Perspectives on Probabilistic Graphical Models

Dong Liu

Information Science and Engineering

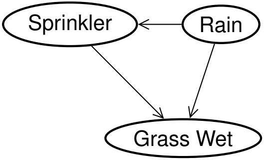
KTH - Royal Institute of Technology

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

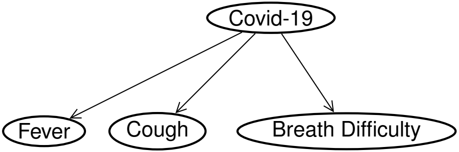
Slide is available at: https://github.com/FirstHandScientist/phdthesis

Why are Probabilistic Graphical Models interested?

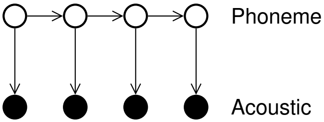
DIRECTED GRAPH REPRESENTATION



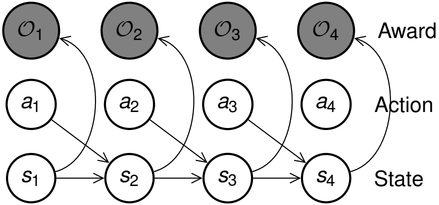
Is the sprinkler working?



Is the person get contiguous by COVID?

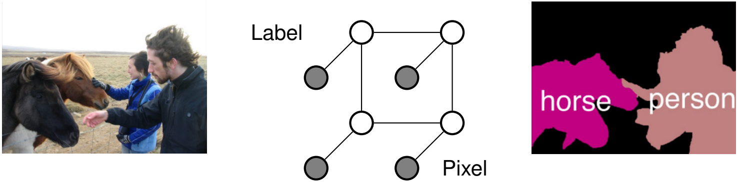


Speech Recognition

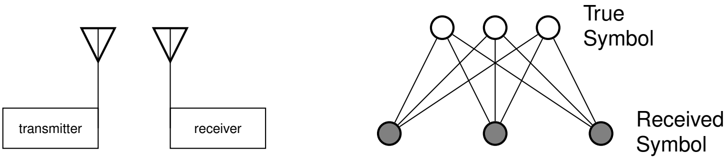


Control, reinforcement learning

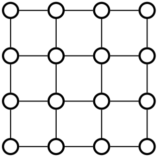
Undirected Graph Representations



Vision Perception



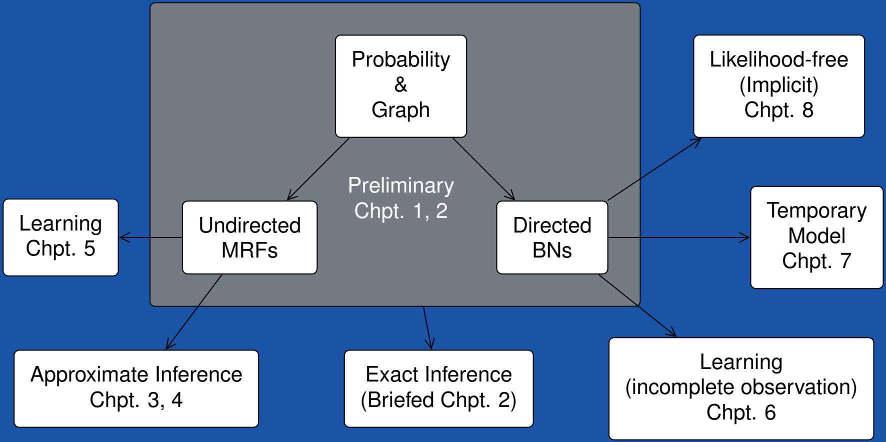
Digital communication



Physics (Ising or Potts model)

- Error-control codes
- Computational biology
- Natural language processing
- etc.

A Guide to This Dissertation



WHAT ARE PROBABILISTIC GRAPHICAL MODELS

Informally...

- attributes of our interests in a system → variable nodes
- relationship of these factors → structures of a graph

Intrinsic property: **reasoning with uncertainty**

A directed/undirected graph encoding dependencies/indepedencies of distribution $p(\boldsymbol{x};\boldsymbol{\theta})$:

- A BN/Generative model is a directed graph
  - $p(\boldsymbol{x};\boldsymbol{\theta}) = \prod_{n=1}^{N} p(x_n|\mathcal{P}(x_n))$
  - $\mathcal{P}(\cdot)$ are parent nodes
  - the local functions are proper distributions
- An MRF denoted by an undirected graph $\mathcal{G}(\mathcal{V},\mathcal{E})$
  - The probability distribution (Gibbs distribution) is $p(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \prod_{a\in\mathcal{I}} \psi_a(\boldsymbol{x}_a;\boldsymbol{\theta}_a)$
  - a indexes potential functions $\mathcal{I} = \{\psi_A, \psi_B, \cdots, \psi_M\}$
  - $Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} \prod_a \psi_a(\boldsymbol{x}_a;\boldsymbol{\theta}_a)$.

