Lecture 7: Probabilistic Graphical Models

André Martins



Deep Structured Learning Course, Fall 2020

Lecture 7: Probabilistic Graphical Models



- Homework 2 is due today!
- Project midterm report is due next week!
- Homework 3 is out, the deadline is December 9. Start early!



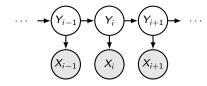
• Vlad Niculae (co-instructor of DSL last year)

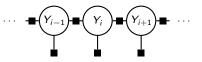
Graphical Models

In this unit, we will formalize & extend these graphical representations encountered in previous lectures.

Directed

Undirected





Outline

Directed Models

Bayes networks Conditional independence and D-separation Causal graphs & the *do* operator

Output: Description of the second second

Markov random fields

Factor graphs

Outline

1 Directed Models

Bayes networks

Conditional independence and D-separation Causal graphs & the *do* operator

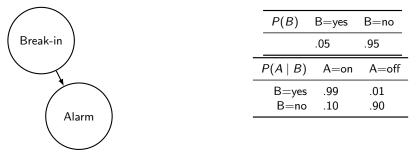
Output: Description of the second second

Markov random fields

Factor graphs

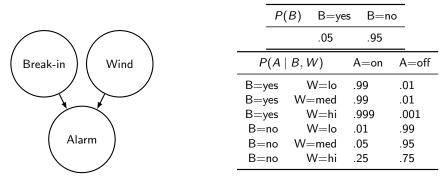
- Common task: Characterize how some related events co-occur. Specifically, in terms of probabilities!
- A car alarm is going off. Was there a break-in?

- Common task: Characterize how some related events co-occur. Specifically, in terms of probabilities!
- A car alarm is going off. Was there a break-in?



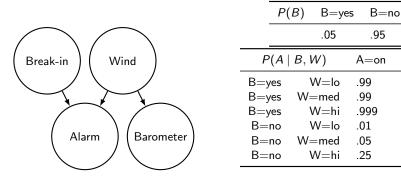
• $\mathbb{P}(B \mid A) = ?$

- Common task: Characterize how some related events co-occur. Specifically, in terms of probabilities!
- A car alarm is going off. Was there a break-in?



• $\mathbb{P}(B \mid A) =$? Can we observe wind? $\mathbb{P}(B \mid A, W) =$?

- Common task: Characterize how some related events co-occur. Specifically, in terms of probabilities!
- A car alarm is going off. Was there a break-in?



P(B | A) =? Can we observe wind? P(B | A, W) =?
 Maybe we're in the basement, but have a barometer.

André Martins (IST)

Lecture 7: Probabilistic Graphical Models

A=off

.01

.01

.001

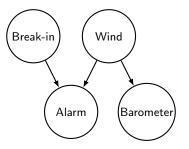
.99

.95

.75

Bayes networks

Toolkit for encoding knowledge about interaction structures between rv's.



Directed acyclic graph (DAG). Nodes = variables. Arrows = statistical dependencies.

In general:
$$\mathbb{P}(X_1, \dots, X_n) = \prod_i \mathbb{P}(X_i \mid \mathsf{parents}(X_i))$$

For example: $\mathbb{P}(Break-in, Wind, Alarm, Barometer)$ = $\mathbb{P}(Break-in)\mathbb{P}(Wind)\mathbb{P}(Alarm | Break-in, Wind)\mathbb{P}(Barometer | Wind)$

IST, Fall 2020 8 / 66

Brk.	Wind	Alarm	Bar.	Р		Brk.	Brk. Wind	Brk. Wind Alarm	Brk. Wind Alarm Bar.
yes	lo	on	lo	0.0243		no	no lo	no lo on	no lo on lo
yes	lo	on	med	0.0002		no	no lo	no lo on	no lo on med
yes	lo	on	hi	0.0002	r	10	no lo	no lo on	no lo on hi
yes	lo	off	lo	0.0002	nc	c	o lo	o lo off	o lo off lo
yes	lo	off	med	2.50e-06	no		lo	lo off	lo off med
yes	lo	off	hi	2.50e-06	no		lo	lo off	lo off hi
yes	med	on	lo	0.0001	no		med	med on	med on lo
yes	med	on	med	0.0146	no		med	med on	med on med
yes	med	on	hi	0.0001	no	1	med	med on	ned on hi
yes	med	off	lo	1.50e-06	no	m	ed	ed off	ed off lo
yes	med	off	med	0.0001	no	me	d	d off	
yes	med	off	hi	1.50e-06	no	med		off	
yes	hi	on	lo	9.99e-05	no	hi		on	on lo
yes	hi	on	med	9.99e-05	no	hi		on	
yes	hi	on	hi	0.0098	no	hi		on	
yes	hi	off	lo	1.00e-07	no	hi		off	
yes	hi	off	med	1.00e-07	no	hi		off	
yes	hi	off	hi	9.80e-06	no	hi		off	off hi

Without any structure, $\mathbb{P}(Break-in, Wind, Alarm, Barometer)$ would have to be stored & estimated like

k .	Wind	Alarm	Bar.	Р	•		Brk.	Brk. Wind	Brk. Wind Alarm	Brk. Wind Alarm Bar.
	lo	on	lo	0.0243			no	no lo	no lo on	no lo on lo
	lo	on	med	0.0002			no	no lo	no lo on	no lo on med
	lo	on	hi	0.0002			no	no lo	no lo on	no lo on hi
	lo	off	lo	0.0002			no	no lo	no lo off	no lo off lo
	lo	off	med	2.50e-06			no	no lo	no lo off	no lo off med
	lo	off	hi	2.50e-06			no	no lo	no lo off	no lo off hi
	med	on	lo	0.0001			no	no med	no med on	no med on lo
	med	on	med	0.0146			no	no med	no med on	no med on med
	med	on	hi	0.0001			no	no med	no med on	no med on hi
	med	off	lo	1.50e-06			no	no med	no med off	no med off lo
	med	off	med	0.0001			no	no med	no med off	
	med	off	hi	1.50e-06		r	10			
	hi	on	lo	9.99e-05		nc)			
	hi	on	med	9.99e-05		nc)			
	hi	on	hi	0.0098		r	10			
	hi	off	lo	1.00e-07			no			
	hi	off	med	1.00e-07			no			
	hi	off	hi	9.80e-06			no	no hi	no hi off	no hi off hi

Without any structure, $\mathbb{P}(\mathsf{Break-in},\mathsf{Wind},\mathsf{Alarm},\mathsf{Barometer})$ would have to be stored & estimated like

 $\mathbb{P}(\text{Break-in}=\text{yes}, \text{Alarm}=\text{on}) = 0.0496$

-				•	-			-	-
_	Alarm	Bar.	Р		_	Brk.	Brk. Wind	Brk. Wind Alarm	Brk. Wind Alarm Bar.
	on	lo	0.0243			no	no lo	no lo on	no lo on lo
	on	med	0.0002			no	no lo	no lo on	no lo on med
	on	hi	0.0002			no	no lo	no lo on	no lo on hi
	off	lo	0.0002			no	no lo	no lo off	no lo off lo
	off	med	2.50e-06			no	no lo	no lo off	no lo off med
	off	hi	2.50e-06			no	no lo	no lo off	no lo off hi
	on	lo	0.0001			no	no med	no med on	no med on lo
	on	med	0.0146			no	no med	no med on	no med on med
	on	hi	0.0001			no	no med	no med on	no med on hi
	off	lo	1.50e-06			no	no med	no med off	no med off lo
	off	med	0.0001			no	no med	no med off	no med off med
	off	hi	1.50e-06			no	no med	no med off	no med off hi
	on	lo	9.99e-05			no	no hi	no hi on	no hi on lo
	on	med	9.99e-05			no	no hi	no hi on	no hi on med
	on	hi	0.0098			no	no hi	no hi on	no hi on hi
	off	lo	1.00e-07			no	no hi	no hi off	no hi off lo
	off	med	1.00e-07			no	no hi	no hi off	no hi off med
	off	hi	9.80e-06			no	no hi	no hi off	no hi off hi

Without any structure, $\mathbb{P}(Break-in, Wind, Alarm, Barometer)$ would have to be stored & estimated like

 $\mathbb{P}(\mathsf{Break-in}{=}\mathsf{yes},\mathsf{Alarm}{=}\mathsf{on}) = 0.0496$

 $\mathbb{P}(\text{Break-in}=no, \text{Alarm}=on) = 0.0665$

Brk.	Wind	Alarm	Bar.	P		Brk.	Brk. Wind	Brk. Wind Alarm	Brk. Wind Alarm Bar.
			1.	0.0242			Ia		no lo on lo
es	lo	on	lo	0.0243		no			
yes	lo	on	med	0.0002		no			
yes	lo	on	hi	0.0002	n	0			
yes	lo	off	lo	0.0002	no		lo	lo off	lo off lo
yes	lo	off	med	2.50e-06	no		lo	lo off	lo off med
yes	lo	off	hi	2.50e-06	no		lo	lo off	lo off hi
yes	med	on	lo	0.0001	no	1	med	med on	med on lo
yes	med	on	med	0.0146	no	n	ned	ned on	ned on med
yes	med	on	hi	0.0001	no	m	ed	ed on	ed on hi
yes	med	off	lo	1.50e-06	no	me	d	d off	d off lo
yes	med	off	med	0.0001	no	med		off	off med
yes	med	off	hi	1.50e-06	no	med		off	off hi
yes	hi	on	lo	9.99e-05	no	hi		on	on lo
yes	hi	on	med	9.99e-05	no	hi		on	on med
yes	hi	on	hi	0.0098	no	hi		on	on hi
yes	hi	off	lo	1.00e-07	no	hi		off	off lo
yes	hi	off	med	1.00e-07	no	hi		off	off med
yes	hi	off	hi	9.80e-06	no	hi		off	off hi

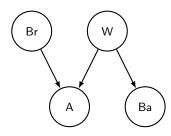
Without any structure, $\mathbb{P}(Break-in, Wind, Alarm, Barometer)$ would have to be stored & estimated like

 $\mathbb{P}(\text{Break-in=yes}, \text{Alarm=on}) = 0.0496$ $\mathbb{P}(\text{Break-in=no}, \text{Alarm=on}) = 0.0665$

$$\mathbb{P}(\text{Break-in=yes} \mid \text{Alarm=on}) = \frac{\mathbb{P}(\text{Break-in=yes}, \text{Alarm=on})}{\sum_{b} \mathbb{P}(\text{Break-in=}b, \text{Alarm=on})}$$

= .427

Knowing the model structure (statistical dependencies), complicated models become manageable.



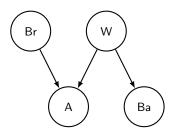
$$\begin{split} \mathbb{P}(\mathsf{Br},\,\mathsf{W},\,\mathsf{A},\,\mathsf{Ba}) \\ = \mathbb{P}(\mathsf{Br})\mathbb{P}(\mathsf{W})\mathbb{P}(\mathsf{A}\mid\mathsf{Br},\,\mathsf{W})\mathbb{P}(\mathsf{Ba}\mid\mathsf{W}) \end{split}$$

	_					_		
	P((Br)	ye	s	no			
			.0	5	.95			
•	P(W	/)	lo	mie	d	hi	-	
-			.5	.3		.2	_	
								_
P	$P(A \mid E)$	Br, W	′)		on		off	_
Br=	yes	W	/=lo		.99		.01	
Br=	yes	W =	med		.99		.01	
Br=	yes	W	/=hi		.999		.001	
Br=	=no	W	/=lo		.01		.99	
Br=	=no	W =	med		.05		.95	
Br=	=no	W	/=hi		.25		.75	
								-
Р(Ba N	N)	lo		mid		hi	
	W=	=lo	.98		.01		.01	
	W=n	nid	.01		.98		.01	
	W=	=hi	.01	. =	.01	-	.98	5
	a da la				T F.		020	10

André Martins (IST)

Lecture 7: Probabilistic Graphical Models

Knowing the model structure (statistical dependencies), complicated models become manageable.

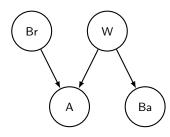


$$\begin{split} \mathbb{P}(\mathsf{Br},\,\mathsf{W},\,\mathsf{A},\,\mathsf{Ba}) \\ = \mathbb{P}(\mathsf{Br})\mathbb{P}(\mathsf{W})\mathbb{P}(\mathsf{A}\mid\mathsf{Br},\,\mathsf{W})\mathbb{P}(\mathsf{Ba}\mid\mathsf{W}) \end{split}$$

 Can estimate parts in isolation e.g. P(Wind) from weather history.

	P(Br	·) ye	s no		
		.0	5.95	5	
	P(W)	lo	mid	hi	
		.5	.3	.2	
					_
F	$P(A \mid Br,$	W)	on	off	
Br=	yes	W=lo	.99	.01	
Br=	yes W	/=med	.99	.01	
Br=	yes	W=hi	.999	.001	
Br=	=no	W=lo	.01	.99	
Br=	=no W	/=med	.05	.95	
Br=	=no	W=hi	.25	.75	
P	$(Ba \mid W)$	lo	mid	hi	_
	W=lo	.98	.01	.01	
	W=mid	.01	.98	.01	
	W=hi	.01	.01	.98	
					÷.

Knowing the model structure (statistical dependencies), complicated models become manageable.



$$\begin{split} \mathbb{P}(\mathsf{Br},\,\mathsf{W},\,\mathsf{A},\,\mathsf{Ba}) \\ = \mathbb{P}(\mathsf{Br})\mathbb{P}(\mathsf{W})\mathbb{P}(\mathsf{A}\mid\mathsf{Br},\,\mathsf{W})\mathbb{P}(\mathsf{Ba}\mid\mathsf{W}) \end{split}$$

- Can estimate parts in isolation e.g. P(Wind) from weather history.
- Can sample by following the graph from roots to leaves.

