Perspectives on Probabilistic Graphical Models

DONG LIU

Doctoral Thesis in Electrical Engineering
Stockholm, Sweden 2020

Devision of Information Science and Engineering
TRITA-EE XXXX KTH, School of Electrical Engineering and and Computer Science
ISSN .... SE-100 44 Stockholm
ISBN .... SWEDEN

Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framlägges till offentlig granskning för avläggande av teknologie doktorsexamen i Elektroteknik onsdag den 13 september 2017 klockan 13.15 i F3, Lindstedtsvägen 26, Stockholm.

Tryck: Universitetsservice US AB

To my beloved

Abstract

Sammanfattning

Acknowledgements

Contents

Acronyms and Notations

Notations

X	random variable
x	realization of the random variable X
$\mathcal{X}$	alphabet of the random variable X
X_i^k	random sequence $(X_i, \ldots, X_k)$
x_i^k	realization of the random sequence X_i^k
$\mathcal{X}_i^k$	alphabet of the random sequence X_i^k
X^k	random sequence $(X_1, \ldots, X_k)$
x^k	realization of the random sequence X^k
$\mathcal{X}^k$	alphabet of the random sequence X^k
$X_i^{k \setminus n}$	random sequence $(X_i, \ldots, X_{n-1}, X_{n+1}, \ldots, X_k)$
$x_i^{k \setminus n}$	realization of the random sequence $X_i^{k \setminus n}$
$\mathcal{X}_i^{k \setminus n}$	alphabet of the random sequence $X_i^{k \setminus n}$
$X^{k \setminus n}$	random sequence $(X_1, \ldots, X_{n-1}, X_{n+1}, \ldots, X_k)$
$x^{k \setminus n}$	realization of the random sequence $X^{k \setminus n}$
$\mathcal{X}^{k \setminus n}$	alphabet of the random sequence $X^{k \setminus n}$
$\|\| \cdot \|\|$	set cardinality
f_X	p.d.f. of the continuous random variable X
p_X	p.m.f. of the discrete random variable X
$\mathcal{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2

$\mathrm{D}(\cdot\|\|\cdot)$	Kullback-Leibler divergence
$\mathrm{D}_\tau(\cdot\|\|\cdot)$	τ-th order Rényi divergence
$\mathrm{C}(\cdot,\cdot)$	Chernoff information
$\mathrm{E}[\cdot]$	expectation
$\partial\cdot$	boundary of a closed set
$\hat{\partial}\cdot$	upper boundary of a two-dimensional closed set
$\check{\partial}\cdot$	lower boundary of a two-dimensional closed set
$\log(\cdot)$	natural logarithm

Chapter 1

Introduction

Motivate the research in probabilistic models.

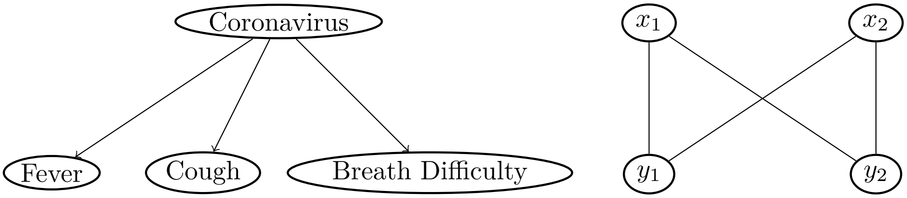
1.1 Motivations

Most tasks conducted by a person or an automated system requires a fundamental ability of *reasoning*, which is always about reaching a conclusion based on available information. At times, a conclusion is not enough and it is also required to know how reliable the conclusion is. Take the coronavirus that started from Wuhan, China at the end of 2019, as example, a doctor needs checks the information about a person to reason if the person is infected by the coronavirus. The relevant information includes symptoms such as fever, cough, breathing difficulties and probably kidney failure in severe cases. (maybe a small figure of coronavirus here.) Even after the doctor has concluded as positive or negative of ccoronavirus for the person, the natural question is why and how *confident* the diagnose is.