What are Probabilistic Graphical Models

Informally...

- attributes of our interests in a system $\rightarrow$ variable nodes
- relationship of these factors $\rightarrow$ structures of a graph

Intrinsic property: **reasoning with uncertainty**

A directed/undirected graph encoding dependencies/indepedencies of distribution $p(\boldsymbol{x};\boldsymbol{\theta})$:

- A BN/Generative model is a directed graph
  - $p(\boldsymbol{x};\boldsymbol{\theta}) = \prod_{n=1}^{N} p(x_n|\mathcal{P}(x_n))$
  - $\mathcal{P}(\cdot)$ are parent nodes
  - the local functions are proper distributions
- An MRF denoted by an undirected graph $\mathcal{G}(\mathcal{V},\mathcal{E})$
  - The probability distribution (Gibbs distribution) is $p(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})}\prod_{a\in\mathcal{I}}\psi_a(\boldsymbol{x}_a;\boldsymbol{\theta}_a)$
  - a indexes potential functions $\mathcal{I} = \{\psi_A, \psi_B, \cdots, \psi_M\}$
  - $Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}}\prod_a \psi_a(\boldsymbol{x}_a;\boldsymbol{\theta}_a)$.

USAGE OF GRAPHICAL MODELS

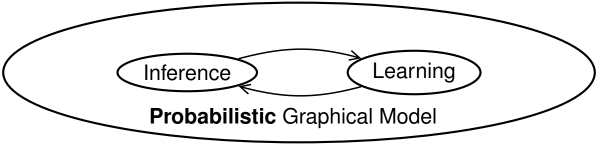
- The common inference problems:
  - Computing the likelihood of observed data.
  - Computing the marginals distribution $p(\boldsymbol{x}_A)$ over particular subset $A \subset \mathcal{V}$ of nodes
  - Computing the conditional distribution $p(\boldsymbol{x}_A|\boldsymbol{x}_B)$,
  - Computing the partition function or the Helmholtz free energy (for MRFs)
- Learning:
  - To model or determine $p(\boldsymbol{x}; \boldsymbol{\theta})$.

Two key components interacting with each other:

Usage of Graphical Models

- The common inference problems:
  - 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 partition function or the Helmholtz free energy (for MRFs)
- Learning:
  - To model or determine $p(\boldsymbol{x}; \boldsymbol{\theta})$.

Two key components interacting with each other:



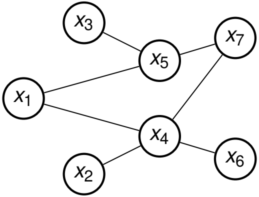
WHAT IS THE STATE OF x?

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.

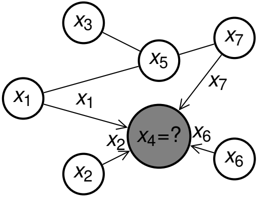


what is the state of x_4

WHAT IS THE STATE OF x?

Gibbs sampling: let us guess by sampling

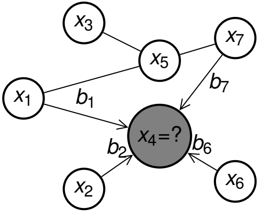
Sample iteratively:
$x_i \sim p(x_i|\boldsymbol{x}_{-i}) \sim p(x_i, \boldsymbol{x}_{-i})$



Queries by collected samples $\{\boldsymbol{x}^n\}_1^N$.

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

Propagating beliefs iteratively



Queries by collected samples $\{b_i\}$.

Intuition from *Gibbs (variational) free energy*

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

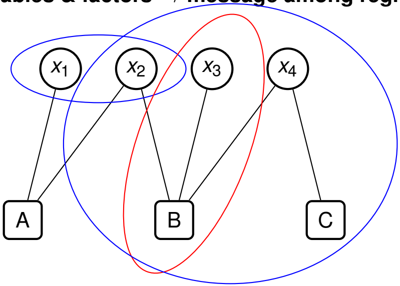
with trial $b(\boldsymbol{x})$. Instance: Bethe free energy.

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

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** $\{b_R\}$.

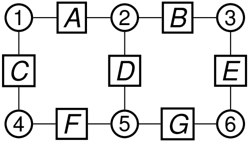
A *region* R is a set V_R of variables nodes and a set A_R of factor nodes, such that if a factor node 'a' belongs to A_R, all the variables nodes neighboring a are in V_R.