	P(B	r) ye	s no		
		.0	5.9	5	
	P(W)	lo	mid	hi	•
		.5	.3	.2	•
					·
F	$P(A \mid Br$, W)	on		off
Br=	yes	W=lo	.99		.01
Br=	=yes V	V=med	.99		.01
Br=	yes	W=hi	.99	9.	.001
Br=	=no	W=lo	.01		.99
Br=	=no V	V=med	.05		.95
Br=	=no	W=hi	.25		.75
Ρ	(Ba W) lo	mic	l ł	ni
	W=le	o .98	.01		01
	W=mi	d .01	.98		01
	W=h	i .01	.01	-= -	98 _
				_	

10/66

Bayes Nets:

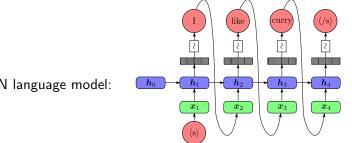
reduce number of parameters & aid estimation let us reason about **independencies** in a model are a building-block for modeling **causality**



Bayes Nets:

are not neural network diagrams encode structure, not parametrization are non-unique for a distribution encode independence **requirements**, not necessarily all

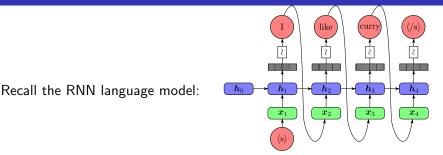
BN are not neural net diagrams



Recall the RNN language model:

In statistical terms, what are we modeling?

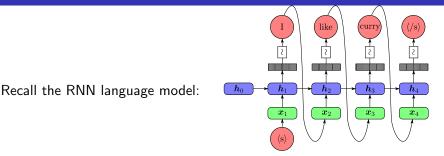
BN are not neural net diagrams



• In statistical terms, what are we modeling?

 $\mathbb{P}(X_1,\ldots,X_n)=\mathbb{P}(X_1)\mathbb{P}(X_2\mid X_1)\mathbb{P}(X_3\mid X_1,X_2)\ldots$

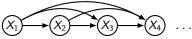
BN are not neural net diagrams



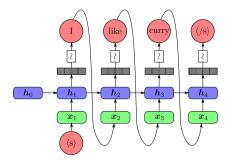
In statistical terms, what are we modeling?

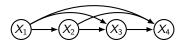
 $\mathbb{P}(X_1,\ldots,X_n)=\mathbb{P}(X_1)\mathbb{P}(X_2\mid X_1)\mathbb{P}(X_3\mid X_1,X_2)\ldots$

• Bayes Net:



• Not useful! Everything conditionally-depends on everything. (more later)





Neural net diagrams (and computation graphs) show **how to compute something** Bayes networks show **how a distribution factorizes** (what is assumed independent)

A BN tells us: **how the distribution decomposes** A BN can't tell us: **what the probabilities are!**

Example: $X \in \mathfrak{X} =$ all English sentences, $Y \in \{$ sports,music,... $\}$.

BN for a generative model:



We must posit what are $\mathbb{P}(Y)$ and $\mathbb{P}(X \mid Y)$. Many possible options!

A BN tells us: how the distribution decomposes A BN can't tell us: what the probabilities are!

Example: $X \in \mathfrak{X} =$ all English sentences, $Y \in \{ \texttt{sports}, \texttt{music}, \dots \}.$

BN for a generative model:



We must posit what are $\mathbb{P}(Y)$ and $\mathbb{P}(X \mid Y)$. Many possible options! $\mathbb{P}(Y)$: uniform: $\mathbb{P}(Y = \text{sports}) = \mathbb{P}(Y = \text{music}) = \frac{1}{|Y|}$,

A BN tells us: how the distribution decomposes A BN can't tell us: what the probabilities are!

Example: $X \in \mathfrak{X} =$ all English sentences, $Y \in \{$ sports,music,... $\}$.

BN for a generative model:



We must posit what are $\mathbb{P}(Y)$ and $\mathbb{P}(X \mid Y)$. Many possible options!

 $\mathbb{P}(Y)$: uniform: $\mathbb{P}(Y = \texttt{sports}) = \mathbb{P}(Y = \texttt{music}) = \frac{1}{|\mathcal{Y}|}$, or estimated from data.

A BN tells us: how the distribution decomposes A BN can't tell us: what the probabilities are!

Example: $X \in \mathcal{X} =$ all English sentences, $Y \in \{\text{sports}, \text{music}, \dots\}$. BN for a generative model: $(Y) \rightarrow (X)$ We must posit what are $\mathbb{P}(Y)$ and $\mathbb{P}(X \mid Y)$. Many possible options! $\mathbb{P}(Y)$: uniform: $\mathbb{P}(Y = \text{sports}) = \mathbb{P}(Y = \text{music}) = \frac{1}{|Y|}$, or estimated from data. $\mathbb{P}(X \mid Y)$ (remember: values of X are sentences)

A BN tells us: how the distribution decomposes A BN can't tell us: what the probabilities are!

Example: $X \in \mathfrak{X} =$ all English sentences, $Y \in \{ \text{sports}, \text{music}, \dots \}$. BN for a generative model: $(Y) \rightarrow (X)$ We must posit what are $\mathbb{P}(Y)$ and $\mathbb{P}(X \mid Y)$. Many possible options! $\mathbb{P}(Y)$: uniform: $\mathbb{P}(Y = \text{sports}) = \mathbb{P}(Y = \text{music}) = \frac{1}{|Y|}$, or estimated from data. $\mathbb{P}(X \mid Y)$ (remember: values of X are sentences) Naive Bayes $\mathbb{P}(X \mid Y) = \prod_{j=1}^{L} \mathbb{P}(X_j \mid Y)$

A BN tells us: how the distribution decomposes A BN can't tell us: what the probabilities are!

Example: $X \in \mathcal{X} =$ all English sentences, $Y \in \{\text{sports}, \text{music}, \dots\}$. BN for a generative model: $Y \longrightarrow X$ We must posit what are $\mathbb{P}(Y)$ and $\mathbb{P}(X \mid Y)$. Many possible options! $\mathbb{P}(Y)$: uniform: $\mathbb{P}(Y = \text{sports}) = \mathbb{P}(Y = \text{music}) = \frac{1}{|Y|}$, or estimated from data. $\mathbb{P}(X \mid Y)$ (remember: values of X are sentences) Naive Bayes $\mathbb{P}(X \mid Y) = \prod_{j=1}^{L} \mathbb{P}(X_j \mid Y)$ Per-class Markov language model $\mathbb{P}(X \mid Y) = \prod_{j=1}^{L} \mathbb{P}(X_j \mid X_{j-1}, Y)$

A BN tells us: how the distribution decomposes A BN can't tell us: what the probabilities are!

Example: $X \in \mathcal{X} =$ all English sentences, $Y \in \{\text{sports}, \text{music}, \dots\}$. BN for a generative model: $Y \longrightarrow X$ We must posit what are $\mathbb{P}(Y)$ and $\mathbb{P}(X \mid Y)$. Many possible options! $\mathbb{P}(Y)$: uniform: $\mathbb{P}(Y = \text{sports}) = \mathbb{P}(Y = \text{music}) = \frac{1}{|Y|}$, or estimated from data. $\mathbb{P}(X \mid Y)$ (remember: values of X are sentences) Naive Bayes $\mathbb{P}(X \mid Y) = \prod_{j=1}^{L} \mathbb{P}(X_j \mid Y)$ Per-class Markov language model $\mathbb{P}(X \mid Y) = \prod_{j=1}^{L} \mathbb{P}(X_j \mid X_{j-1}, Y)$ Per-class recurrent NN language model $\mathbb{P}(X \mid Y) = \text{LSTM}(x_1, \dots, x_L; w_Y)$

A BN tells us: how the distribution decomposes A BN can't tell us: what the probabilities are!

Example: $X \in \mathcal{X} =$ all English sentences, $Y \in \{\text{sports}, \text{music}, \dots\}$. BN for a generative model: $Y \longrightarrow X$ We must posit what are $\mathbb{P}(Y)$ and $\mathbb{P}(X \mid Y)$. Many possible options! $\mathbb{P}(Y)$: uniform: $\mathbb{P}(Y = \text{sports}) = \mathbb{P}(Y = \text{music}) = \frac{1}{|\mathcal{Y}|}$, or estimated from data. $\mathbb{P}(X \mid Y)$ (remember: values of X are sentences) Naive Bayes $\mathbb{P}(X \mid Y) = \prod_{j=1}^{L} \mathbb{P}(X_j \mid Y)$ Per-class Markov language model $\mathbb{P}(X \mid Y) = \prod_{j=1}^{L} \mathbb{P}(X_j \mid X_{j-1}, Y)$ Per-class recurrent NN language model $\mathbb{P}(X \mid Y) = \text{LSTM}(x_1, \dots, x_L; w_Y)$

 $\mathbb{P}(X \mid Y)$ need not be parametrized as a table.

A BN tells us: **how the distribution decomposes** A BN can't tell us: **what the probabilities are!**

 $\mathbb{P}(X \mid Y)$ need not be parametrized as a table.

rv's need not be discrete! mixture of Gaussians: $\mathbb{P}(X \mid Y = y) \sim \mathcal{N}(\mu_y, \Sigma_y)$.

André Martins (IST)

Equivalent factorizations

There are many possible factorizations! $\mathbb{P}(X, Y) =$

Equivalent factorizations

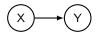
There are many possible factorizations! $\mathbb{P}(X, Y) =$



 $\mathbb{P}(X)\mathbb{P}(Y\mid X)$

Equivalent factorizations

There are many possible factorizations! $\mathbb{P}(X, Y) =$

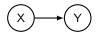


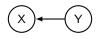


 $\mathbb{P}(X)\mathbb{P}(Y\mid X)$

 $\mathbb{P}(Y)\mathbb{P}(X\mid Y)$

There are many possible factorizations! $\mathbb{P}(X, Y) =$

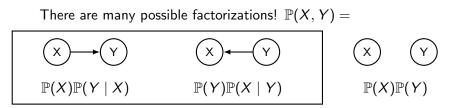




 $\mathbb{P}(X)\mathbb{P}(Y\mid X)$

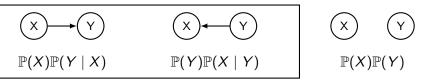
 $\mathbb{P}(Y)\mathbb{P}(X\mid Y)$





The first two are valid Bayes nets for any $\mathbb{P}(X, Y)$!

There are many possible factorizations! $\mathbb{P}(X, Y) =$



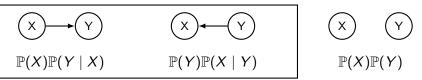
The first two are valid Bayes nets for any $\mathbb{P}(X, Y)$!

In fact, recall generative vs discriminative classifiers!

Generative (e.g. naïve Bayes): (X) ← (Y)
To classify, we would compute P(Y | X) via Bayes' rule.
Discriminative (e.g. logistic regression) (X) → (Y)

in LR, we don't model $\mathbb{P}(X)$, we assume X is always observed (gray).

There are many possible factorizations! $\mathbb{P}(X, Y) =$



The first two are valid Bayes nets for any $\mathbb{P}(X, Y)$!

In fact, recall generative vs discriminative classifiers!

• Generative (e.g. naïve Bayes):

To classify, we would compute $\mathbb{P}(Y \mid X)$ via Bayes' rule.

• Discriminative (e.g. logistic regression)

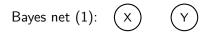
in LR, we don't model $\mathbb{P}(X)$, we assume X is always observed (gray). Some arrow direction choices are harder to estimate.

Х

Some make more sense (why?): (Barmtr.) (Wind) vs. (Barmtr.) (Wind)

Recall, we say $X \perp Y$ iff. P(X, Y) = P(X)P(Y)Let X = grade in DSL, Y = month you were born.

Recall, we say $X \perp Y$ iff. P(X, Y) = P(X)P(Y)Let X = grade in DSL, Y = month you were born.



Example parametrization:				
P(X)	A+	А	В	
	.01	.02	.04	
P(Y)	Jan	Feb	Mar	
	.10	.12	.09	

Recall, we say $X \perp Y$ iff. P(X, Y) = P(X)P(Y)Let X = grade in DSL, Y = month you were born.

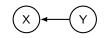
Bayes net (1):
$$(X)$$
 (Y)

Example parametrization:				
P(X)	A+	А	В	
	.01	.02	.04	
P(Y)	Jan	Feb	Mar	
	.10	.12	.09	

BN (1) imposes $X \perp Y$ in any parametrization.

Recall, we say $X \perp Y$ iff. P(X, Y) = P(X)P(Y)Let X = grade in DSL, Y = month you were born.

Bayes net (1): (x)



Example parametrization:				
P(X)	A+	А	В	
	.01	.02	.04	
P(Y)	Jan	Feb	Mar	
	.10	.12	.09	

Does it mean $X \not\perp Y$ necessarily?

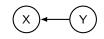
BN (1) imposes $X \perp Y$ in any parametrization.

Recall, we say $X \perp Y$ iff. P(X, Y) = P(X)P(Y)Let X = grade in DSL, Y = month you were born.

Bayes net (1): (X)

$$\bigcirc \quad \bigtriangledown$$

Bayes net (2):



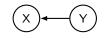
Example parametrization:				
P(X)	A+	А	В	
	.01	.02	.04	
P(Y)	Jan	Feb	Mar	
	.10	.12	.09	

BN (1) imposes $X \perp Y$ in any parametrization.