Two problems are inevitable to conduct the reasoning:

- How should we specify the relationship between a conclusion and the available information? In the coronavirus example, the counterpart question to answer is how the doctor should relate coronavirus infection with the symptoms. This step is called *modeling* which represents a reasoning problem abstractly by specifying the relationship between known information and unknown part, in preparation of answer query on it.
- With the model, how a conclusion should be made? This process of reaching a answer to the query is called *inference*. something about coronavirus

As times, a model is not totally fixed since one may not be sure the correctness of the assumptions about the model. A typical strategy is to leave some freedom in the configuration of the model at beginning. By using previous observations or information, the model is adjusted to be able to explain the observation in more reasonable way. This adds the following problem in reasoning:



(a) A directed graphical model for Example 1.

(b) An undirected graphical model.

Figure 1.1: Different perspectives on probabilistic graphic models. 1.1a A toy Bayesian network. 1.1b A toy Markov random field.

- Instead of having a fixed model at the first step, a set of model is given. We then need to choose one model based previous observations to do inference in order to make conclusion or answer query. This phase of choosing a model is called *learning.*

With all the discussed problems above, modelling, inference and learning, our purpose is to carry out reasoning with being aware of how confident a conclusion or answer is. These problems can be treated nicely with probabilistic models. Probabilistic models is built on the fundamental calculus of probability theory that is natural to accommodate the *uncertainty*, which is desired in reasoning. In additional, the probabilistic models offers rich space to modeling problems, where inference can be carried on either exactly or approximately. **More importantly, the modeling or modeling learning part is not necessarily coupled with inference algorithm.** This proper separation allows free that a certain family of general inference algorithms can be applies to a broad class probabilistic models. It offers the freedom of trying different models of a class without the need of replacing inference algorithm.

Example 1. *Consider the coronavirus infection problem. Using probabilistic modelk, we are able to model the problem in a rigid way. Additionally, we can make query more formally in probabilistic model framework. Assume each symptom among fever, cough and breathing difficulty can take value from* $\{\text{True}, \text{False}\}$. *Also the coronavirus infection is either true or false. One exemplified query can be*

$$P(\text{Infection} = \text{True} | \text{Fever} = \text{True}, \text{Cough} = \text{False}, \text{BreathingDifficulty} = \text{True}),$$

which is asking how likely the patient is infected by coronavirus if symptoms of both fever and breathing difficulty are observed but no sight of cough.

Example 2. *digital communication? Y, X do we need this example*

Todo:
Explain probabilistic graphical models with figure illustrated.
basic problems: 1. modeling, 2. inference, 3. learning?

what graphical models do and why
what learning means
the inference diagram here
the learning diagram here

- Structural learning
- parameter learning

the learning principle:

- Maximal likelihood estimation (MLE)
- Bayesian estimation
- Maximal conditional likelihood
- Maximal " ‘Margin‘"
- Maximum entropy

1.2 Scope and Thesis Outline

We organizing the general Bayesian problems and/or methods as the follows, based on which we would discuss the part that this thesis would cover.

- Inference
  - Exact Inference
    - Variable elimination
    - Belief probation
    - Junction tree method
    - ...
  - Approximate Inference
    - Stochastic Approximation (Particle methods)
      - Gibbs Samples
      - Importance sampling
      - Markov Chain Monte Carlo
      - ...
    - Deterministic Approximation (Variational methods)
      - Mean field
      - Loopy belief propagation
      - Expectation Propagation
      - ...

- Learning
    - Structure Learning
        - Directed graphical models
        - Undirected graphical models
    - Parameter Learning
        - Directed graphical models
        - Undirected graphical models

The main work of this thesis would about the deterministic approximate inference methods and parameter learning of graphical models, although the rest topics or issues would be referred in context in the above category.

Summary of Contributions

[1]

Publications

Tools (code) developed:

Chapter 2

Background

Background on probabilistic graphical models

2.1 Graphical representation and inference

.

graphical models

It is clear that undirected graphs should be explained, since Part I is work on it. For HMM, it can be viewed either a dynamic Bayesian network (chapter 6, Koller) or condition random field(introduction to CRF, Sutton).

intro to inference methods

2.2 Learning principles

It may be better to discuss the learning principle here.

Cited from 10-708 lecture6 note:

UNOBSERVED VARIABLES:

A variable can be unobserved or latent because it is a(n):

-Abstract or imaginary quantity meant to simplifiy the data generation process, e.g. speech recognition models, mixture models. -A real-world object that is difficult or impossible to measure, e.g. the temperature of a star, causes of disease, evolutionary ancestors. -A real-world object that was not measured due to missed samples, e.g. faulty sensors.

