Perspectives on Probabilistic Graphical Models

Dong Liu

Information Science and Engineering

KTH - Royal Institute of Technology

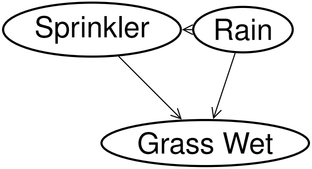
Profile page: https://firsthandscientist.github.io/

Slide is available at: https://github.com/FirstHandScientist/phdthesis

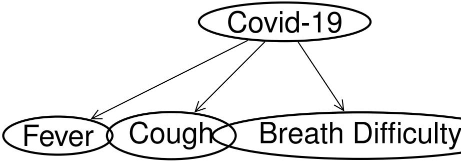
Why are Probabilistic Graphical Models interested?

PROBABILISTIC GRAPHICAL MODELS

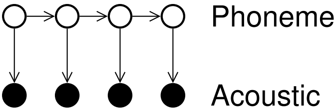
Directed: Bayesian Networks



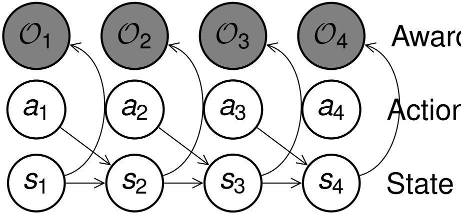
Is the sprinkler working?



Is the person get contiguous by COVID?



Speech Recognition

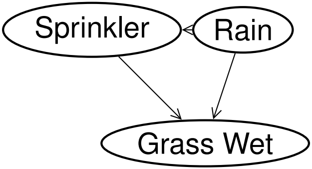


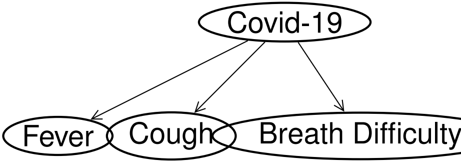
Communication system

Two key aspects to encode in a graphical model:

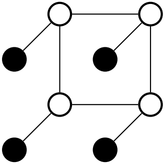
Probabilistic Graphical Models

Directed: Bayesian Networks

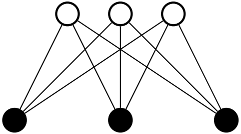




Undirected: Markov Random Field



Computer vision



Control, reinforcement learning

Two key aspects to encode in a graphical model:

- attributes of our interests in a system → variable nodes
- relationship of these factors (dependencies or independencies) → structures of a graph

Markov Random Field

Markov Random Field (MRF) and Fundamental Inference Problems

Let us walk through via MRF

- An MRF can be represented by a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ with each node $i \in \mathcal{V}$ is associated with a random variable X_i
- The probability distribution (Gibbs distribution) is

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \prod_a \psi_a(\boldsymbol{x}_a; \boldsymbol{\theta}_a),$$

where $\boldsymbol{x} = \{x_1, x_2, \cdots, x_N\}$, a indexes potential functions $\mathcal{I} = \{\psi_A, \psi_B, \cdots, \psi_M\}$ and $\boldsymbol{\theta}$ is set of potential function parameters. $Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} \prod_a \psi_a(\boldsymbol{x}_a; \boldsymbol{\theta}_a)$.

What do we do with graphical models?

USAGE OF GRAPHICAL MODELS

In general:

- Representation
  - In place of real systems
  - Abstraction of complex problems or systems (with subjective bias)
- Answer queries
  Evidence (observation) $\rightarrow$?? $\rightarrow$ Answers

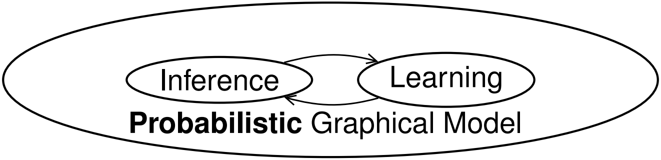
Two components interacting with each other:

USAGE OF GRAPHICAL MODELS

In general:

- Representation
  - In place of real systems
  - Abstraction of complex problems or systems (with subjective bias)
- Answer queries
  Evidence (observation) $\rightarrow$?? $\rightarrow$ Answers

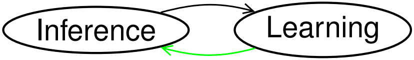
Two components interacting with each other:



Usage of Graphical Models

Why impact in two direction?