Does it mean $X \not\perp Y$ necessarily? NO! P(Y)Feb Mar Jan .10 .12 .09 $P(X \mid Y)$ 20 19 18 Y=Jan .01 .02 .04 Y=Feb 01 .02 .04 Y=Mar 01 02 .04

Recall, we say $X \perp Y$ iff. P(X, Y) = P(X)P(Y)Let X = grade in DSL, Y = month you were born.

Bayes net (1): (X)



Example parametrization:				
P(X)	A+	А	В	
	.01	.02	.04	
P(Y)	Jan	Feb	Mar	
	.10	.12	.09	

BN (1) imposes $X \perp Y$ in any parametrization.

Does it mean $X \not\perp Y$ necessarily? NO! P(Y)Feb Mar Jan .12 .10 .09 $P(X \mid Y)$ 20 19 18 Y=Jan .01 .02 .04 Y=Feb 01 .02 .04 Y=Mar 01 02 .04

A BN expresses which independences **must exist**, but there can be additional ones.

André Martins (IST)

IST, Fall 2020 17 / 66

Outline

1 Directed Models

Bayes networks

Conditional independence and D-separation

Causal graphs & the do operator

Output: Description of the second second

Markov random fields

Factor graphs

Conditional independence in Bayes nets

Identifying independences in a distribution is generally hard.

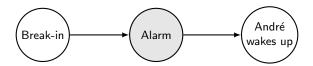
Bayes nets let us reason about it via graph algorithms!

Definition (conditional independence)

A is independent of B given a set of variables $C = \{C_1, \ldots, C_n\}$, denoted

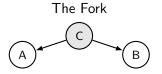
 $A \perp\!\!\!\perp B \mid C,$

iff $\mathbb{P}(A, B \mid C_1, \ldots, C_n) = \mathbb{P}(A \mid C_1, \ldots, C_n)\mathbb{P}(B \mid C_1, \ldots, C_n)$. Note. Equivalently, $\mathbb{P}(A \mid B, C_1, \ldots, C_n) = \mathbb{P}(A \mid C_1, \ldots, C_n)$. Intuitively: if we observe *C*, does observing *B* too bring us more info about *A*?



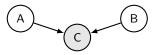
You want to assess if I'm awake. Does it matter if there really was a break-in?

André Martins (IST)



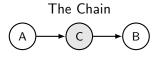
The Chain $(A \rightarrow C \rightarrow B)$

The Collider

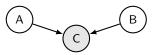


The Fork

 $A \perp B \mid C$ Given C, A and B are independent. Example: Alarm \leftarrow Wind \rightarrow Barometer



The Collider



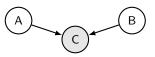
The Fork

 $A \perp \!\!\!\perp B \mid C$ Given C, A and B are independent. Example: Alarm \leftarrow Wind \rightarrow Barometer

The Chain $(A) \rightarrow (C) \rightarrow (B)$

 $\begin{array}{c} A \perp B \mid C \\ \text{After observing } C, \\ \text{further observing } A \text{ would not tell us about } B. \\ \text{Example: Burglary} \rightarrow \text{Alarm} \rightarrow \text{André wakes up} \end{array}$

The Collider



The Fork

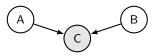
The Chain

В

 $A \perp\!\!\!\!\perp B \mid C$ Given C, A and B are independent. Example: Alarm \leftarrow Wind \rightarrow Barometer

 $A \perp \!\!\!\perp B \mid C$ After observing *C*, further observing *A* would not tell us about *B*. Example: Burglary \rightarrow Alarm \rightarrow André wakes up

The Collider



Surprisingly, $A \perp B$ but not $A \perp B \mid C$ Example:Burglary \rightarrow Alarm \leftarrow WindBurglaries occur regardless how windy it is.If alarm rings, hearing wind makes burglary less likely!Burglary is "explained away" by wind.

Definition: A and B are **d-separated** given set C if for any path P from A to B at least one holds:

Definition: A and B are **d-separated** given set C if for any path P from A to B at least one holds:

1 *P* includes a fork with observed parent:

$$X \leftarrow Z \rightarrow Y$$
 (with $Z \in C$)

Definition: A and B are **d-separated** given set C if for any path P from A to B at least one holds:

1 *P* includes a fork with observed parent:

$$X \leftarrow Z \rightarrow Y$$
 (with $Z \in C$)

2 *P* includes a chain with observed middle:

$$X \to Z \to Y$$
 or $X \leftarrow Z \leftarrow Y$ (with $Z \in C$)

Definition: A and B are **d-separated** given set C if for any path P from A to B at least one holds:

1 *P* includes a fork with observed parent:

$$X \leftarrow Z \rightarrow Y$$
 (with $Z \in C$)

2 *P* includes a chain with observed middle:

$$X \to Z \to Y$$
 or $X \leftarrow Z \leftarrow Y$ (with $Z \in C$)

3 *P* includes a collider with unobserved descendants:

 $X \rightarrow Z \leftarrow Y$ (with neither Z nor any of its descendants $\in C$)

Definition: A and B are **d-separated** given set C if for any path P from A to B at least one holds:

1 *P* includes a fork with observed parent:

$$X \leftarrow Z \rightarrow Y$$
 (with $Z \in C$)

2 *P* includes a chain with observed middle:

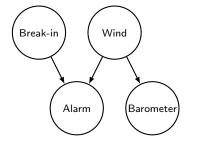
$$X \to Z \to Y$$
 or $X \leftarrow Z \leftarrow Y$ (with $Z \in C$)

3 *P* includes a collider with unobserved descendants:

 $X \rightarrow Z \leftarrow Y$ (with neither Z nor any of its descendants $\in C$)

Theorem: A and B d-separated given $C \implies A \perp B \mid C$.





Wind $\bot\!\!\!\!\bot$ Barometer?

André Martins (IST)

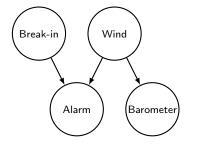
Lecture 7: Probabilistic Graphical Models

IST, Fall 2020 22 / 66

æ

<ロト < 同ト < ヨト < ヨト



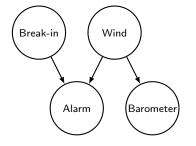


Wind \bot Barometer? No

(日)

æ

< ∃⇒



Wind ⊥ Barometer? **No** Break-in ⊥ Wind?

André Martins (IST)

Lecture 7: Probabilistic Graphical Models

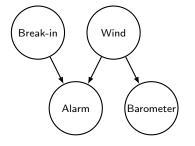
IST, Fall 2020

э

<ロト < 同ト < ヨト < ヨト

22 / 66



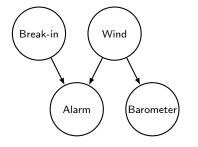


Wind ⊥ Barometer? No Break-in ⊥ Wind? Yes

э

< ∃⇒

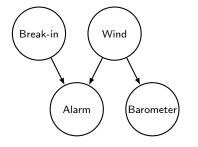
(日)



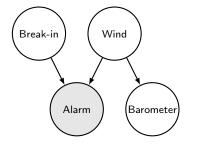
Wind ⊥ Barometer? No Break-in ⊥ Wind? Yes Break-in ⊥ Barometer?

< D > < A > < B > < B >

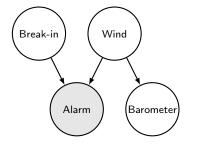
< ∃ →



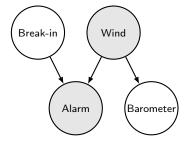
Wind \parallel Barometer? No Break-in L Wind? Yes Break-in *I* Barometer? **Yes**



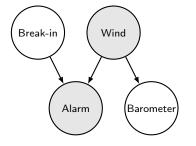
Wind \parallel Barometer? No Break-in L Wind? Yes Break-in II Barometer? Yes Break-in \bot Barometer | Alarm?



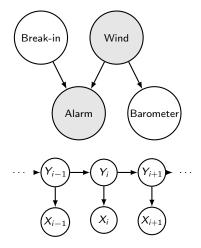
Wind \parallel Barometer? No Break-in L Wind? Yes Break-in II Barometer? Yes Break-in L Barometer | Alarm? No



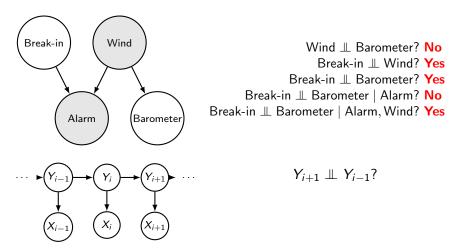
Wind ⊥ Barometer? No Break-in ⊥ Wind? Yes Break-in ⊥ Barometer? Yes Break-in ⊥ Barometer | Alarm? No Break-in ⊥ Barometer | Alarm, Wind?



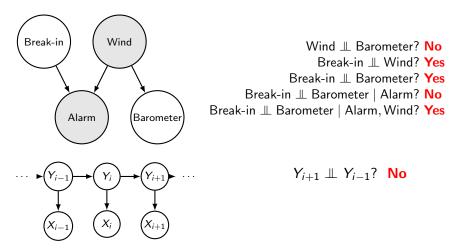
Wind ⊥ Barometer? No Break-in ⊥ Wind? Yes Break-in ⊥ Barometer? Yes Break-in ⊥ Barometer | Alarm? No Break-in ⊥ Barometer | Alarm, Wind? Yes



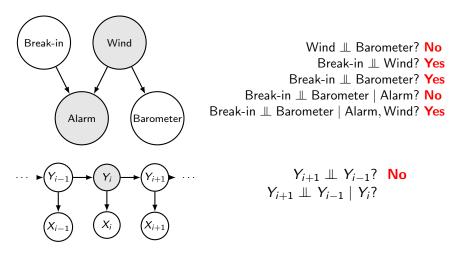
Wind ⊥ Barometer? No Break-in ⊥ Wind? Yes Break-in ⊥ Barometer? Yes Break-in ⊥ Barometer | Alarm? No Break-in ⊥ Barometer | Alarm, Wind? Yes



IST, Fall 2020 22 / 66

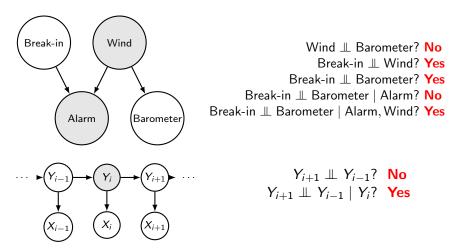


IST, Fall 2020 22 / 66



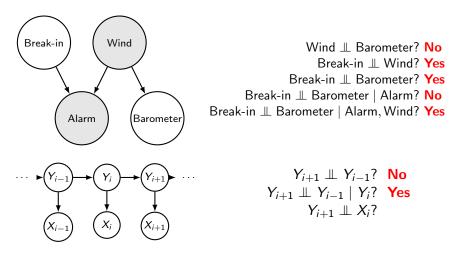
IST, Fall 2020

22 / 66



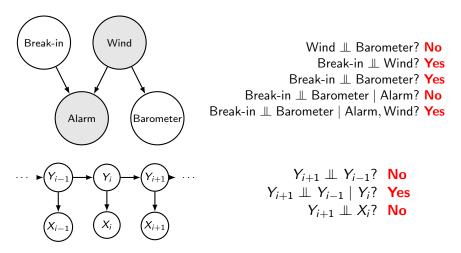
IST, Fall 2020

22 / 66

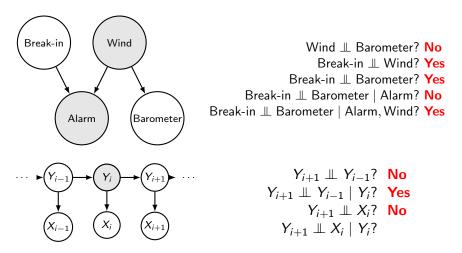


IST, Fall 2020

22 / 66

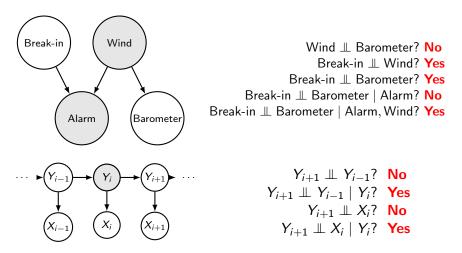


IST, Fall 2020 22 / 66



IST, Fall 2020

22 / 66



IST, Fall 2020

22 / 66

Generative stories and plate notation

In papers, you'll see statistical models defined through generative stories:

 $egin{aligned} \mu &\sim \mathsf{Uniform}([-1,1]) \ \sigma &\sim \mathsf{Uniform}([1,2]) \ X \mid \mu, \sigma &\sim \mathsf{Normal}(\mu,\sigma) \end{aligned}$

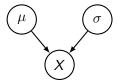
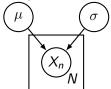


Plate notation is a way to denote repetition of templates:

$$egin{aligned} &\mu \sim \mathsf{Uniform}([-1,1]) \ &\sigma \sim \mathsf{Uniform}([1,2]) \ &X_n \mid \mu, \sigma \sim \mathsf{Normal}(\mu,\sigma) \quad i=1,\ldots,N \end{aligned}$$



Outline

1 Directed Models

Bayes networks

Conditional independence and D-separation

Causal graphs & the do operator

Output: Description of the second second

Markov random fields

Factor graphs

Correlation does not imply causation; but then, *what does?*

André Martins (IST)

Lecture 7: Probabilistic Graphical Models

IST, Fall 2020 25 / 66

Bayes nets only model independence assumptions.

The correlation between the a barometer reading B and wind strength W can be represented either way:

Bayes nets only model independence assumptions.

The correlation between the a barometer reading B and wind strength W can be represented either way:

Seeing that the barometer reading is high, we can forecast wind.

lo	mid	hi
.98	.01	.01
.01	.98	.01
.01	.01	.98
	.98 .01	.98 .01 .01 .98

Bayes nets only model independence assumptions.

