependence for the given graph topology - $\|\mathbf{\Theta}\|_\infty \triangleq \max_i \|\theta_i\|_1$ where $\|\theta_i\|_1$ represents the 1-norm of each row of the matrix. The complete definition is an upper triangular matrix as

$$\theta_{i,j} \triangleq \begin{cases} 1 & i = j \\ \eta_{i,j} & i < j \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

3 Generalization bounds

In this section we realize a few important properties - algorithmic stability, weak dependence and the rademacher complexity of the hypotheses set defined in the preceding section.

For a formal definition of stability, we observe how a function behaves when only a single dimension is perturbed. In learning theory, this notion of stability has been traditionally applied to observe the behavior of a learning algorithm with the addition or removal of training instances.

Definition 4. *Let $\mathcal{F} \triangleq \{f : \mathcal{Z}^n \to \mathbb{R}^N\}$ be the class of vector valued functions. We say $\mathcal{F}$ exhibits uniform collective stability β, if for any two inputs $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}^n$ that differ only at a single coordinate, $\sup_{f \in \mathcal{F}} \|f(\mathbf{z}) - f(\mathbf{z}')\|_1 \leq \beta$*

We also will have an updated definition of Rademacher Complexity because our random variables $\mathbf{Z}$ are not independent now, which is different from the canonical definition [2].

Definition 5. *Let $\mathbf{Z} \triangleq \{Z_i\}_{i=1}^n$ ($Z_i \in \mathcal{Z}$) be a set of random variables. Let $\{\sigma_i\}_{i=1}^n$ be independent random variables known as the Rademacher variables taking values $\{\pm 1\}$. We define the Empirical Rademacher Complexity as*

$$\hat{\mathfrak{R}}(\mathcal{F}, \mathbf{Z}) \triangleq \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f_i(\mathbf{Z}) | \mathbf{Z}\right] \tag{5}$$

In light of this, we subsequently have the Rademacher Complexity as $\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}[\hat{\mathfrak{R}}(\mathcal{F}, \mathbf{Z})]$.

To accomodate a variety of loss functions, we also have an admissibility constraint over the loss functions as defined below.

Definition 6. *A loss function $\ell : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R}$ is (M, λ)-admissible if there exist finite constants M and λ such that for any $y, y' \in \mathcal{Y}$ and $\hat{y}, \hat{y}' \in \hat{\mathcal{Y}}$, we have $|\ell(y, \hat{y}) - \ell(y', \hat{y})| \leq M$ and $|\ell(y, \hat{y}) - \ell(y, \hat{y}')| \leq \lambda \|\hat{y} - \hat{y}'\|_1$*

We finally present the main result from [6].

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

$$R_G(h) \leq \hat{R}_G(h, \mathbf{Z}) + 2\lambda \sum_{j=1}^{k} \mathfrak{R}_n(\mathcal{H}^j) + (M + \lambda\beta)\|\mathbf{\Theta}_n\|_\infty \sqrt{\frac{ln(1/\delta)}{2n}} \tag{6}$$

The proofs of this theorem are achieved via series of technical Lemmas which we leave out for brevity and we invite the reader to go through an extended work of [6]. Instead, we will focus on our analysis of this generalization bound and make some comments on its applicability.

4 Extended Discussion on Theorem 2

The concentration inequality given by Theorem 2 presents how the empirical error for a given graph and realization of $\mathbf{Z}$ concentrates around the "true" risk which is what we sought after in the beginning. This bound is affected by three influential properties - *Rademacher Complexity* of the graph-based hypotheses, *uniform collective stability* of the loss functions and the *slowest decay of dependence* in the graph. We present our analysis for each of these terms below.

4.1 Rademacher Complexity of graph-based Hypotheses

Intuitively, the complexity of the hypothesis set will be governed by the complexity of dependencies modelled in the graph. Quite naturally, a simplistic model might not capture the dependencies in the underlying data whereas an overly complex model might just overfit the data and prove to be bad for generalization. Hence, we must trade-off between this complexity and modelling the dependencies. This is captured by the measure of the Rademacher complexity term above.

Subsequently, this also brings us to an important practical observation - given a large graph representing the data e.g. a large social network, it might be possible to prune some of the dependencies in the graph to generalize better. However, one should be careful as this acts like a regularization technique and appropriate trade-offs must be considered.