- Learning to Inference:



  A graphical model
  - built by expert knowledge, or
  - built by extracting information from evidence (empirical data).

- Inference to Learning:

  Model learning: an error trial process that compares inferred 'fact' and actual fact (evidence).
  Model learning usually needs inference as a subroutine, which sometimes are replaced by sampling in particle based methods.

USAGE OF GRAPHICAL MODELS

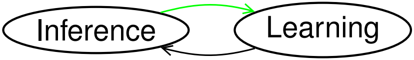
Why impact in two direction?

- Learning to Inference:

A graphical model

- built by expert knowledge, or
- built by extracting information from evidence (empirical data).

- Inference to Learning:

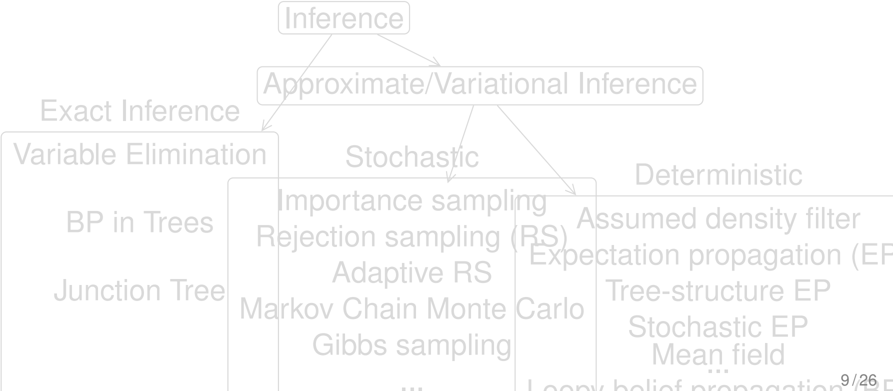


Model learning: an error trial process that compares inferred 'fact' and actual fact (evidence).
Model learning usually needs inference as a subroutine, which sometimes are replaced by sampling in particle based methods.

COMMON INFERENCE PROBLEMS

The common inference problems in a MRF $\mathcal{G}(\mathcal{V}, \mathcal{E})$:

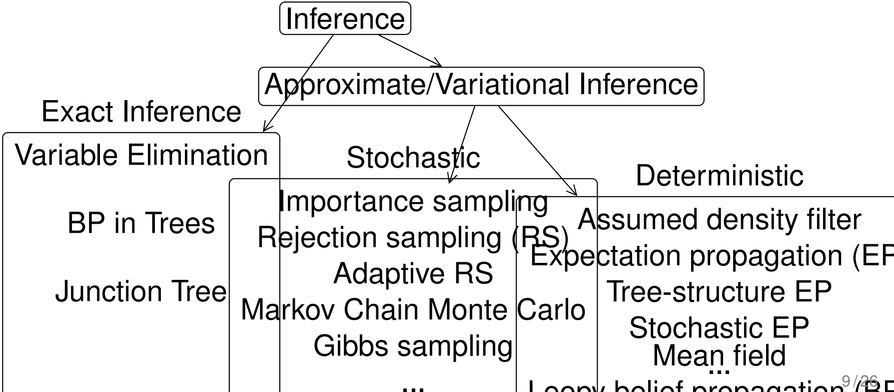
- Computing the likelihood of observed data.
- Computing the marginals distribution $p(\boldsymbol{x}_A)$ over particular subset $A \subset \mathcal{V}$ of nodes
- Computing the conditional distribution $p(\boldsymbol{x}_A | \boldsymbol{x}_B)$,
- Computing the most likely configuration $\operatorname{argmax}_{\boldsymbol{x}} p(\boldsymbol{x})$