The correlation between the a barometer reading B and wind strength W can be represented either way:

Seeing that the barometer reading is high, we can forecast wind.

$\mathbb{P}(W \mid B)$	lo	mid	hi
B = lo	.98	.01	.01
B = mid	.01	.98	.01
B = hi	.01	.01	.98

But setting the barometer needle to high manually won't cause wind! We write: $\mathbb{P}(W \mid do(B = hi)) =?$

IST, Fall 2020

26 / 66

Setting the barometer needle to high manually won't cause wind!

Setting the barometer needle to high manually **won't cause wind!** Two reasons why doing \neq seeing:

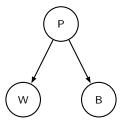
- the direction does not express a causal relationship
- we missed some confounding factor

If we created wind with a ceiling fan, does it alter the barometer?

Setting the barometer needle to high manually **won't cause wind!** Two reasons why doing \neq seeing:

- the direction does not express a causal relationship
- we missed some confounding factor

If we created wind with a ceiling fan, does it alter the barometer? No! **Pressure** is a confounding factor.



Causal models

Definition (Pearl 2000)

A causal model is a DAG \mathcal{G} with vertices X_1, \ldots, X_N representing events. Almost like a BN. However, paths are **causal**.

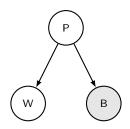
- A causes B only if A is an ancestor of B in G.
- $A \rightarrow B$ means A is a direct cause of B.

A good model is essential.

Wrong causal assumptions \Rightarrow wrong conclusions.

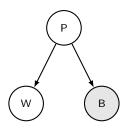
(We won't cover how to assess if the model is right. This is a bit *chicken-and-egg*, but domain knowledge helps, and we are allowed to reason about *unobserved* causes.)

Seeing (*observational*): $\mathbb{P}(W \mid B = hi)$



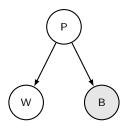
Seeing (*observational*): $\mathbb{P}(W \mid B = hi)$

Measure the world for a while (or call IPMA)					
Date	P	ressure	Wind	Baro	meter
1977-01-01		hi	hi		hi
1977-0	1-02	hi	mid		hi
1977-0	1-02	mid	mid		mid
2019-1	1-03	hi	hi		hi
gives:	$\mathbb{P}(W \mid E)$	3) lo	mid	hi	
	B =	hi .01	.01	.98	-



Seeing (*observational*): $\mathbb{P}(W \mid B = hi)$

Measure the world for a while (or call IPMA)					
Date F		essure	Wind	Baro	meter
1977-01-01		hi	hi		hi
1977-0	1977-01-02		mid		hi
1977-01-02		mid	mid		mid
2019-11-03		hi	hi		hi
gives:	$\mathbb{P}(W \mid B$) lo	mid	hi	
	B = h	i .01	.01	.98	

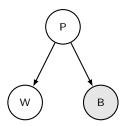


Doing (*interventional*): $\mathbb{P}(W \mid do(B = hi))$

Set the needle to high, breaking inbound arrows; re-generate **new** data in this **new** DAG (or estimate what that would give.)

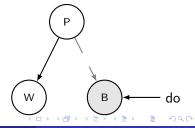
Seeing (*observational*): $\mathbb{P}(W \mid B = hi)$

Measure the world for a while (or call IPMA)					
Date		essure	Wind	Baro	meter
1977-01-01		hi	hi		hi
1977-01-02		hi	mid		hi
1977-01-02		mid	mid		mid
2019-11-03		hi	hi		hi
gives:	$\mathbb{P}(W \mid E$	3) lo	mid	hi	
gives.	B = I	ni .01	.01	.98	



Doing (*interventional*): $\mathbb{P}(W \mid do(B = hi))$

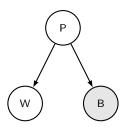
Set the needle to high, breaking inbound arrows; re-generate **new** data in this **new** DAG (or estimate what that would give.)



IST, Fall 2020 29 / 66

Seeing (*observational*): $\mathbb{P}(W \mid B = hi)$

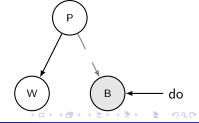
Measure the world for a while (or call IPMA)					
Date	Pre	ssure	Wind	Baro	meter
1977-01-01		hi	hi		hi
1977-01-02		hi	mid		hi
1977-01-02		mid	mid		mid
2019-11-03		hi	hi		hi
gives:	$\mathbb{P}(W \mid B)$	lo	mid	hi	
	B = h	i .01	.01	.98	



Doing (*interventional*): $\mathbb{P}(W \mid do(B = hi))$

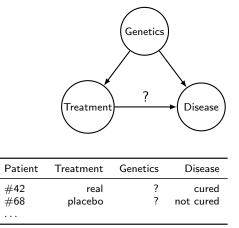
Set the needle to high, breaking inbound arrows; re-generate **new** data in this **new** DAG (or estimate what that would give.)

$$\mathbb{P}(W \mid \mathsf{do}(B = \mathsf{hi})) = \mathbb{P}(W)$$



Randomized controlled trials

Try to actually implement the *do* operator in real life.



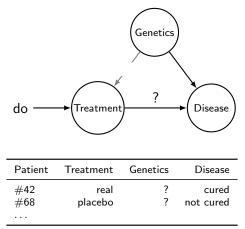
No need to be able to measure genetics as long as we can sample A LOT OF test subjects with no/little bias.

André Martins (IST)

IST, Fall 2020 30 / 66

Randomized controlled trials

Try to actually implement the *do* operator in real life.



No need to be able to measure genetics as long as we can sample A LOT OF test subjects with no/little bias.

André Martins (IST)

IST, Fall 2020 30 / 66

Do-calculus

RCTs are powerful, but often unethical, always expensive.

Do-calculus: use the **causal DAG assumptions** to draw causal conclusions from observational data.

- Apply transformations to $\mathbb{P}(X \mid do(Y))$ until the "do" goes away. (Not always possible!)
- Quantities without "do" can be estimated observationally.
- Transformation: 3 rules.

Pearl's 3 rules

Notation: $\begin{array}{l} \mathfrak{G}_{\bar{X}}\\ \mathfrak{G}_{\bar{X}}\\ Z(\bar{X})\\ y; \operatorname{do}(x) \end{array}$

disjoint sets of events (sets of nodes); may be empty the graph with all edges **into** X removed. the graph with all edges **out of** X removed. subset of nodes in Z which are not ancestors of X. shorthand for Y = y; respectively do(X = x).

1 Ignoring observations:

X, Y, Z, W

 $\mathbb{P}(y \mid \mathsf{do}(x), z, w) = \mathbb{P}(y \mid \mathsf{do}(x), w) \quad \text{if} \quad (Y \perp \!\!\!\!\perp Z \mid X, W)_{\mathfrak{G}_{\bar{X}}}$

2 Action/observation exchange: the back-door criterion

 $\mathbb{P}(y \mid do(x), do(z), w) = \mathbb{P}(y \mid do(x), z, w) \quad \text{if} \quad (Y \perp Z \mid X, W)_{\mathfrak{G}_{\bar{X}, Z}}$

3 Ignoring actions

 $\mathbb{P}(y \mid do(x), do(z), w) = \mathbb{P}(y \mid do(x), w) \quad \text{if} \quad (Y \perp \!\!\!\!\perp Z \mid X, W)_{\mathfrak{G}_{\bar{X}, Z(\bar{W})}}$

- 4 同 2 4 日 2 4 日 2 - 日

Examples 1,2: Pressure and barometer



Rule 3: $\mathbb{P}(P = hi \mid do(B = hi)) = \mathbb{P}(P = hi)$ since $(P \perp B)_{\mathcal{G}_{\bar{B}}}$

Examples 1,2: Pressure and barometer



Rule 3: $\mathbb{P}(P = hi \mid do(B = hi)) = \mathbb{P}(P = hi)$ since $(P \perp B)_{\mathcal{G}_{\bar{B}}}$



Rule 2: $\mathbb{P}(B = hi \mid do(P = lo)) = \mathbb{P}(B = hi \mid P = lo)$ since $(B \perp P)_{\mathcal{G}_{\underline{P}}}$

Examples 1,2: Pressure and barometer



Rule 3: $\mathbb{P}(P = hi \mid do(B = hi)) = \mathbb{P}(P = hi)$ since $(P \perp B)_{\mathcal{G}_{\bar{B}}}$

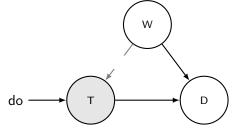


Rule 2: $\mathbb{P}(B = hi \mid do(P = lo)) = \mathbb{P}(B = hi \mid P = lo)$ since $(B \perp P)_{\mathcal{G}_{\underline{P}}}$

Good check: we get the intuitively correct results.

Example 3: Measurable confounder

T: treatment, D: disease. The confounder is W: wealth.



Condition on wealth (which thus needs to be measurable)

$$\mathbb{P}(D = \text{cured} \mid \text{do}(T = y)) = \mathbb{P}(D = \text{cured} \mid \text{do}(T = y), W = y)\mathbb{P}(W = y \mid \text{do}(T = y))$$

$$+ \mathbb{P}(D = \text{cured} \mid \text{do}(T = y), W = n)\mathbb{P}(W = n \mid \text{do}(T = y))$$

$$= \mathbb{P}(D = \text{cured} \mid \text{do}(T = y), W = y)\mathbb{P}(W = y)$$

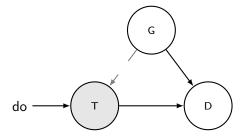
$$+ \mathbb{P}(D = \text{cured} \mid \text{do}(T = y), W = n)\mathbb{P}(W = n) \quad (R3)$$

$$= \mathbb{P}(D = \text{cured} \mid T = y, W = y)\mathbb{P}(W = y)$$

$$+ \mathbb{P}(D = \text{cured} \mid T = y, W = n)\mathbb{P}(W = n) \quad (R2)$$

T: treatment, D: disease.

The confounder is *G*: genetics (impractical to measure and estimate)

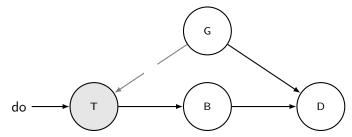


Without more info or more assumptions, we're stuck!

Example 4: a surprisingly possible one

T: treatment, D: disease, B: blood cell count.

The confounder is G: genetics (still hidden)



"The front-door criterion:" conditioning on *B* lets us remove dos!

(I won't show you how, derivation is a bit longer. Try it at home.)

$$\mathbb{P}(D = \text{cured} \mid \text{do}(T = y) = \sum_{b} \mathbb{P}(B = b \mid T = y) \sum_{t} \mathbb{P}(D = \text{cured} \mid T = t, B = b) \mathbb{P}(T = t)$$

Directed models: summary

- Bayes nets: specify & estimate fine-grained distributions over interdependent events.
- Under a specified model, algorithm to decide conditional independence: d-separation
- Bestowing a DAG with causal assumptions lets us reason about interventions.

Further reading: (Pearl, 1988; Koller and Friedman, 2009; Pearl, 2000, 2012; Dawid, 2010) Slides on causal inference and learning causal structure (links):

- Sanna Tyrväinen, Introduction to Causal Calculus
- Ricardo Silva, Causality
- Dominik Janzing & Bernhard Schölkopf, Causality

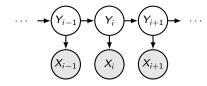
Highly recommended online course: https://www.bradyneal.com/causal-inference-course

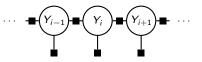
Graphical Models

In this unit, we will formalize & extend these graphical representations encountered in previous lectures.

Directed

Undirected





Outline

Directed Models

Bayes networks

Conditional independence and D-separation

Causal graphs & the do operator

O Undirected Models

Markov random fields

Factor graphs

Outline

1 Directed Models

- Bayes networks
- Conditional independence and D-separation
- Causal graphs & the do operator

2 Undirected Models

Markov random fields

Factor graphs

Modeling friendships

- Four students: An, Bo, Chris, Dee are voting on a Yes/No ballot.
- Friendship pairs: An-Bo, Bo-Chris, Chris-Dee, Dee-An.
- Friends are 100x more likely to vote the same way.

Modeling friendships

- Four students: An, Bo, Chris, Dee are voting on a Yes/No ballot.
- Friendship pairs: An-Bo, Bo-Chris, Chris-Dee, Dee-An.
- Friends are 100x more likely to vote the same way.



• An's vote is a random variable A with values $a \in \{Y, N\}$, and so on.

$$\mathbb{P}(a, b, c, d) \propto f(a, b) \cdot f(b, c) \cdot f(c, d) \cdot f(d, a)$$

For any $X, Y \in \{A, B, C, D\}, f$ is the **compatibility function**:

$$f(x,y) = \begin{cases} 100 & \text{if } x = y = \text{Yes or } x = y = \text{No} \\ 1 & \text{otherwise.} \end{cases}$$

Modeling friendships

- Four students: An, Bo, Chris, Dee are voting on a Yes/No ballot.
- Friendship pairs: An-Bo, Bo-Chris, Chris-Dee, Dee-An.
- Friends are 100x more likely to vote the same way.



• An's vote is a random variable A with values $a \in \{Y, N\}$, and so on.

 $\mathbb{P}(a, b, c, d) \propto f(a, b) \cdot f(b, c) \cdot f(c, d) \cdot f(d, a)$

For any $X, Y \in \{A, B, C, D\}, f$ is the **compatibility function**:

$$f(x,y) = \begin{cases} 100 & \text{if } x = y = \text{Yes or } x = y = \text{No} \\ 1 & \text{otherwise.} \end{cases}$$

Can we represent this exact factorization in a Bayes net?