4.2 Uniform Collective Stability of loss functions

As mentioned earlier, we perform joint inference over all the inputs. This obviously leads to a situation where any change of one random variable could adversely affect many others. We quantify this effect via the notion of Uniform Collective Stability. We have a favorable bound when $\beta = O(1)$. In other words, the effect of changing a single node in the graph should converge to a constant. However, this condition is overly idealistic on which we comment later.

4.3 Decay of dependence in the graph

This is perhaps the most important term in the bound and relates directly to the original problem setting where the input random variables have inherent dependencies among them. Ideally, we would like $\|\mathbf{\Theta}_n\|_\infty = O(1)$ or in other words be

independent of n. This represents how far out in the graph does a random variable influence another random variable (captured by the notion of total variation distance). This would make our bound very favorable and follow the order of $O(1/\sqrt{n})$. Even in the case where $\|\mathbf{\Theta}_n\|_\infty$ is sub-logarithmic, we still will have a favorable bound as it will converge to zero in limit of n. We believe that a bound on this term has a direct impact on the Rademacher Complexity of the hypothesis set because modelling long-range dependencies in the graph will require more complex hypotheses.

4.4 Practical Considerations

All the conditions that we point out above, analyse the bound in a fairly idealistic expectation in the hope that right hand side of the bound follows the order of $O(1/\sqrt{n})$. We would also require that the number of parameters to be trained in this model don't grow with the size of the structure to make sure we have a uniform convergence bound.

Luckily, owing to the analyses presented in [6], it turns out that a class of popular graphical models called the *Templated Structured Models* admit favorable constraints of $\beta = O(1)$ and $\mathfrak{R}_n(\mathcal{H}) = O\left(\sqrt{lg(n)/n}\right)$ which makes our idealistic expectations somewhat achievable. However, this does not end our concerns. It turns out that the TSMs that satisfy these constrainsts are only the ones whose inference objectives are strongly convex and the set of clique templates obey $|\mathcal{T}| = O(1)$.

In practice, it is hard to qualify these constraints and hence we introduce another approach to this analysis which considers a verifiable metric based on similar intuitions (under certain compromises).

5 Relational Learning in large graphs

In light of setups we had earlier, relational learning involves a similar large single graph with increasing size and complex dependencies across data over the graph. We introduce suitable assumptions on features and an intuitive dependence assumption i.e. finite distance dependence as presented in [4]. To characterize the complex dependencies, a combinational method is presented and PAC guarantees for learning via MLE (Maximum Likelihood Estimation) is given.

5.1 Basic Markov Networks setup

We assume a standard Markov model definition with the well-known log-linear potential functions for maximal cliques as

$$p(y|x) = \frac{1}{Z(z)} \prod_{T \in \mathcal{T}} \prod_{c \in C_T} \psi(x_c, y_c; \theta_T) \tag{7}$$

$$\psi(x_c, y_c; \theta_T) = exp(\langle \theta_T, \phi(x_c, y_c) \rangle) \tag{8}$$

where θ_T is the real-valued parameter vector and (x_c, y_c) are assignments of x and y over the clique c. $\mathcal{T}$ represents the set of all clique templates and each template has the same parameterization.

For a single clique template, the full Markov Network model is written as

$$p_\theta(y|x) = \frac{1}{Z_n(\theta; x)} exp \left\{ \langle \theta_T, \sum_{c \in C_G} \phi(x_c, y_c) \rangle \right\} \tag{9}$$

where $Z_n(\theta; x) = \int_{\mathcal{Y}} exp \left\{ \langle \theta_T, \sum_{c \in C_G} \phi(x_c, y_c) \rangle \right\} dy$ is the partition function. Our aim to maximize the log-likelihood function and get as close as possible to the true parameter θ^*.

$$\ell^{(n)}(\theta; x) = \frac{1}{n} \sum_{c \in C_G} \langle \theta_T, \phi(x_c, y_c) \rangle - \frac{1}{n} \log Z_n(\theta; x) \tag{10}$$

Now, our maximum likelihood estimate is represented by $\hat{\theta}_n = \underset{\theta}{\text{argmax}}\ \ell^{(n)}(\theta; x)$

5.2 Constraints and Assumptions

5.2.1 Finite Distance Dependence

Definition 7. *Let $\psi(v, v')$ be the minimal path length between any two nodes v and v' in the graph. The distance between two cliques c and c' is defined as $\psi(c, c') \triangleq \max_{v \in c, v' \in c'} \psi(v, v')$*

Definition 8. *There exists a constant $\lambda^* > 0$ such that the correlation coefficient $\rho_j(c, c') = 0$ for any feature vector element ϕ_j between cliques c and c′ provided that $\psi(c, c') > \lambda^*$.*

This definition captures the intuition that the cliques are independent if they are far enough. We require that the non-neighbors of a node are not overly dependent on the neighbors of the node.