Common Inference Problems

The common inference problems in a MRF $\mathcal{G}(\mathcal{V}, \mathcal{E})$:

- Computing the likelihood of observed data.
- Computing the marginals distribution $p(\boldsymbol{x}_A)$ over particular subset $A \subset \mathcal{V}$ of nodes
- Computing the conditional distribution $p(\boldsymbol{x}_A | \boldsymbol{x}_B)$,
- Computing the most likely configuration $\text{argmax}_{\boldsymbol{x}}\, p(\boldsymbol{x})$



INFERENCE ROUTINE IN LEARNING

What is $\boldsymbol{\theta}$ in $p(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})}\prod_a \psi_a(\boldsymbol{x}_a;\boldsymbol{\theta}_a)$?
A direct view:

$$\max_{\boldsymbol{\theta}} \log p(\boldsymbol{x};\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}} \sum_a \log \psi_a(\boldsymbol{x}_a;\boldsymbol{\theta}_a) \quad \underbrace{-\log Z(\boldsymbol{\theta})}_{\textit{Helmholtz free energy}} ,$$

An alternative view:

$$\frac{\partial \log p(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_a} = \frac{\partial \log \varphi_a(\boldsymbol{x}_a;\boldsymbol{\theta}_a)}{\partial \boldsymbol{\theta}_a} - \mathbb{E}_{p(\boldsymbol{x}_a;\boldsymbol{\theta})}\left[\frac{\partial \log \varphi_a(\boldsymbol{x}_a;\boldsymbol{\theta}_a)}{\partial \boldsymbol{\theta}_a}\right].$$

Remark:

- This essentially requires estimation of Helmholtz free energy or marginal probabilities.
- Stationary points translate into moment matching.

Inference Routine in Learning

What is $\boldsymbol{\theta}$ in $p(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})}\prod_a \psi_a(\boldsymbol{x}_a;\boldsymbol{\theta}_a)$?
A direct view:

$$\max_{\boldsymbol{\theta}} \log p(\boldsymbol{x};\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}} \sum_a \log \psi_a(\boldsymbol{x}_a;\boldsymbol{\theta}_a) \quad \underbrace{-\log Z(\boldsymbol{\theta})}_{Helmholtz\ \ free\ \ energy} \quad ,$$

An alternative view:

$$\frac{\partial \log p(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_a} = \frac{\partial \log \varphi_a(\boldsymbol{x}_a;\boldsymbol{\theta}_a)}{\partial \boldsymbol{\theta}_a} - \mathbb{E}_{p(\boldsymbol{x}_a;\boldsymbol{\theta})}\left[\frac{\partial \log \varphi_a(\boldsymbol{x}_a;\boldsymbol{\theta}_a)}{\partial \boldsymbol{\theta}_a}\right].$$

Remark:

- This essentially requires estimation of Helmholtz free energy or marginal probabilities.
- Stationary points translate into moment matching.

Play with **Gibbs (variational) free energy**

$$F_V(b) = \mathrm{KL}(b(\boldsymbol{x})||p(\boldsymbol{x};\theta)) - \log Z(\theta)$$

with trial $b(\boldsymbol{x})$.

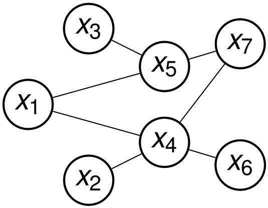
WHAT IS THE STATE OF x?

A TOY EXAMPLE

Assume that we are interested into the state of node i in an MRF, it can be answered by

- the probability $p(x_i)$, or
- an empirical version, a collection of samples $\{x_i^n\}_{n=1}^N$

It is similar for the case when $\boldsymbol{x}$ is of interests, instead of x_i.