Discrete latent variables can used to partition or cluster data into sub-groups

Continuous latent variables (factors) can be used for dimensionality reduction (e.g. factor analysis, etc)

Dealing with latent variables

about clamping node

clamping node gives conditional distribution.

about ELBO bound

1. the bound used by EM

2. talk about ELBO/variational inference Variational Inference, which is closely related the bound used in EM.

Part I

Inference

Chapter 3

An alternative view of belief propagation

Content:

1. α Belief Propagation as Fully Factorized Approximation, GlobalSIP 2019.
2. α Belief Propagation for Approximate Bayesian Inference, under review.

3.1 α belief propagation

3.2 Convergence study

3.3 Experimental results

3.4 Summary

Chapter 4

Region-based Energy Neural Network Model

work in Region-based Energy Neural Network for Approximate Inference, under, review

4.1 Region-based graph and energy

4.2 RENN model for Approximate Inference

4.3 RENN model for markov random field training

4.4 Experimental results

4.5 Summary

Part II

Learning

Chapter 5

Learning with inference

5.1 learning Undirected graphical models/ MRF

move the MRF learning by using RENN here

I should read lecture note 7 of 10-708 again when writing this seciton.

5.2 Amortized/Neural Variational Learning and Inference of partial observed MRF

1. TRW as upper bound to partition function

2. Mean field or negative TRW as lower bound to partition function

combining above together, we can obtain two different lower bound of likelihood. Consider if worthy a paper.

- The log-likelihood of partial observed MRF is non-convex in general (log-sum-exp is convex, but the difference of two log-sum-exp functions might not be). This combination convert the original non-convex learning into convex optimization with regarding to MRF parameter? should be, but need a confirmation.
- 1. The speed of training can be improved by directly optimizing amortized beliefs.
- The bound becomes tighter by using clamping of variable, clamping can be done with or without selection of variables. No sampling is needed in training or inference.
- If need more contribution, use tree-reweighted hyper graph to obtain tighter bound.
- Not necessarily done here: the bound can also be further improved by important sampling.

Reference:

- 1. Wainwright, 2003, Tree-reweighted belief propagation algorithms and approximate ML estimation by pseudo-moment matching
- 2. Weller, 2015,Clamping Improves TRW and Mean Field Approximations
- 3. Mnih, 2014, Neural Variational Inference and Learning in Belief Networks, which describes a neural variational method for belief network. The major difference is the belief network as a DAG do not have the problem of partition function difficulty as MRF or partial observed MRF.

5.3 Notation

Random variable $\boldsymbol{v} \in \mathcal{X}_v$ that can be observed. Random variable $\boldsymbol{h} \in \mathcal{X}_h$ that is hidden variable and can not be observed.

5.4 Model and Problem Definition

We define the conditional probabilistic model as

$$p(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \tilde{p}(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta}), \tag{5.1}$$

with

$$Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} \tilde{p}(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta}) \tag{5.2}$$

$$\tilde{p}(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta}) = \exp\left\{-E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{\theta})\right\} \tag{5.3}$$

where $E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})$ is the average energy: $\mathcal{X}_v \times \mathcal{X}_h \to \mathbb{R}$.

We want to maximize the marginal likelihood:

$$\max_{\boldsymbol{\theta}} \log \sum_{\boldsymbol{h}} p(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta}) = \max_{\boldsymbol{\theta}} \log\ Z(\boldsymbol{v}, \boldsymbol{\theta}) - \log\ Z(\boldsymbol{\theta}), \tag{5.4}$$

where $Z(\boldsymbol{v}, \boldsymbol{\theta}) = \sum_{\boldsymbol{h}} \tilde{p}(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})$

5.5 A lower bound of the marginal likelihood

Denote $A(\boldsymbol{\theta}) = \log Z(\boldsymbol{\theta})$ and $A(\boldsymbol{v}, \boldsymbol{\theta}) = \log Z(\boldsymbol{v}, \boldsymbol{\theta})$

$$E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{\theta}) = -\langle \boldsymbol{\theta}, \boldsymbol{\varphi}(\boldsymbol{v}, \boldsymbol{h}) \rangle \tag{5.5}$$

and

$$\boldsymbol{\mu} = \mathbb{E}_{p(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})}[\boldsymbol{\varphi}(\boldsymbol{v}, \boldsymbol{h})]. \tag{5.6}$$

In case of overcomplete representation of $\boldsymbol{\varphi}$, $\boldsymbol{\mu}$ is the set of marginal distributions.