Modeling friendships

- Four students: An, Bo, Chris, Dee are voting on a Yes/No ballot.
- Friendship pairs: An-Bo, Bo-Chris, Chris-Dee, Dee-An.
- Friends are 100x more likely to vote the same way.



• An's vote is a random variable A with values $a \in \{Y, N\}$, and so on.

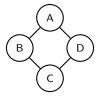
 $\mathbb{P}(a, b, c, d) \propto f(a, b) \cdot f(b, c) \cdot f(c, d) \cdot f(d, a)$

For any $X, Y \in \{A, B, C, D\}, f$ is the **compatibility function**:

$$f(x,y) = \begin{cases} 100 & \text{if } x = y = \text{Yes or } x = y = \text{No} \\ 1 & \text{otherwise.} \end{cases}$$

Can we represent this exact factorization in a Bayes net? No!

Markov random fields



Definition

Let \mathcal{G} be an *undirected* graph with nodes corresponding to random variables X_1, \ldots, X_N . Let $C(\mathcal{G})$ denote the set of *cliques* (fully connected subgraphs) of \mathcal{G} . A MRF is a distribution of the form

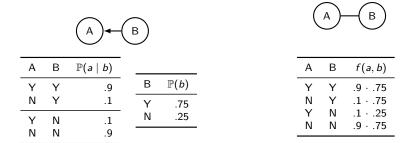
$$\mathbb{P}(x_1,\ldots,x_n)=\frac{1}{Z}\prod_{c\in C}f_c(\boldsymbol{x}_c)$$

where for each clique c, f_c is a non-negative compatibility function.

2 Convert all arcs $A \rightarrow B$ or $A \leftarrow B$ into undirected edges A - B.



2 Convert all arcs $A \rightarrow B$ or $A \leftarrow B$ into undirected edges A - B.



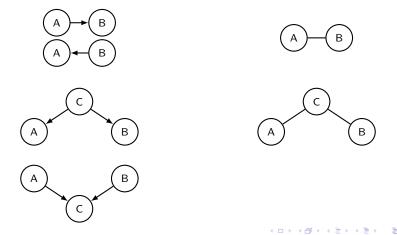
▶ < ∃ >

2 Convert all arcs $A \rightarrow B$ or $A \leftarrow B$ into undirected edges A - B.

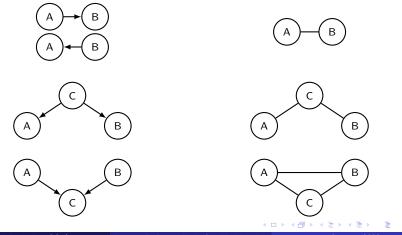




2 Convert all arcs $A \rightarrow B$ or $A \leftarrow B$ into undirected edges A - B.

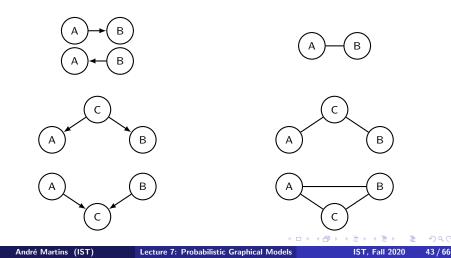


2 Convert all arcs $A \rightarrow B$ or $A \leftarrow B$ into undirected edges A - B.



Lecture 7: Probabilistic Graphical Models

First, add edge A - C for any collider structure A → B ← C;
 Convert all arcs A → B or A ← B into undirected edges A - B.

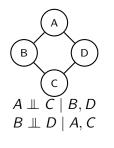


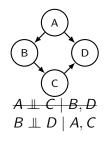
Loose conversion

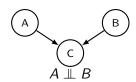
Similarly, we can convert a MRF to a BN (we won't cover it.) However, **independences may be lost** in either direction.

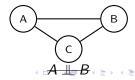


То









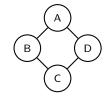
Bayes vs Markov

Bayes network

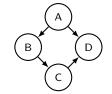
- Factors are conditionals (normalized)
- Easy to sample
- Can be made causal
- Can easily find $\mathbb{P}(x_1,\ldots,x_n)$.

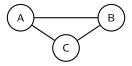
Markov networks

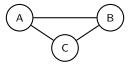
- Factors are cliques (unnormalized)
- No directional ambiguity
- Often more compact
- More symmetric notation



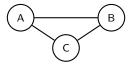
$$\mathbb{P}(a, b, c, d) = 1/2 f_1(a, b) f_2(b, c) f_3(c, d) f_4(d, a)$$



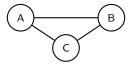




Single clique: $\{A, B, C\}$, so $\mathbb{P}(a, b, c) = \frac{1}{Z}f(a, b, c)$.



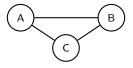
Single clique: $\{A, B, C\}$, so $\mathbb{P}(a, b, c) = \frac{1}{Z}f(a, b, c)$. No way to represent $\mathbb{P}(a, b, c) = \frac{1}{Z}f_1(a, b)f_2(b, c)f_3(c, a)$.



Single clique: $\{A, B, C\}$, so $\mathbb{P}(a, b, c) = \frac{1}{Z}f(a, b, c)$.

No way to represent $\mathbb{P}(a, b, c) = 1/2 f_1(a, b)f_2(b, c)f_3(c, a)$.

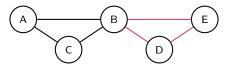
Pairwise MRF: Like a MRF, but factors are edges rather than cliques.



Single clique: $\{A, B, C\}$, so $\mathbb{P}(a, b, c) = \frac{1}{Z}f(a, b, c)$.

No way to represent $\mathbb{P}(a, b, c) = 1/Z f_1(a, b)f_2(b, c)f_3(c, a)$.

Pairwise MRF: Like a MRF, but factors are edges rather than cliques. But what if we want to mix them?



 $\mathbb{P}(a, b, c, d, e) = \frac{1}{Z} f_1(a, b) f_2(b, c) f_3(c, a) f_4(b, d, e)$

Outline

1 Directed Models

- Bayes networks
- Conditional independence and D-separation
- Causal graphs & the do operator

O Undirected Models

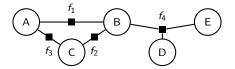
Markov random fields

Factor graphs

Factor graphs

Explicitly represent factors in the graph to remove ambiguity.

 $\mathbb{P}(a, b, c, d, e) = \frac{1}{Z} f_1(a, b) f_2(b, c) f_3(c, a) f_4(b, d, e)$



Definition (Factor graph)

A FG is a bipartite graph \mathcal{G} with vertices in $\mathcal{V} \cup \mathcal{F}$, where $X_1, \ldots, X_n \in \mathcal{V}$ are random variables and $\alpha \in \mathcal{F}$ are factors, inducing a distribution

$$\mathbb{P}(x_1,\ldots,x_n)=\frac{1}{Z}\prod_{\alpha\in\mathcal{F}}f_{\alpha}(\boldsymbol{x}_{\alpha})$$

where $f_{\alpha} \geq 0$, and X_{α} is the set of variables with an edge to factor α .

Factor graphs

- Any MRF can be mapped exactly to a FG (clique \rightarrow factor).
- Any Pairwise MRF can be mapped exactly to a FG (edge \rightarrow factor).
- FGs are more general / more *fine-grained*.

Algorithms

- Inference: Given a FG with compatibility functions, answer queries
 - Maximization: Find most likely assignment x₁,..., x_N (possibly given evidence x_i : i ∈ E).

$$\operatorname{argmax}_{x_1,\ldots,x_M} \mathbb{P}(x_1,\ldots,x_N \mid \boldsymbol{x}_{\mathcal{E}})$$

 Marginalization: Find the marginal probability of some partial assignment over x_j : j ∈ M (possibly given evidence x_i : i ∈ E)

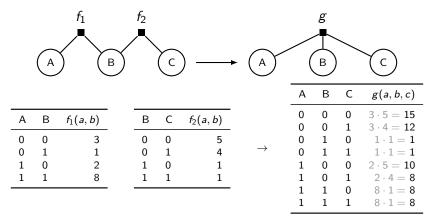
$$\mathbb{P}(\mathbf{x}_{\mathcal{M}} \mid \mathbf{x}_{\mathcal{E}})$$

- NP-hard / #P-hard in general, but doable for tree-shaped graphs with dynamic programming.
- Learning: Given a dataset, estimate the compatibility tables (or, in general a model that produces them.)

Since $\text{BN} \rightarrow \text{MRF} \rightarrow \text{FG},$ it suffices to study inference algorithms for $\text{FG}.^1$

Multiplying factors

A core operation: combining factors by multipliying them.

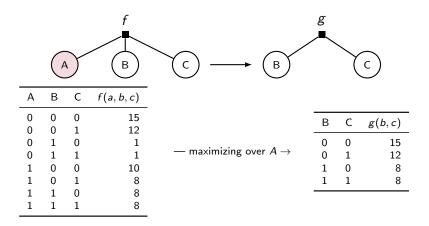


Distribution is preserved:

$$f_1(a,b) \cdot f_2(b,c) \cdot f_3(\dots) \cdot \dots = g(a,b,c) \cdot f_3(\dots) \cdot \dots$$

IST, Fall 2020 51/66

Maximizing over a variable



$$\max_{a} f(a, b, c) \cdot \underbrace{f_4(\ldots) \cdot \ldots}_{A-\text{free}} = g(b, c) \cdot f_4(\ldots) \cdot \ldots$$

IST, Fall 2020 52 / 66

Marginalizing over a variable

	A	Y	f B	$C \longrightarrow B$	g		С
А	В	С	f(a, b, c)				
0	0	0	15		В	С	g(b, c)
0	0	1	12				8(0,0)
0	1	0	1		0	0	25
0	1	1	1	— summing over $A ightarrow$	0	1	20
1	0	0	10		1	0	9
1	0	1	8		1	1	9
1	1	0	8				
1	1	1	8				

$$\sum_{a} f(a, b, c) \cdot \underbrace{f_4(\ldots) \cdot \ldots}_{A-\text{free}} = g(b, c) \cdot f_4(\ldots) \cdot \ldots$$

IST, Fall 2020 53 / 66

$$\begin{array}{c|c} A & f_{AB} & B & f_{BC} & C & f_{CD} \\ \hline \end{array} \\ \hline \end{array} \\ \hline \end{array} \\ \begin{array}{c|c} & & \\ & & & \\ & & & \\$$

Query: $\max_{a,b,c,d} \mathbb{P}(a,b,c,d) =?$

ΑB	$f_{AB}(a, b)$
0 0	10
0 1	2
1 0	3
1 1	9
ВC	$f_{BC}(b,c)$
0 0	1
0 1	3
1 0	1
1 1	2
C D	$f_{CD}(c,d)$
0 0	4
0 1	2
1 0	1
1 1	•
$1 \ 1$	3

$$\begin{array}{c|c} A & f_{AB} & B & f_{BC} & C & f_{CD} \\ \hline \end{array} \\ \hline \end{array} \\ \hline \end{array} \\ \begin{array}{c|c} & & \\ & & & \\ & & & \\$$

Query: $\max_{a,b,c,d} \mathbb{P}(a,b,c,d) =?$

ΑB	$f_{AB}(a,b)$
00	10
$\begin{array}{c} 0 & 1 \\ 1 & 0 \end{array}$	2 3
$1 \ 0 \ 1 \ 1 \ 1$	9 9
ВC	$f_{BC}(b,c)$
0 0	1
0 1	3
10	1
1 1	2
C D	$f_{CD}(c,d)$
0 0	4
0 1	2
$\begin{array}{c}1 \\ 1 \end{array}$	1 3

1 Pick order: D, C, B, A

< D > < A > < B > < B >

André Martins (IST)

Lecture 7: Probabilistic Graphical Models

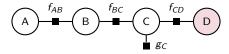
IST, Fall 2020 54 / 66

$$\begin{array}{c|c} A & f_{AB} & B & f_{BC} & f_{CD} \\ \hline A & \bullet & B & \bullet & C & \bullet & D \\ \end{array}$$

Query: $\max_{a,b,c,d} \mathbb{P}(a,b,c,d) =?$

ΑB	$f_{AB}(a, b)$
$\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}$	10 2
$\begin{array}{c} 0 & 1 \\ 1 & 0 \end{array}$	23
$1 \ 1 \ 1$	9
	9
ВC	$f_{BC}(b,c)$
0 0	1
0 1	3
1 0	1
11	2
C D	$f_{CD}(c,d)$
0 0	4
0 1	2
1 0	1
$1 \ 1$	3

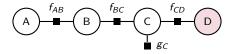
1	Pick order: D, C, B, A
2	Maximize over D ($f_{CD} \rightarrow g_C$)



Query: $\max_{a,b,c,d} \mathbb{P}(a, b, c, d) =?$

		-	
ΑB	$f_{AB}(a,b)$		
0 0	10		
01	2 3		
$\begin{array}{c} 1 & 0 \\ 1 & 1 \end{array}$	3 9		
	9		
ВC	$f_{BC}(b,c)$	_	
0 0	1		
0 1	3		
1 0	1		
1 1	2		
СD	$f_{CD}(c,d)$	_	
0 0	4	C	$g_C(c)$
0 1	2	0	4 ^{D=0} 3 ^{D=1}
$\begin{array}{c} 1 & 0 \\ 1 & 1 \end{array}$	1 3	1	3D=1

Pick order: D, C, B, A
 Maximize over D (f_{CD} → g_C)