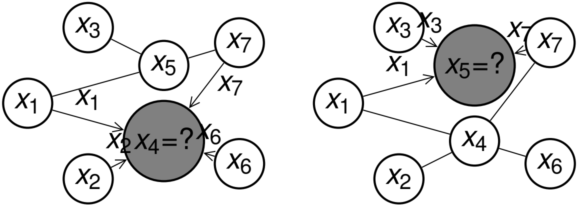


what is the state of x_4

WHAT IS THE STATE OF x?

GIBBS SAMPLING: LET US GUESS BY SAMPLING

We can approximately sample iteratively:
$x_i \sim p(x_i | \boldsymbol{x}_{-i}) \sim p(x_i, \boldsymbol{x}_{-i})$



This coordinate-wise sampling algorithm is called Gibbs sampling, which answers queries by collected samples $\{\boldsymbol{x}^n\}_1^N$.

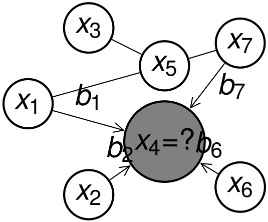
Gibbs sampling is named after the physicist Josiah Willard Gibbs, which was described by brothers Stuart and Donald Geman in 1984, some eight decades after the death of Gibbs.

$$x_i \sim \exp\{\sum_{a \in \mathrm{ne}_i} \log \varphi_a(x_i, \boldsymbol{x}_{a-i}; \boldsymbol{\theta}_a)\}$$

where ne_i gives the neighboring potential factors of node i.

WHAT IS THE STATE OF x?

Naive Mean Field: **message in form of sample values $\rightarrow$ message in form of belief**



Corresponding to minimization of **variational free energy** $F_v(b)$ **with trial** b **in fully-factorized form for univariant** $\{b_i\}$.

Iterative sampling $\rightarrow$ iterative belief update via

$$\log b_i(x_i) \propto \sum_{a \in \mathrm{ne}_i} \sum_{\boldsymbol{x}_a \setminus x_i} \prod_{j \in a \setminus i} b_j(x_j) \log \varphi_a(\boldsymbol{x}_a; \boldsymbol{\theta}_a).$$

What is the state of x?

Belief propagation (BP): let us guess by propagating belief

Proposed by Pearl (1982) for Bayesian networks (tree-structured graphs), which then widely used for general graphs (loopy BP).

Yedidia, et al, connected the loopy BP with stationary points of **Bethe free energy**

$$F_{Bethe}(b) = \sum_{a\in\mathcal{F}} \sum_{\boldsymbol{x}_a} b_a(\boldsymbol{x}_a) \log \frac{b_a(\boldsymbol{x}_a)}{\varphi_a(\boldsymbol{x}_a)} - \sum_{i=1}^{N} (|\mathrm{ne}_i| - 1) \sum_{x_i} b_i(x_i) \log b_i(x$$

Corresponding to minimization of approximated **variational free energy $F_v(b)$ with trial b includes $\{b_i\}$ and $\{b_a\}$.**

$$\text{msg : factor to variable } m_{a\to i}(x_i) \propto \sum_{\boldsymbol{x}_a\setminus x_i} \varphi_a(\boldsymbol{x}_a) \prod_{j\in a\setminus i} m_{j\to a}(x_j),$$

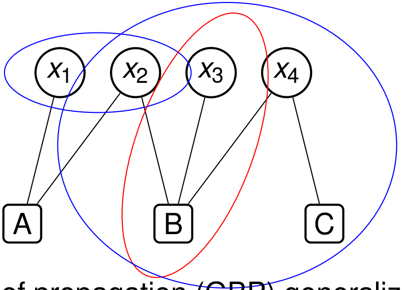
$$\text{msg : variable to factor } m_{j\to a}(x_j) \propto \prod_{a'\in \mathrm{ne}_j\setminus a} m_{a'\to j}(x_j)$$

See, D. Liu, M. T. Vu, Z. Li, and Lars K. Rasmussen. α belief propagation for approximate inference. 2020
D. Liu, N. N. Moghadam, L. K. Rasmussen, etc. α belief propagation as fully factorized approximation. In GlobalSIP, 2019. for alternative view to loopy BP.