With mean field approximation,

$$A_M(\boldsymbol{v},\boldsymbol{\theta}) = \max_{\boldsymbol{\mu_v}\in\mathcal{M}_M} \langle\boldsymbol{\theta},\boldsymbol{\mu_v}\rangle + H(\boldsymbol{\mu_v}), \tag{5.7}$$

where $\mathcal{M}_M$ is the subspace of distributions where each variable is independent. And we have

$$A_M(\boldsymbol{v},\boldsymbol{\theta}) \leqslant A(\boldsymbol{v},\boldsymbol{\theta}). \tag{5.8}$$

With tree-reweighted approximation, TRW,

$$A_T(\boldsymbol{\theta}) = \max_{\boldsymbol{\mu}\in\mathcal{M}_T} \langle\boldsymbol{\theta},\boldsymbol{\mu}\rangle + H(\boldsymbol{\mu}), \tag{5.9}$$

where $\mathcal{M}$ is the subspace of distributions where each variable is independent. And we have

$$A_T(\boldsymbol{\theta}) \geqslant A(\boldsymbol{\theta}). \tag{5.10}$$

We define the lower bound of marginal loglikelihood:

$$\mathcal{L}(\boldsymbol{\theta}) = A_M(\boldsymbol{v},\boldsymbol{\theta}) - A_T(\boldsymbol{\theta}) \leqslant \log\sum_{\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta}). \tag{5.11}$$

Connection to RBM:

$$p(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})}\exp\boldsymbol{vWh}+\boldsymbol{vb}+\boldsymbol{va} \tag{5.12}$$

Note $p(\boldsymbol{h}|\boldsymbol{h})$ is exactly independent, and thus the $A_M(\theta) = A(\boldsymbol{\theta})$ can be achieved, then how tight the lower bound $\mathcal{L}(\boldsymbol{theta})$ would dependent only on the TRW bound.

5.6 Experiment

- Start with standard RBM section 4.2 in Amortized learning of MRFs
  - Try to break the conditinal independence by connecting nodes of $\boldsymbol{h}$
  - Extent to conditional RBM training for denoising and data completion
- high-order HMMs

Chapter 6

Powering the expectation maximization method by neural networks

content: Neural Network based Explicit Mixture Models and Expectation-maximization based Learning, under review

section/chapter transition text: mixture model could be obtained from clamping and condition on a discrete variable, ref to Geier, Locally conditional belief propagation. Weller, clamping variables and approximate inference

Remark 6.1. *RELATIONSHIP TO K-MEANS CLUSTERING Big picture: The EM algorithm for mixtures of Gaussians is like a soft version of the K-means algorithm.*

Remark 6.2. *EM lower bound + entropy of posterior of latent variable if a free energy. ref to 10-708 lecture6 note. EM using posterior of latent variable is equivalent to fully observable MLE where statistics are replaced by their expectations w.r.t the posterior.*

Can be viewed as two-node graphical model learning. 10-708lecture5-note

6.1 Normalizing flow

6.2 expectation maximization of neural network based mixture models

6.3 An alternative construction method

6.4 Experiments

6.5 Summary

Chapter 7

Powering Hidden Markov Model by Neural Network based Generative Models

content:

1. Powering Hidden Markov Model by Neural Network based Generative Models, ECAI 2020
2. Antoine Honore, Dong Liu, Hidden Markov Models for sepsis detection in preterm infants, ICASSP, 2020

HMM is an instance of 2-time-slice Bayesian network(2-TBN) (section 6.2.2 Koller). Also, it can be argued from CRF.

7.1 Hidden Markov Model

7.2 GenHMM

7.3 Application to phone recognition

7.4 Application to sepsis detection in preterm infants

7.5 Summary

Chapter 8

An implicit probabilistic generative model

content: Entropy-regularized Optimal Transport Generative Models, ICASSP 2019

8.1 Modeling data without explicit probabilistic distribution

8.2 Employing EOT for modeling

8.3 Experimental results

8.4 Summary

Part III

Epilogue

Chapter 9

Conclusion and Discussions

Bibliography

[1] Dong Liu, Baptiste Cavarec, Lars K Rasmussen, and Jing Yue. On dominant interference in random networks and communication reliability. In *ICC 2019-2019 IEEE International Conference on Communications (ICC)*, pages 1–7. IEEE, 2019.