In our previous analysis, we came across the notion of weak dependence. However, that condition has been relaxed in this work. Interestingly, the dependence need not decay as the distance from a random variable in the graph increases but all we require is that the dependence is bounded by the distance, no matter how strong.

5.2.2 Bounded Features

Definition 9. *The magnitude of any feature vector element is upper bounded: $\phi_{max} \triangleq \max_{j,x_c,y_c} |\phi_j(x_c, y_c)|$. In addition we assume that node degrees remain bounded as n grows. $deg(v) \leq d \leq \infty$.*

5.2.3 Minimum Curvature

Definition 10. *The minimum eigenvalue of the feature covariance matrix Λ_{min} $(var_{\theta^*[\phi|x]}) \geq C_{min} > 0$*

This assumption manifests a well known result where if all the eigenvalues of the Hessian are positive, the Hessian is *positive-definite* and we have a unique solution for the MLE due to the convexity. In other cases, we wouldn't otherwise be able to achieve a unique consistent solution.

Our goal is to find the sample complexity of MLE. Given the convergence rate of gradients of $\hat{\theta}_n$ we can obtain the sample complexity of MLE by Taylor's expansion about the true parameter θ^*. This estimation is done by upper bounding the number of "close-by" cliques for a given clique which can be solved by the combinatorial analysis of interaction nodes and cliques over the graph.

6 PAC Generalization bounds for MLE

We first present a few definitions required for the analysis.

Definition 11. *We define $C(c, \lambda)$ as the set of cliques within a distance of λ. More formally $C(c, \lambda) = \{c' \in C_G : 1 \leq \psi(c, c') \leq \lambda\}$.*

Definition 12. *We define $V(c, \lambda) = \{v' \in V : v' \in c', v \notin c, c \neq c', \psi(c, c') \leq \lambda\}$ represent the set of vertices in those cliques, in which distances to the given clique c are not larger than λ, excluding vertices in the clique itself.*

This definition is just a continuation of the previous one to make counting of vertices convenient. We first provide the results for upper bound on $|C(c, \lambda)|$.

Lemma 3.

$$|V(c, \lambda)| < 3d^{\lambda} \tag{11}$$

where d is the maximal degree of nodes in V.

Lemma 4.

$$|C(c, \lambda)| < 3\binom{d}{m-1}d^{\lambda} \tag{12}$$

where m is the number of vertices in a clique.

These results have been obtained by a naive couting method which counts duplicates. The reader is referred to the proofs presented in [4]. A tighter bound is given by the following theorem.

Theorem 5. *For any clique $c \in C_G$ and positive integer $\lambda > m$, d is the maximal degree of vertices in V and m is the number of vertices in a clique, we have*

$$|C(c, \lambda)| < \frac{3}{2}\left(\frac{1}{m} + 1\right)\binom{d}{m-1}d^{\lambda} \tag{13}$$

6.1 Sample Complexity Bounds

Under the assumptions defined above, we have the following lemma for the convergence bounds of θ.

Lemma 6. *Under the assumptions defined above, for any $t > 0$*

$$\mathbb{P}\left[\|\nabla \ell^{(n)}(\theta^*)\|_2 > t\right] < \frac{r^2\phi_{max}^2}{nt^2}\left[\frac{3}{2}\left(\frac{1}{m} + 1\right)\binom{d}{m-1}d^{\lambda^*} + 1\right] \tag{14}$$

where r is the total number of parameters (dimension of θ) and n is the total number of examples.

The PAC learning bound follows as a consequence by the Taylor's expansion of the normalized log-likelihood around θ^*. The detailed proof can be read in [4].

Using the above Lemma's and Taylor's expansion we arrive at

$$\mathbb{P}\left[\|\hat{\theta}_n - \theta^*\|_2 > \epsilon\right] \leq \mathbb{P}\left[\|\nabla \ell^{(n)}(\theta^*)\| > \frac{\epsilon C_{\min}}{2}\right] \tag{15}$$

and consequently the following theorem.