WHAT IS THE STATE OF x?

YEDIDIA, FREEMAN, WEISS: A STEP TO GENERALIZATION

Message among variables & factors $\rightarrow$ message among regions



Generalized belief propagation (GBP) generalizes loopy BP

- usual better approximation than LBP
- higher complexity
- sensitive to scheduling of region messages

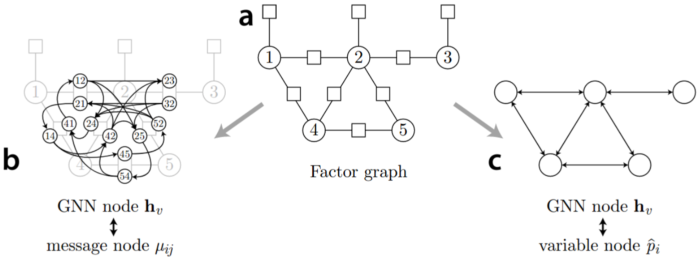
Corresponding to minimization of approximated **variational free energy $F_v(b)$ with trial b including $\{b_R\}$.**

A *region* R is a set V_R of variables nodes and a set A_R of factor nodes, such that if a factor node 'a' belongs to A_R, all the variables nodes neighboring a are in V_R.

Attempts with neural networks: an imitation game of message passing, or trials under free energy?

LEARN THE MESSAGE UPDATE RULE BY NN

An end-to-end learning process: Factor graph $\rightarrow$ converted graph representation $\rightarrow$ GNN $\rightarrow$ Output



- sum-product update rule (in BP) $\rightarrow$ NN, to learn
- pseudo probability (belief aggregation) $\rightarrow$ NN, to learn
- end-to-end learning that requires true marginal probability, which BP, GPB and mean field do no require

For related methods, see:
Heess et al, Learning to Pass Expectation Propagation Messages
Yoon, et al. 2019, Inference in Probabilistic Graphical Models by Graph Neural Networks

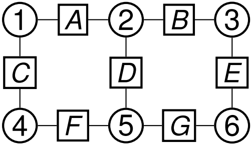
RENN

Region revisited

- If you cannot collect true targets ($p(x_i)$)
- If you are unwilling to be restricted to pre-defined inference

Factor graph representation of MRF (2-by-3 grid) with factor nodes.

MRF → region graph:



An alternative region graph of the same MRF:

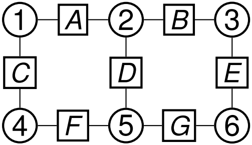
RENN

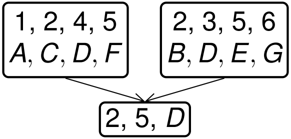
Region revisited

- If you cannot collect true targets ($p(x_i)$)
- If you are unwilling to be restricted to pre-defined inference

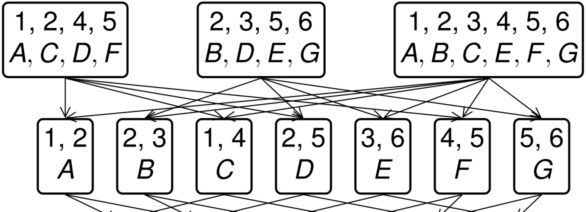
Factor graph representation of MRF (2-by-3 grid) with factor nodes.