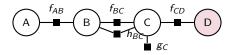


Query: $\max_{a,b,c,d} \mathbb{P}(a,b,c,d) = ?$

A B	$f_{AB}(a, b)$		
0 0	10		
01	2 3		
10 11	3 9		
1 1	9		
ВC	$f_{BC}(b,c)$	-	
0 0	1		
0 1	3		
1 0	1		
1 1	2	_	
СD	$f_{CD}(c,d)$	_	
0 0	4	C	$g_C(c)$
0 1	2 - 1 3	0	4 ^{D=0}
1 0 1 1	1	1	4 ^{D=0} 3 ^{D=1}
			-

1 Pick order: D, C, B, A

- **2** Maximize over $D(f_{CD} \rightarrow g_C)$
- 3 Multiply f_{BC} with g_C giving h_{BC}



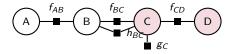
Query: $\max_{a,b,c,d} \mathbb{P}(a,b,c,d) =?$

		_	
ΑB	$f_{AB}(a,b)$	_	
$\begin{array}{ccc} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{array}$	10 2 3 9	_	
ВC	$f_{BC}(b,c)$	-	
0 0 0 1 1 0 1 1	1 3 1 2	-	
СD	$f_{CD}(c,d)$		
0 0	4	С	gc
$\begin{array}{c} 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{array}$	4 2 1 3	0 1	4 3

ВC	$h_{BC}(b,c)$
0 0 0 1 1 0 1 1	$\begin{array}{l} 1 \cdot 4 = 4^{D=0} \\ 3 \cdot 3 = 9^{D=1} \\ 1 \cdot 4 = 4^{D=0} \\ 2 \cdot 3 = 6^{D=1} \end{array}$

- 1 Pick order: D, C, B, A
- **2** Maximize over $D(f_{CD} \rightarrow g_C)$

3 Multiply f_{BC} with g_C giving h_{BC}



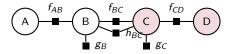
Query: $\max_{a,b,c,d} \mathbb{P}(a,b,c,d) =?$

	-		
	_	$f_{AB}(a,b)$	ΑB
	-	10 2 3 9	0 0 0 1 1 0 1 1
	•	$f_{BC}(b,c)$	ВC
	-	1 3 1 2	0 0 0 1 1 0 1 1
		$f_{CD}(c,d)$	CD
$g_C(c)$	С	4	0 0
4 ^{D=0} 3 ^{D=1}	0 1	2 ⁻ 1 3	$egin{array}{ccc} 0 & 1 \ 1 & 0 \ 1 & 1 \end{array}$

ВC	$h_{BC}(b,c)$
0 0 0 1 1 0 1 1	$\begin{array}{l} 1 \cdot 4 = 4^{D=0} \\ 3 \cdot 3 = 9^{D=1} \\ 1 \cdot 4 = 4^{D=0} \\ 2 \cdot 3 = 6^{D=1} \end{array}$

1 Pick order: D, C, B, A

- **2** Maximize over $D(f_{CD} \rightarrow g_C)$
- 3 Multiply f_{BC} with g_C giving h_{BC}
- 4 Maximize over C ($h_{BC} \rightarrow g_B$)

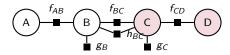


Query: $\max_{a,b,c,d} \mathbb{P}(a,b,c,d) =?$

	$f_{AB}(a,b)$	AΒ
	10	0 0
	2 3	$\begin{array}{c} 0 & 1 \\ 1 & 0 \end{array}$
	9	1 1
3	$f_{BC}(b,c)$	ВC
5	1	0 0
)	3	0 1
1	1	1 0
	2	1 1
	$f_{CD}(c,d)$	сd
2	4	0 0
)	2 - 1	0 1
,	1	10
-	2	1 1

ВC	$h_{BC}(b,c)$
0 0 0 1 1 0 1 1	$1 \cdot 4 = 4^{D=0} \\ 3 \cdot 3 = 9^{D=1} \\ 1 \cdot 4 = 4^{D=0} \\ 2 \cdot 3 = 6^{D=1}$

- 1 Pick order: D, C, B, A
- **2** Maximize over $D(f_{CD} \rightarrow g_C)$
- 3 Multiply f_{BC} with g_C giving h_{BC}
- 4 Maximize over $C (h_{BC} \rightarrow g_B)$



Query: $\max_{a,b,c,d} \mathbb{P}(a,b,c,d) =?$

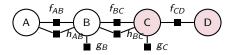
	• -	f _{AB} (a, b) 10 2 3 9	A B 0 0 0 1 1 0 1 1
B $g_B(b)$	B	$f_{BC}(b,c)$	ВC
$ \begin{array}{ccc} 0 & 9^{C=1} \\ 1 & 6^{C=1} \end{array} $	-	1 3 1	0 0 0 1 1 0
	-	$f_{CD}(c,d)$	1 1 C D
$C g_C(c)$	С	4	0 0
0 4 ^{D=0} 1 3 ^{D=1}	0	2 ⁻ 1 3	0 1 1 0 1 1

ВC	$h_{BC}(b,c)$
0 0	$1 \cdot 4 = 4^{D=0}$ $3 \cdot 3 = 9^{D=1}$
$\begin{array}{c} 0 & 1 \\ 1 & 0 \end{array}$	$3 \cdot 3 \equiv 9^{D-1}$ $1 \cdot 4 = 4^{D=0}$
1 1	$2 \cdot 3 = 6^{D=1}$

- 1 Pick order: D, C, B, A
- **2** Maximize over $D(f_{CD} \rightarrow g_C)$
- 3 Multiply f_{BC} with g_C giving h_{BC}
- 4 Maximize over C ($h_{BC} \rightarrow g_B$)
- **5** Multiply f_{AB} with g_B giving h_{AB}

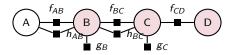
IST, Fall 2020

54 / 66



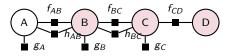
Query: $\max_{a,b,c,d} \mathbb{P}(a,b,c,d) =?$

A B 0 0 0 1 1 0 1 1	f _{AB} (a, b) 10 2 3 9		A B 0 0 0 1 1 0 1 1	$ \begin{array}{c} h_{AB}(a,b) \\ \hline 10 \cdot 9 = 90^{C=1} \\ 2 \cdot 6 = 12^{C=1} \\ 3 \cdot 9 = 27^{C=1} \\ 9 \cdot 6 = 54^{C=1} \end{array} $	1 Pick order: D, C, B, A 2 Maximize over $D (f_{CD} \rightarrow g_C)$ 3 Multiply f_{BC} with g_C giving h_{BC}
B C 0 0 0 1 1 0 1 1	f _{BC} (b,c)	$\begin{array}{c} B & g_B(b) \\ 0 & 9^{C=1} \\ 1 & 6^{C=1} \end{array}$		$ \begin{array}{c} h_{BC}(b,c) \\ \hline 1 \cdot 4 = 4^{D=0} \\ 3 \cdot 3 = 9^{D=1} \\ 1 \cdot 4 = 4^{D=0} \\ 2 \cdot 3 = 6^{D=1} \end{array} $	4 Maximize over $C (h_{BC} \rightarrow g_B)$ 5 Multiply f_{AB} with g_B giving h_{AB}
C D 0 0 0 1 1 0 1 1	$f_{CD}(c,d) = \begin{pmatrix} 4 \\ 2 \\ 1 \\ 3 \\ 3 \end{pmatrix}$	$ \begin{array}{ccc} C & g_C(c) \\ 0 & 4^{D=0} \\ 1 & 3^{D=1} \end{array} $		2.3 - 0	(ロ) (句) (ミ) (ミ) ミ うへの



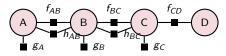
Query: $\max_{a,b,c,d} \mathbb{P}(a,b,c,d) =?$

ΑB	$f_{AB}(a, b)$		ΑB	$h_{AB}(a,b)$	• 1 Pick order: D, C, B, A
0 0 0 1 1 0 1 1	10 2 3 9		0 0 0 1 1 0 1 1	$\begin{array}{l} 10 \cdot 9 = 90^{C=1} \\ 2 \cdot 6 = 12^{C=1} \\ 3 \cdot 9 = 27^{C=1} \\ 9 \cdot 6 = 54^{C=1} \end{array}$	 2 Maximize over D (f_{CD} → g_C) 3 Multiply f_{BC} with g_C giving h_{BC}
ВC	$f_{BC}(b,c)$	B $g_B(b)$	ВC	$h_{BC}(b,c)$	4 Maximize over $C (h_{BC} \rightarrow g_B)$
0 0 0 1 1 0 1 1	1 · 3 1	$ \begin{array}{ccc} 0 & 9^{C=1} \\ 1 & 6^{C=1} \end{array} $		$1 \cdot 4 = 4^{D=0}$ $3 \cdot 3 = 9^{D=1}$ $1 \cdot 4 = 4^{D=0}$	5 Multiply f_{AB} with g_B giving h_{AB}
C D	$f_{CD}(c,d)$	<u>.</u>	1 1	$2 \cdot 3 = 6^{D=1}$	6 Maximize over B ($h_{AB} \rightarrow g_A$)
0 0 0 1 1 0 1 1	4 2 ⁻ 1 3	$ \begin{array}{ccc} C & g_C(c) \\ 0 & 4^{D=0} \\ 1 & 3^{D=1} \end{array} $			< ロ > < 御 > < 三 > < 三 > 、 三 > 、 三 、 の Q (4)



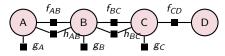
Query: $\max_{a,b,c,d} \mathbb{P}(a,b,c,d) =?$

		_				
ΑB	$f_{AB}(a, b)$	А	$g_A(a)$	ΑB	$h_{AB}(a, b)$	1 Pick order: D, C, B, A
0 0 0 1 1 0 1 1	10 2 3 9	0 1	90 ^{B=0} 54 ^{B=1}	0 0 0 1 1 0 1 1	$\begin{array}{l} 10 \cdot 9 = 90^{C=1} \\ 2 \cdot 6 = 12^{C=1} \\ 3 \cdot 9 = 27^{C=1} \\ 9 \cdot 6 = 54^{C=1} \end{array}$	 2 Maximize over D ($f_{CD} \rightarrow g_C$) 3 Multiply f_{BC} with g_C giving h_{BC}
B C 0 0 0 1	$f_{BC}(b,c)$ 1 3	- B 0	$g_B(b)$ 9 ^{C=1}	B C 0 0	$h_{BC}(b,c)$ $1 \cdot 4 = 4^{D=0}$	4 Maximize over $C (h_{BC} \rightarrow g_B)$ 5 Multiply f_{AB} with g_B giving
1 0 1 1 C D	$f_{CD}(c,d)$	1	6 ^{C=1}	0 1 1 0 1 1	$3 \cdot 3 = 9^{D=1}$ $1 \cdot 4 = 4^{D=0}$ $2 \cdot 3 = 6^{D=1}$	h_{AB} (b) Maximize over $B(h_{AB} \rightarrow g_A)$
0 0 0 1 1 0 1 1	4 2 1 3	C 0 1	$g_C(c)$ $4^{D=0}$ $3^{D=1}$			< ロ > (得 > < 三 > 、 ミ > 、 三 、 の Q Q



Query: $\max_{a,b,c,d} \mathbb{P}(a, b, c, d) =?$

ΑB	$f_{AB}(a, b)$	А	$g_A(a)$		ΑB	h _{AB} (a, b)	1	Pick order: D, C, B, A
0 0 0 1 1 0 1 1	10 2 3 9	0 1	90 ^{B=0} 54 ^{B=1}		0 0 0 1 1 0 1 1	$10 \cdot 9 = 90^{C=1}$ 2 \cdot 6 = 12^{C=1} 3 \cdot 9 = 27^{C=1} 9 \cdot 6 = 54^{C=1}	3	Maximize over $D(f_{CD} \rightarrow g_C)$ Multiply f_{BC} with g_C giving
B C	$f_{BC}(b,c)$	- B	$g_B(b)$	-	BC	$h_{BC}(b,c)$	_	h_{BC} Maximize over C $(h_{BC} ightarrow g_B)$
0 1 1 0	1 3 1	0 1	$9^{C=1}$ $6^{C=1}$		0 0 0 1 1 0	$1 \cdot 4 = 4^{D=0}$ $3 \cdot 3 = 9^{D=1}$ $1 \cdot 4 = 4^{D=0}$	5	Multiply f_{AB} with g_B giving h_{AB}
1 1 C D	$f_{CD}(c, d)$	_		_	1 1	$1 \cdot 4 = 4$ $2 \cdot 3 = 6^{D=1}$	-	Maximize over B $(h_{AB} \rightarrow g_A)$ Maximize over A $(g_A \rightarrow \emptyset)$
0 0 0 1 1 0 1 1		C 0 1	$g_C(c) = \frac{4^{D=0}}{3^{D=1}}$					□ > < @ > < ≥ > < ≥ > < ≥ < ≥ の < @



Query: $\max_{a,b,c,d} \mathbb{P}(a,b,c,d) =?$

			_			
A B $f_{AB}(a, b)$	А	$g_A(a)$	ΑB	$h_{AB}(a, b)$	1 Pi	ck order: D, C, B, A
0 0 10 0 1 2 1 0 3 1 1 9	0 1	90 ^{B=0} 54 ^{B=1}	0 0 0 1 1 0	$10 \cdot 9 = 90^{C=1}$ 2 \cdot 6 = 12^{C=1} 3 \cdot 9 = 27^{C=1}	-	aximize over $D \; (f_{CD} o g_C)$ ultiply f_{BC} with g_C giving
$\begin{array}{c} 1 & 1 & 9 \\ \hline \\ B & C & f_{BC}(b,c) \end{array}$			11	$9 \cdot 6 = 54^{C=1}$	h _E	
$\frac{1}{0} \frac{1}{0} \frac{1}$	– B	$g_B(b)$	B C	$h_{BC}(b,c)$	_	aximize over C $(h_{BC} ightarrow g_B)$
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	0 1	$9^{C=1}$ $6^{C=1}$	0 0 0 1	$1 \cdot 4 = 4^{D=0}$ $3 \cdot 3 = 9^{D=1}$	5 M h _A	ultiply <i>f_{AB}</i> with <i>g_B</i> giving
1 1 2			- 10 11	$1 \cdot 4 = 4^{D=0}$ $2 \cdot 3 = 6^{D=1}$	_	aximize over $B~(h_{AB} ightarrow g_A)$
$C D f_{CD}(c, d)$					🕜 M	aximize over $A\;(g_A o \emptyset)$
	С	$g_C(c)$	_		🚯 Ju	st like Viterbi!
0 1 2 1 0 1 1 1	0 1	$4^{D=0}$ $3^{D=1}$				ne max is $90/z$.
<u>11</u> 3				ushahilistis CusulissI M		 ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●