Theorem 7. *Under the assumptions defined above, for any $\epsilon, \delta > 0$, the MLE learns the parameters within L_2 error ϵ with probability at least $1 - \delta$, provided that*

$$n > \frac{4r^2\phi_{max}^2}{\epsilon^2 \delta C_{min}^2}\left[\frac{3}{2}\left(\frac{1}{m} + 1\right)\binom{d}{m-1}d^{\lambda^*} + 1\right] \tag{16}$$

7 Extended discussion on Theorem 7

This result shows us the importance of λ^*, which denotes the largest distance dependency in the given data graph. Increasing this affects the sample complexity exponentially which falls in line with our intuition as longer range dependencies would require more samples to generalize well. This would require a more complex hypothesis set.

From the sample complexity bound, we see that larger C_{min} will make the learning easier which follow from standard results on the relationship between the Hessian of a matrix and the time to convergence of convex objectives. The result for the size of the clique also falls in line with our intuition that for larger clique sizes denoted by m, we will require more number of samples for generalization. Even for r and d, more number of parameters and a larger degree in the graph leads to a requirement of more samples.

[4] also observes that the convergence rates are of the order of $O(1/n)$ in the case of relational learning whereas [3] reports results in the order of $O(1/e^n)$. This points to an inherent difficulty in learning with relational data.

7.1 Comparisons with Theorem 2

The approaches taken to explain generalization bound within the framework of Markov Networks in both Theorem 2 and 7 present an interesting contrast. While

one takes a more fine grained approach by analyzing bounds on the random variables in a much generalized setting, the other makes various compromises to make for more verifiable bounds, which we highlight below.

Graph distance seems to be an imperative quantity in both analyses however their treatment varies. The former approach proposed the notion of "decay" of dependence over the graph being a strict requirement for generalization. This was encapsulated in the matrix norm of $\|\mathbf{\Theta}\|_\infty$ which is relatively a more difficult constraint to qualify. However in the latter approach, we observe that the analysis does not put much emphasis on weak or strong dependence but only expects the dependence to vanish after a bounded graph distance λ^*. This constraint is more relaxed and interpretable in the bound. It should also be noted that the dependence need not decay in the latter case with the graph distance.

But, it should also be noted that the combinatorial analysis is provided in a more restricted setup to arrive at the PAC bounds which include the choice of L2 error and limiting the the number of clique templates to just one. This makes the analysis seemingly incomplete though the authors claim that this can be extended to other estimation methods as well.

The degree of the graph is explicitly mentioned in the combinatorial bound whereas the former results don't seem to be dependent on the degree of the graph. This can be more or less explained by the fact that the decay of dependence obivates the need to have such a term in the bound. Moreover, for graphs with a high number of connections, the combinatorial analysis will be vacuous. In that respect, the former analysis seem to hold more ground.

8 Challenges

It is striking that the confidence requirement is not logarithmic anymore in the combinatorial analysis and has a greater impact on the sample complexity in the order of $O(1/\delta)$ instead of $O(\log 1/\delta)$ in the former analysis. While hinting at a harder learning problem, this poses a contradictory relationship with δ as presented by the result in Theorem 2.

Surprisingly, we started with the notion of β-stability for joint inference tasks whereas the combinatorial approach does not report any such requirement. This brings us to the question of whether the concept of stability is relevant or not? It is unclear whether the notion of stability manifests itself in the bounds in another form. It could be alternatively viewed as the fact that bounding the empirical parameter estimate to the true Bayes estimate (θ^*) subsumes the idea of stability.

This analysis diverges from the the former which uses stability over the predictions - predictive stability.

As the number of parameters to be learned, the clique size and the maximum degree in the graph increase with the size of the graph, then the bounds become vacuous in both cases (TSMs in the former analysis).

9 Conclusion

In this work, we presented our learnings on two approaches to generalization within the framework of Markov Networks. While we were able to observe certain intuitive ideas manifest in the bounds as expected, there were certain contrasts that were not fully explainable. The trade-offs in the analyses are apparent to achieve tight bounds in restricted settings. We've outlined a few challenges we believe to be of importance to be able to make stronger claims about the generalization for large structured examples.

References

[1] David Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, New York, NY, USA, 2012.

[2] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, March 2003.

[3] Joseph Bradley and Carlos Guestrin. Sample complexity of composite likelihood. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 136–160, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

[4] Peng He and Changshui Zhang. Non-Asymptotic Analysis of Relational Learning with One Network. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 320–327, Reykjavik, Iceland, 22–25 Apr 2014. PMLR.

[5] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[6] Ben London, Bert Huang, Ben Taskar, and Lise Getoor. Collective stability in structured prediction: Generalization from one example. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 828–836, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[7] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012.