MRF $\rightarrow$ region graph:





An alternative region graph of the same MRF:



RENN

The region-based free energy of a region graph is

$$F_R(\mathcal{B};\boldsymbol{\theta}) = \sum_{R\in\mathcal{R}} \underbrace{c_R}_{counting\ number} \sum_{\boldsymbol{x}_R} b_R(\boldsymbol{x}_R)(\underbrace{E_R(\boldsymbol{x}_R;\boldsymbol{\theta}_R)}_{region\ average\ energy} + \ln b_R(\boldsymbol{x}_R)$$

- counting number: balance the contribution of each region
- region average energy: $-\sum_{a\in A_R} \ln \varphi_a(\boldsymbol{x}_a;\boldsymbol{\theta}_a)$

Denote

by

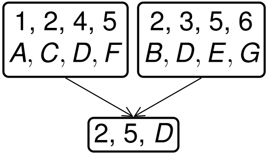
Amortizing root be-liefs

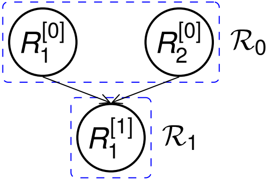
RENN

The region-based free energy of a region graph is

$$F_R(\mathcal{B};\boldsymbol{\theta}) = \sum_{R\in\mathcal{R}} \underbrace{c_R}_{counting\ number} \sum_{\boldsymbol{x}_R} b_R(\boldsymbol{x}_R)(\underbrace{E_R(\boldsymbol{x}_R;\boldsymbol{\theta}_R)}_{region\ average\ energy} + \ln b_R(\boldsymbol{x}_R)$$

- counting number: balance the contribution of each region
- region average energy: $-\sum_{a\in A_R} \ln \varphi_a(\boldsymbol{x}_a;\boldsymbol{\theta}_a)$

Denote 

by 

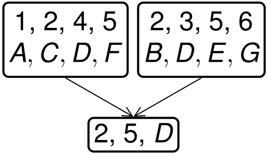
RENN

The region-based free energy of a region graph is

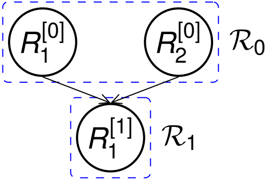
$$F_R(\mathcal{B};\boldsymbol{\theta}) = \sum_{R\in\mathcal{R}} \underbrace{c_R}_{counting\ number} \sum_{\boldsymbol{x}_R} b_R(\boldsymbol{x}_R)(\underbrace{E_R(\boldsymbol{x}_R;\boldsymbol{\theta}_R)}_{region\ average\ energy} + \ln b_R(\boldsymbol{x}_R)$$

- counting number: balance the contribution of each region
- region average energy: $-\sum_{a\in A_R} \ln \varphi_a(\boldsymbol{x}_a;\boldsymbol{\theta}_a)$

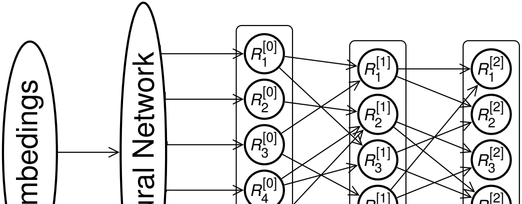
Denote



by



Amortizing
root
be-
liefs:



RENN

Objective of RENN[1]:

$$\min \textbf{region-based free energy}(F_R) + \underbrace{\textbf{panelty on belief consistency}}_{\textit{along region graph struture}}$$

Inference only

MRF → [Construct Region Graph → Optimization] → Read out beliefs

RENN

Learning alternatives of MRFs

learn with customized optm.

$$\frac{\partial \log p(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_a} = \frac{\partial \log \varphi_a(\boldsymbol{x}_a;\boldsymbol{\theta}_a)}{\partial \boldsymbol{\theta}_a} - \mathbb{E}_{p(\boldsymbol{x}_a;\boldsymbol{\theta})}\left[\frac{\partial \log \varphi_a(\boldsymbol{x}_a;\boldsymbol{\theta}_a)}{\partial \boldsymbol{\theta}_a}\right].$$

learn with auto-grads

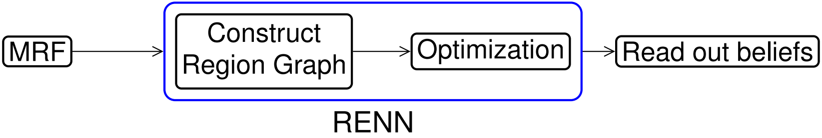
$$\max_{\boldsymbol{\theta}} \log p(\boldsymbol{x};\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}} \sum_a \log \psi \ldots - \log Z$$