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:



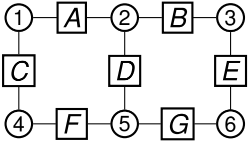
An alternative region graph of the same MRF:

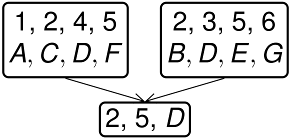
RENN

Region revisited

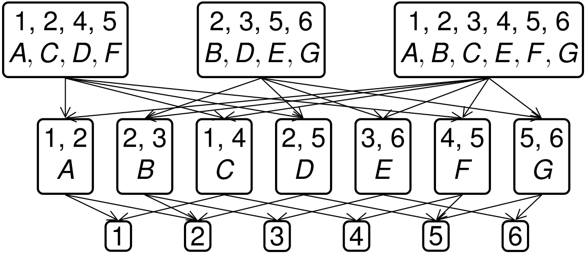
- If you cannot collect true targets ($p(x_i)$)
- If you are unwilling to be restricted to pre-defined inference

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 beliefs:

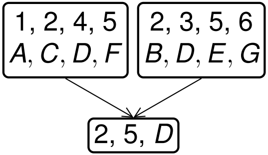
RENN

The region-based free energy of a region graph is

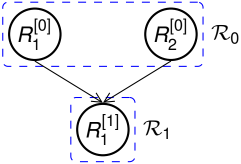
$$F_R(\mathcal{B};\boldsymbol{\theta}) = \sum_{R\in\mathcal{R}} \underbrace{c_R}_{\text{counting number}} \sum_{\boldsymbol{x}_R} b_R(\boldsymbol{x}_R)\Big(\underbrace{E_R(\boldsymbol{x}_R;\boldsymbol{\theta}_R)}_{\text{region average energy}} + \ln b_R(\boldsymbol{x}_R)\Big),$$

- 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 beliefs:

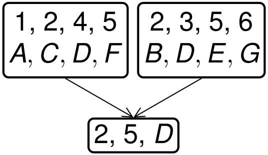
RENN

The region-based free energy of a region graph is

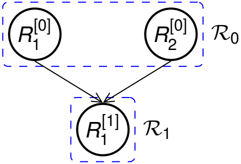
$$F_R(\mathcal{B};\boldsymbol{\theta}) = \sum_{R\in\mathcal{R}} \underbrace{c_R}_{\textit{counting number}} \sum_{\boldsymbol{x}_R} b_R(\boldsymbol{x}_R)(\underbrace{E_R(\boldsymbol{x}_R;\boldsymbol{\theta}_R)}_{\textit{region average energy}} + \ln b_R(\boldsymbol{x}_R)),$$

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

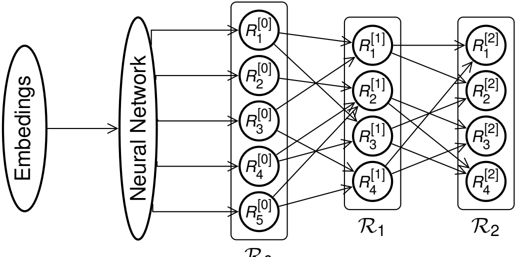
Denote



by



Amortizing root beliefs:



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.

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

learn with auto-grads

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

by $-\log Z(\boldsymbol{\theta}) \simeq F_R$.

[1] More detail on RENN? Refer to, Dong Liu, Ragnar Thobaben, and Lars K. Rasmussen. Region-based energy neural network for approximate inference. arxiv, 2020.

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.

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

learn with auto-grads

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

by $-\log Z(\boldsymbol{\theta}) \simeq F_R$.

[1] More detail on RENN? Refer to, Dong Liu, Ragnar Thobaben, and Lars K. Rasmussen. Region-based energy neural network for approximate inference. arxiv, 2020.

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.

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

learn with auto-grads

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

by $-\log Z(\boldsymbol{\theta}) \simeq F_R$.

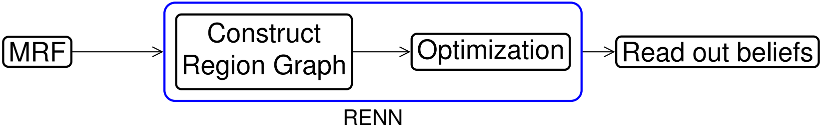
[1] More detail on RENN? Refer to, Dong Liu, Ragnar Thobaben, and Lars K. Rasmussen. Region-based energy neural network for approximate inference. arxiv, 2020.

RENN

Objective of RENN[1]:

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

Inference only



Learning alternatives of MRFs

learn with customized optm.

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

learn with auto-grads

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

by $-\log Z(\boldsymbol{\theta}) \simeq F_R$.

[1] More detail on RENN? Refer to, Dong Liu, Ragnar Thobaben, and Lars K. Rasmussen. Region-based energy neural network for approximate inference. arxiv, 2020.

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

Inference Net: Wiseman, Kim, Amortized Bethe Free Energy Minimization for Learning MRFs, 2019.

LEARNING MRFS

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

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**

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:

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

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:

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

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.

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.