André Martins (IST)

Lecture 7: Probabilistic Graphical Models

IST, Fall 2020 54 / 66

$$\begin{array}{c|c} A & f_{AB} & B & f_{BC} & C & f_{CD} \\ \hline A & \bullet & B & \bullet & C & \bullet & D \\ \end{array}$$

Query:
$$Z = \sum_{a,b,c,d} f(a,b,c,d) =?$$

ΑB	$f_{AB}(a, b)$	
0 0	10	
0 1	2	
1 0	3	
1 1	9	
ВC	$f_{BC}(b,c)$	
0 0	1	
0 1	3	
1 0	1	
1 1	2	
C D	$f_{CD}(c,d)$	
0 0	4	
0 1	2	
1 0	1	
1 1	3	

$$\begin{array}{c|c} A & f_{AB} & B & f_{BC} & C & f_{CD} \\ \hline A & \bullet & B & \bullet & C & \bullet & D \\ \end{array}$$

Query:
$$Z = \sum_{a,b,c,d} f(a,b,c,d) = ?$$

ΑB	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9
ВC	$f_{BC}(b,c)$
0 0	1
0 1	3
1 0	1
1 1	2
C D	$f_{CD}(c,d)$
0 0	4
0 1	2
1 0	1
$1 \ 1$	3

1 Pick order: D, C, B, A

André Martins (IST)

Lecture 7: Probabilistic Graphical Models

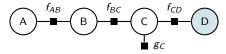
IST, Fall 2020 55 / 66

$$A \xrightarrow{f_{AB}} B \xrightarrow{f_{BC}} C \xrightarrow{f_{CD}} D$$

Query:
$$Z = \sum_{a,b,c,d} f(a,b,c,d) = ?$$

$f_{AB}(a,b)$
10 2
3
9
$f_{BC}(b,c)$
1
3
1
-
1
1 2
$\frac{1}{2}$ $f_{CD}(c,d)$ $\frac{4}{2}$
$\frac{1}{2}$ $f_{CD}(c,d)$ 4

1	Pick order: D, C, B, A
2	Sum over $D(f_{CD} \rightarrow g_C)$

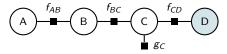


Query:	Z =	$\sum_{a,b,c,d}$	f(a,	<i>b</i> , <i>c</i> ,	, d) =?
--------	-----	------------------	------	-----------------------	---------

	ΑB	$f_{AB}(a,b)$	_	
	0 0	10	-	
	0 1	2		
	1 0	3		
_	1 1	9	_	
-	ВC	$f_{BC}(b,c)$	-	
-	0 0	1	-	
	0 1	3		
	1 0	1		
	1 1	2		
-			-	
-	1 1 C D	$f_{CD}(c,d)$	- C	g _C (c)
-	1 1 C D 0 0	$f_{CD}(c,d)$	-	g _C (c)
-	1 1 C D 0 0 0 1	f _{CD} (c, d) - 4 - 2	0	6
-	1 1 C D 0 0	$f_{CD}(c,d)$	-	

D	Pick order: D, C, B, A
2	Sum over $D(f_{CD} \rightarrow g_C)$

э



Query:
$$Z = \sum_{a,b,c,d} f(a,b,c,d) =$$
?

ΑB	$f_{AB}(a,b)$	_	
0 0	10	-	
0 1	2		
1 0	3		
1 1	9	_	
ВC	$f_{BC}(b,c)$	•	
0 0	1		
0 1	3		
1 0	1		
	-		
1 1	2	_	
1 1 C D	$f_{CD}(c, d)$.	-	
СD	$f_{CD}(c,d)$	- C	g _C (c)
C D 0 0	$f_{CD}(c,d)$	-	,
C D 0 0 0 1	f _{CD} (c, d) - 4 - 2	0	6
C D 0 0	$f_{CD}(c,d)$	-	,

0	Pick	order:	D,	С,	В, А	
---	------	--------	----	----	------	--

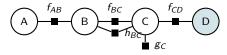
2 Sum over $D(f_{CD} \rightarrow g_C)$

3 Multiply f_{BC} with g_C giving h_{BC}

André Martins (IST)

Lecture 7: Probabilistic Graphical Models

IST, Fall 2020 55 / 66



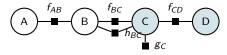
Query:
$$Z = \sum_{a,b,c,d} f(a,b,c,d) =?$$

	$f_{AB}(a,b)$	ΑB
	10	0 0
	2 3	$\begin{array}{c} 0 & 1 \\ 1 & 0 \end{array}$
	5 9	1 0 1 1
	$f_{BC}(b,c)$	ВC
	1	0 0
	3	0 1
	1	1 0
	2	1 1
	$f_{CD}(c,d)$ -	СD
-C gc	4 -	0 0
0	2	0 1
1	1	1 0
	3 -	1 1

ВC	$h_{BC}(b,c)$
0 0 0 1 1 0 1 1	$1 \cdot 6 = 6 \\ 3 \cdot 4 = 12 \\ 1 \cdot 6 = 6 \\ 2 \cdot 4 = 8$

- 1 Pick order: D, C, B, A
- **2** Sum over $D(f_{CD} \rightarrow g_C)$
- **3** Multiply f_{BC} with g_C giving h_{BC}

(c) 6 4

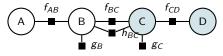


Query:
$$Z = \sum_{a,b,c,d} f(a,b,c,d) = ?$$

	_	$f_{AB}(a, b)$	ΑB
	-	10	0 0
		2	01
		3 9	$\begin{smallmatrix}1&0\\1&1\end{smallmatrix}$
		$f_{BC}(b,c)$	ВC
	-	1	0.0
		3	0 1
		1	1 0
	_	2	1 1
		$f_{CD}(c,d)$ -	СD
$g_C(c)$	- C	4	0.0
6	0	2	0 1
4	1	1	1 0
		3 -	1 1

ВC	$h_{BC}(b,c)$
0 0 0 1 1 0 1 1	$1 \cdot 6 = 6$ $3 \cdot 4 = 12$ $1 \cdot 6 = 6$ $2 \cdot 4 = 8$

- Pick order: D, C, B, A
- **2** Sum over $D(f_{CD} \rightarrow g_C)$
- **3** Multiply f_{BC} with g_C giving h_{BC}
- **4** Sum over C ($h_{BC} \rightarrow g_B$)

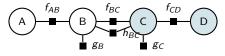


Query: $Z = \sum_{a,b,c,d} f(a, b, c, d) =?$

ΑB	$f_{AB}(a,b)$	_	
$\begin{array}{ccc} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{array}$	10 2 3 9	_	
ВC	$f_{BC}(b,c)$	B	$g_B(b)$
0 0 0 1 1 0	1 3 1	0 1	18 14
1 1 C D	$f_{CD}(c, d)$.	-	
0 0 0 1 1 0 1 1	4. 2 1 3	-C 0 1	g _C (c) 6 4

ВC	$h_{BC}(b,c)$
0 0 0 1 1 0 1 1	$1 \cdot 6 = 6$ $3 \cdot 4 = 12$ $1 \cdot 6 = 6$ $2 \cdot 4 = 8$

- Pick order: D, C, B, A
- **2** Sum over $D(f_{CD} \rightarrow g_C)$
- **3** Multiply f_{BC} with g_C giving h_{BC}
- **4** Sum over C ($h_{BC} \rightarrow g_B$)

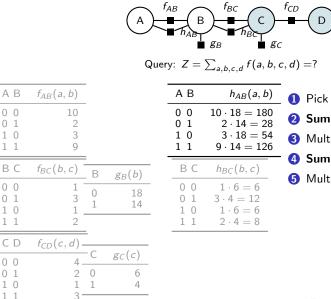


Query: $Z = \sum_{a,b,c,d} f(a, b, c, d) =?$

A B 0 0 0 1 1 0	f _{AB} (a, b) 10 2 3	-			
1 1	9	-			
ВC	$f_{BC}(b,c)$	В	$g_B(b)$	-	В
0 0 0 1 1 0	1 3 1	0 1	18 14	-	0 0 1
1 1	2				1
СD	$f_{CD}(c,d)$ -		()		
0 0 0 1 1 0 1 1	4 - 2 1 3 -	-C 0 1	g _C (c) 6 4		

3 C	$h_{BC}(b,c)$
0	$1 \cdot 6 = 6$
) 1	$3 \cdot 4 = 12$ $1 \cdot 6 = 6$
. 1	$2 \cdot 4 = 8$

- Pick order: D, C, B, A
- **2** Sum over $D(f_{CD} \rightarrow g_C)$
- **3** Multiply f_{BC} with g_C giving h_{BC}
- 4 Sum over C ($h_{BC} \rightarrow g_B$)
- **5** Multiply f_{AB} with g_B giving h_{AB}



Pick order: D, C, B, A

2 Sum over $D(f_{CD} \rightarrow g_C)$

- **3** Multiply f_{BC} with g_C giving h_{BC}
- 4 Sum over C $(h_{BC} \rightarrow g_B)$
- **(5)** Multiply f_{AB} with g_B giving h_{AB}

1

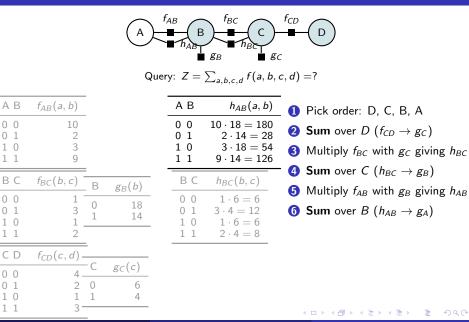
1 0

0 1

1 1

1 0

1 1

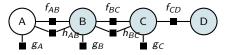


0 1

0 1

1 1

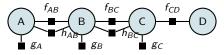
1 1



Query: $Z = \sum_{a,b,c,d} f(a,b,c,d) =?$

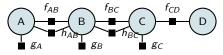
1 Pick orde	$h_{AB}(a, b)$	ΑB	$g_A(a)$	А	$f_{AB}(a, b)$	ΑB
2 Sum over3 Multiply #	$10 \cdot 18 = 180 2 \cdot 14 = 28 3 \cdot 18 = 54 9 \cdot 14 = 126$	0 0 0 1 1 0 1 1	208 180	0 1	10 2 3 9	0 0 0 1 1 0 1 1
 Sum over Multiply <i>i</i> Sum over 	$ \begin{array}{r} h_{BC}(b,c) \\ \hline 1 \cdot 6 = 6 \\ 3 \cdot 4 = 12 \\ 1 \cdot 6 = 6 \\ 2 \cdot 4 = 8 \end{array} $	B C 0 0 0 1 1 0 1 1	g _B (b) 18 14	B 0 1	f _{BC} (b,c)	B C 0 0 0 1 1 0 1 1
< • > < (5)			g _C (c) 6 4	- C 0 1	f _{CD} (c, d) - 4 - 2 1 3 -	C D 0 0 0 1 1 0 1 1

- Pick order: D, C, B, A
- **2** Sum over $D(f_{CD} \rightarrow g_C)$
- **3** Multiply f_{BC} with g_C giving h_{BC}
- **4** Sum over C ($h_{BC} \rightarrow g_B$)
- **5** Multiply f_{AB} with g_B giving h_{AB}
- **6** Sum over $B(h_{AB} \rightarrow g_A)$



Query: $Z = \sum_{a,b,c,d} f(a, b, c, d) =?$

						_
ΑB	$f_{AB}(a, b)$	А	$g_A(a)$	ΑB	$h_{AB}(a, b)$	1 Pick order: D, C, B, A
00	10 2	0 1	208 180	$\begin{array}{c} 0 & 0 \\ 0 & 1 \end{array}$	$10 \cdot 18 = 180$ $2 \cdot 14 = 28$	2 Sum over $D(f_{CD} \rightarrow g_C)$
$ 1 0 \\ 1 1 $	3 .	-	100	1 0 1 1	$3 \cdot 18 = 54$ $9 \cdot 14 = 126$	3 Multiply f_{BC} with g_C giving h_{BC}
		-				• 4 Sum over C $(h_{BC} ightarrow g_B)$
B C	$f_{BC}(b,c)$	В	$g_B(b)$	B C	$h_{BC}(b,c)$	5 Multiply f_{AB} with g_B giving h_{AB}
0 0 0 1	1 3	0	18 14	$\begin{array}{c} 0 & 0 \\ 0 & 1 \end{array}$	$\begin{array}{c} 1 \cdot 6 = 6 \\ 3 \cdot 4 = 12 \end{array}$	6 Sum over $B(h_{AB} \rightarrow g_A)$
1 0 1 1	1.	1	14	$\begin{array}{cc} 1 & 0 \\ 1 & 1 \end{array}$	$1 \cdot 6 = 6$ $2 \cdot 4 = 8$	7 Sum over $A(g_A \rightarrow \emptyset)$
		-			2 1 - 0	
СD	$f_{CD}(c,d)$ -	- C				
0 0	4 _	C	$g_C(c)$			
0 1	2	0	6			
1 0	1	1	4			
	3 "					