RENN

Objective of RENN[1]:

$$\min \textbf{region-based free energy}(F_R) + \underbrace{\textbf{panelty on belief consistency}}_{\textit{along region graph struture}}$$

Inference only



Learning alternatives of MRFs

learn with customized optm.

$$\frac{\partial \log p(\boldsymbol{x};\theta)}{\partial \theta_a} = \frac{\partial \log \varphi_a(\boldsymbol{x}_a;\theta_a)}{\partial \theta_a} - \mathbb{E}_{p(\boldsymbol{x}_a;\theta)}\left[\frac{\partial \log \varphi_a(\boldsymbol{x}_a;\theta_a)}{\partial \theta_a}\right].$$

learn with auto-grads

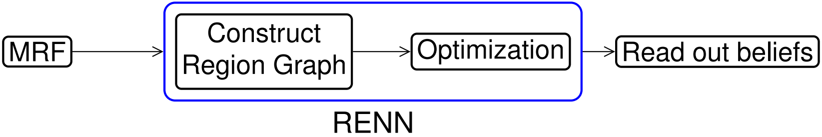
$$\max_{\theta} \log p(\boldsymbol{x};\theta) = \max_{\theta} \sum_a \log \psi \ldots - \log Z$$

RENN

Objective of RENN[1]:

$$\min \textbf{region-based free energy}(F_R) + \underbrace{\textbf{panelty on belief consistency}}_{\textit{along region graph struture}}$$

Inference only



Learning alternatives of MRFs

learn with customized optm.

$$\frac{\partial \log p(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_a} = \frac{\partial \log \varphi_a(\boldsymbol{x}_a;\boldsymbol{\theta}_a)}{\partial \boldsymbol{\theta}_a} - \mathbb{E}_{p(\boldsymbol{x}_a;\boldsymbol{\theta})}\left[\frac{\partial \log \varphi_a(\boldsymbol{x}_a;\boldsymbol{\theta}_a)}{\partial \boldsymbol{\theta}_a}\right].$$

learn with auto-grads

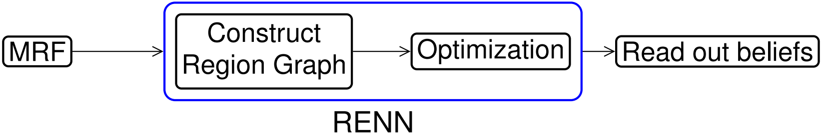
$$\max_{\boldsymbol{\theta}} \log p(\boldsymbol{x};\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}} \sum_a \log \psi \quad - \log Z$$

RENN

Objective of RENN[1]:

$$\min \textbf{region-based free energy}(F_R) + \underbrace{\textbf{panelty on belief consistency}}_{\textit{along region graph struture}}$$

Inference only



Learning alternatives of MRFs

learn with customized optm.

$$\frac{\partial \log p(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_a} = \frac{\partial \log \varphi_a(\boldsymbol{x}_a;\boldsymbol{\theta}_a)}{\partial \boldsymbol{\theta}_a} - \mathbb{E}_{p(\boldsymbol{x}_a;\boldsymbol{\theta})}\left[\frac{\partial \log \varphi_a(\boldsymbol{x}_a;\boldsymbol{\theta}_a)}{\partial \boldsymbol{\theta}_a}\right].$$

learn with auto-grads

$$\max_{\boldsymbol{\theta}} \log p(\boldsymbol{x};\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}} \sum_a \log \psi \ldots - \log Z \ldots$$

INFERENCE RESULTS

Ising model: $p(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\sum_{(i,j)\in\mathcal{E}_F} J_{ij}x_ix_j + \sum_{i\in\mathcal{V}} h_ix_i\right)$, $\boldsymbol{x} \in \{-1,1\}^N$,

- J_{ij} is the pairwise log-potential between node i and j, $J_{ij} \sim \mathcal{N}(0,1)$
- h_i is the node log-potential for node i, $h_i \sim \mathcal{N}(0,\gamma^2)$

Inference on grid graph ($\gamma = 0.1$).

Metric	n	Mean Field	Loopy BP	Damped BP	GBP	Inference Net	RENN
ℓ_1 error	25	0.271 ± 0.051	0.086 ± 0.078	0.084 ± 0.076	0.057 ± 0.024	0.111 ± 0.072	$\mathbf{0.049} \pm 0.078$
	100	0.283 ± 0.024	0.085 ± 0.041	0.062 ± 0.024	0.064 ± 0.019	0.074 ± 0.034	$\mathbf{0.025} \pm 0.011$
	225	0.284 ± 0.019	0.100 ± 0.025	0.076 ± 0.025	0.073 ± 0.013	0.073 ± 0.012	$\mathbf{0.046} \pm 0.011$
	400	0.279 ± 0.014	0.110 ± 0.016	0.090 ± 0.016	0.079 ± 0.009	0.083 ± 0.009	$\mathbf{0.061} \pm 0.009$
Correlation ρ	25	0.633 ± 0.197	0.903 ± 0.114	0.905 ± 0.113	0.923 ± 0.045	0.866 ± 0.117	$\mathbf{0.951} \pm 0.112$
	100	0.582 ± 0.112	0.827 ± 0.134	0.902 ± 0.059	0.899 ± 0.043	0.903 ± 0.049	$\mathbf{0.983} \pm 0.012$
	225	0.580 ± 0.080	0.801 ± 0.078	0.863 ± 0.088	0.869 ± 0.037	0.873 ± 0.037	$\mathbf{0.949} \pm 0.022$
	400	0.596 ± 0.054	0.779 ± 0.059	0.822 ± 0.047	0.852 ± 0.024	0.841 ± 0.028	$\mathbf{0.912} \pm 0.025$
$\log Z$ error	25	2.512 ± 1.060	0.549 ± 0.373	0.557 ± 0.369	$\mathbf{0.169} \pm 0.142$	0.762 ± 0.439	0.240 ± 0.140
	100	13.09 ± 2.156	1.650 ± 1.414	1.457 ± 1.365	$\mathbf{0.524} \pm 0.313$	2.836 ± 2.158	1.899 ± 0.495
	225	29.93 ± 4.679	3.348 ± 1.954	3.423 ± 2.157	$\mathbf{1.008} \pm 0.653$	3.249 ± 2.058	4.344 ± 0.813
	400	51.81 ± 4.706	5.738 ± 2.107	5.873 ± 2.211	$\mathbf{1.750} \pm 0.869$	3.953 ± 2.558	7.598 ± 1.146

- ℓ_1 error of beliefs v.s. true
- correlation ρ between true and approximate marginals,
- $\log Z$ error, true v.s. free energy approximation

LEARNING MRFS

What is θ in $p(\boldsymbol{x};\theta)$?

Table of negative log-likelihood of learned MRFs

n	True	Exact	Mean Field	Loopy BP	Damped BP	GBP	Inference Net	RENN
				Grid Graph				
25	9.000	9.004	9.811	9.139	9.196	10.56	9.252	**9.048**
100	19.34	19.38	23.48	19.92	20.02	28.61	20. 29	**19.76**
225	63.90	63.97	69.01	66.44	66.25	92.62	68.15	**64.79**
				Complete Graph				
9	3.276	3.286	9.558	5.201	5.880	10.06	5.262	**3.414**
16	4.883	4.934	28.74	13.64	18.95	24.45	13.77	**5.178**

learning

SUMMARY

- Brief on probabilistic graphic models
- Overview of inference methods
- A focus on the message-passing
- Transition to inference methods with NN

Thank you for your attention.
Q&A.