Query: $Z = \sum_{a,b,c,d} f(a, b, c, d) =?$

						_
ΑB	$f_{AB}(a, b)$	А	$g_A(a)$	ΑB	$h_{AB}(a, b)$	1 Pick order: D, C, B, A
00	10 2	0 1	208 180	$\begin{array}{c} 0 & 0 \\ 0 & 1 \end{array}$	$10 \cdot 18 = 180$ $2 \cdot 14 = 28$	2 Sum over $D(f_{CD} \rightarrow g_C)$
1 0	3	-	100	1 0	$3 \cdot 18 = 54$	3 Multiply f_{BC} with g_C giving h_{BC}
	9	-		1 1	9 · 14 = 126	- 4 Sum over C ($h_{BC} \rightarrow g_B$)
B C	$f_{BC}(b,c)$	В	$g_B(b)$	B C	$h_{BC}(b,c)$	5 Multiply f_{AB} with g_B giving h_{AB}
0 0 0 1	1 3	0	18 14	0 0 0 1	$\begin{array}{c} 1 \cdot 6 = 6 \\ 3 \cdot 4 = 12 \end{array}$	6 Sum over $B(h_{AB} \rightarrow g_A)$
$\begin{array}{c} 1 & 0 \\ 1 & 1 \end{array}$	1.2	-	17	$\begin{array}{c} 1 & 0 \\ 1 & 1 \end{array}$	$1 \cdot 6 = 6$ $2 \cdot 4 = 8$	7 Sum over $A(g_A \rightarrow \emptyset)$
C D	$f_{CD}(c,d)$	-				8 Just like the Forward
0.0	4	- C	$g_C(c)$			algorithm! $Z = 388$.
0 1	2	0	6			so $\mathbb{P}(0,0,1,1) = {}^{90}\!/z \approx .23$
10	1	1	4			For free: $\mathbb{P}(A = 0) = \frac{208}{388} \approx .54$.
~ -	0	_				▲日▼▲雪▼▲車▼▲車▼ 前 もののの

André Martins (IST)

Lecture 7: Probabilistic Graphical Models

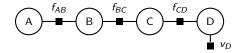
IST, Fall 2020 55 / 66

$$(A) \xrightarrow{f_{AB}} (B) \xrightarrow{f_{BC}} (C) \xrightarrow{f_{CD}} (D)$$

Query: $\mathbb{P}(a, c \mid D = 1) = ?$

ΑB	$f_{AB}(a,b)$
00	10
$\begin{array}{c} 0 & 1 \\ 1 & 0 \end{array}$	2 3
$1 \ 0 \ 1 \ 1 \ 1$	9
	9
ВC	$f_{BC}(b,c)$
0 0	1
0 1	3
1 0	1
1 1	2
C D	$f_{CD}(c,d)$
0 0	4
0 1	2
1 0	1
11	3

Lecture 7: Probabilistic Graphical Models

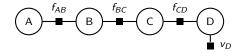


Query: $\mathbb{P}(a, c \mid D = 1) = ?$

		_	
ΑB	$f_{AB}(a, b)$		
0 0 0 1 1 0 1 1	10 2 3 9	- -	
ВC	$f_{BC}(b,c)$	_	
0 0 0 1 1 0 1 1	1 3 1 2		
C D	$f_{CD}(c,d)$	_	
00001	4	D	$v_D(d)$
$\begin{array}{c} 0 \\ 1 \\ 1 \\ 1 \end{array}$	2 1 3	0 1	0 1



André Martins (IST)

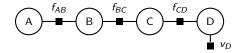


Query: $\mathbb{P}(a, c \mid D = 1) = ?$

		_	
ΑB	$f_{AB}(a,b)$		
$\begin{array}{c} 0 & 0 \\ 0 & 1 \end{array}$	10 2		
1 0	3		
11	9		
1 1	9	-	
ВC	$f_{BC}(b,c)$	_	
0 0	1		
0 1	3		
10	1		
11	2		
		<u>.</u>	
C D	$f_{CD}(c,d)$	_	
00	4	D	$v_D(d)$
$\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}$	2 · 1	0	0
	1	1	1
$1 \ 1$	3	-	1

1 Introduce evidence!

Pick order: D, C, B, A

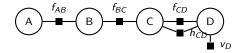


Query: $\mathbb{P}(a, c \mid D = 1) = ?$

ΑB	$f_{AB}(a,b)$	-	
$\begin{array}{c} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{array}$	10 2 3 9	- -	
ВC	$f_{BC}(b,c)$		
0 0 0 1 1 0 1 1	1 3 1 2	- -	
C D	$f_{CD}(c,d)$	_	
$\begin{array}{c} 0 & 0 \\ 0 & 1 \end{array}$	4	D	$v_D(d)$
$ \begin{array}{c} 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{array} $	1 3	0 1	0 1

- Introduce evidence!
- Pick order: D, C, B, A
- 3 Multiply all D factors

IST, Fall 2020 56 / 66

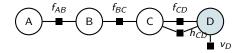


Query: $\mathbb{P}(a, c \mid D = 1) = ?$

А

56 / 66

ΑB	$f_{AB}(a,b)$			0	Introduce evidence!
0 0 0 1 1 0 1 1	10 2 3 9			2	Pick order: D, C, B, A Multiply all <i>D</i> factors
ВC	$f_{BC}(b,c)$:			
0 0 0 1 1 0 1 1	1 3 1 2				
СD	$f_{CD}(c,d)$			C D	$h_{CD}(c,d)$
0 0 0 1 1 0 1 1	4 2 - 1 3	D 0 1	v _D (d) 0 1	0 0 0 1 1 0 1 1	0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 2 0 0 0 2 0 0 2 0 0 2 0
	André Martins	(IST))	Lecture 7: Probabilistic Graphical Models	IST, Fall 2020



Query: $\mathbb{P}(a, c \mid D = 1) = ?$

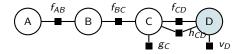
ΑB	$f_{AB}(a,b)$			
$\begin{array}{c} 0 & 0 \\ 0 & 1 \end{array}$	10			
$\begin{array}{c} 0 & 1 \\ 1 & 0 \end{array}$	2 3			
1 1	9			
ВC	$f_{BC}(b,c)$			
0 0	1			
$\begin{array}{c} 0 & 1 \\ 1 & 0 \end{array}$	3			
1 1	2			
СD	$f_{CD}(c,d)$			
00001	4 2 -	D	$v_D(d)$	
1 0	1	0	0	
1 1	3	1	1	
	André Martins	(IST)		Lecture 7: Probabilistic Graphic

Introduce evidence!

- 2 Pick order: D, C, B, A
- **3** Multiply all *D* factors

4 Sum over $D(h_{CD} \rightarrow g_C)$

$f_{CD}(c,d)$			C	D	$h_{CD}(c,d)$	-		
4	D	$v_D(d)$	0	0	0	-		
∠ — 1	0	0	0	1	2			
1	1	1	1		0			
	-	±	1	1 ▶	₽ <≣3	< ≣ > □	2	500
André Martins	(IST)		Lecture 7: Probabilistic Graphical Models		IST,	Fall 2020		56 / 66



Query: $\mathbb{P}(a, c \mid D = 1) = ?$

ΑB	$f_{AB}(a,b)$	-	
0 0	10	-	
$\begin{array}{c} 0 & 1 \\ 1 & 0 \end{array}$	2 3		
$1 \ 1 \ 1$	9	_	
ВC	$f_{BC}(b,c)$	-	
00001	1 3	С	$g_C(c)$
$\begin{array}{c} 0 \\ 1 \\ 1 \\ 1 \end{array}$	1 2	0 1	2 3
СD	$f_{CD}(c,d)$		
00001	4	D	$v_D(d)$
1 0	1	0	0
1 1	3	1	1

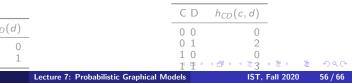
André Martins (IST)

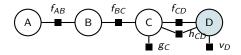
Introduce evidence!

Pick order: D, C, B, A

3 Multiply all D factors

4 Sum over $D(h_{CD} \rightarrow g_C)$



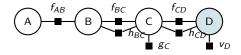


Query: $\mathbb{P}(a, c \mid D = 1) = ?$

ΑB	$f_{AB}(a,b)$		
0 0	10	-	
01	2		
$\begin{array}{c} 1 & 0 \\ 1 & 1 \end{array}$	3		
1 1	9		
ВC	$f_{BC}(b,c)$	_	
$\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}$	1 3	С	$g_C(c)$
10	1	0	2
	-	0 1	2 3
1 0	1	-	
1 0 1 1 C D 0 0	$\frac{1}{f_{CD}(c,d)}$	-	
1 0 1 1 C D 0 0 0 1	$\frac{1}{f_{CD}(c,d)}$	1	3
1 0 1 1 C D 0 0	$\frac{1}{f_{CD}(c,d)}$	1 D	3 v _D (d)

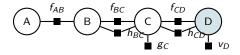
- Introduce evidence!
- 2 Pick order: D, C, B, A
- In Multiply all D factors
- 4 Sum over $D(h_{CD} \rightarrow g_C)$
- **5** Multiply all C factors





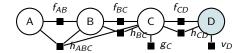
Query: $\mathbb{P}(a, c \mid D = 1) = ?$

ΑB	$f_{AB}(a, b)$					0	Introduce evidence!
$\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}$	10 2					2	Pick order: D, C, B, A
10 1	3					3	Multiply all D factors
	9						Sum over D $(h_{CD} \rightarrow g_C)$
ВC	$f_{BC}(b,c)$					_	
0 0	1	С	$g_C(c)$			6	Multiply all C factors
1 0	1	0	2				
1 1	2	1	3				
СD	$f_{CD}(c,d)$	-		ВC	$h_{BC}(b,c)$	C D	$h_{CD}(c,d)$
0 0	4	D	$v_D(d)$	0 0	2	0.0	0
0 1	2 -	-	0(1)	0 1	9	0 0	2
1 0	1	0	0	1 0	2	1 0	2
1 1	3	1	1	1 1	6	. 1 T	▶ < @ ▶ < E3 < E ▶ E のQ@
	André Martins	(IST))	Lecture 7: Pro	babilistic Graphica	I Models	IST, Fall 2020 56 / 66



Query: $\mathbb{P}(a, c \mid D = 1) = ?$

	André Martins	(IST)		Lecture 7: Prob	abilistic Graphical	Models	IST, Fall 2020	56 / 66
0 0 0 1 1 0 1 1	4 2 - 1 3	D 0 1	$v_D(d)$ 0 1	$\begin{array}{ccc} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{array}$	2 9 2 6	0 0 0 1 1 0 1 1	0 2 0	E 990
СD	$f_{CD}(c,d)$			ВC	$h_{BC}(b,c)$	СD	$h_{CD}(c,d)$	
0 0 0 1 1 0 1 1	1 3 1 2	C 0 1	g _C (c) 2 3			-	lultiply all <i>C</i> facto lultiply all <i>B</i> facto	
ВC	$f_{BC}(b,c)$					-	um over <i>D</i> (<i>h_{CD}</i> –	-
$\begin{array}{c} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{array}$	10 2 3 9					3 M	ick order: D, C, B Iultiply all <i>D</i> facto	ors
AB	<i>f_{AB}(a, b)</i>					1 In	troduce evidence!	



Query: $\mathbb{P}(a, c \mid D = 1) = ?$

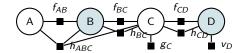
ΑB	$f_{AB}(a, b)$	_		A	ВС	$h_{ABC}(a, b, c)$	1 Introduce evidence!
0 0 0 1 1 0 1 1	10 2 3 9			0 0 0 0	1 0 1 1	20 90 4 12	2 Pick order: D, C, B, A3 Multiply all D factors
ВC	$f_{BC}(b,c)$	-		1	$\begin{array}{c} 0 & 0 \\ 0 & 1 \end{array}$	6 18	4 Sum over $D(h_{CD} \rightarrow g_C)$
0 0 0 1 1 0 1 1	1 3 1 2	C 0 1	g _C (c) 2 3	1	1 0 1 1	18 54	 Multiply all C factors Multiply all B factors
СD	$f_{CD}(c,d)$	_			ВС	$h_{BC}(b,c)$	$C D h_{CD}(c, d)$
0 0 0 1 1 0 1 1	4 2 - 1 3	D 0 1	v _D (d) 0 1	_	0 0 0 1 1 0 1 1	2 9 2 6	$\begin{array}{c} \hline 0 & 0 & 0 \\ 0 & 1 & 2 \\ 1 & 0 & 0 \\ 1 & 1 & \diamond & 0 \\ \end{array}$
	Andró Martine	(IST)	Locture	7. Drob	bilistic Craphical Mo	IST Fall 2020 56 / 6

André Martins (IST)

=

Lecture 7: Probabilistic Graphical Models

IST, Fall 2020 56



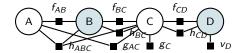
Query: $\mathbb{P}(a, c \mid D = 1) = ?$

A B $f_{AB}(a, b)$	_	A B C	$h_{ABC}(a, b, c)$	1 Introduce evidence!
0 0 10 0 1 2		$\begin{smallmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{smallmatrix}$	20 90	2 Pick order: D, C, B, A
10 3 11 9		$\begin{smallmatrix}&0&1&0\\&0&1&1\end{smallmatrix}$	4 12	3 Multiply all <i>D</i> factors
B C $f_{BC}(b,c)$	-	$1 0 0 \\ 1 0 1$	6 18	4 Sum over $D(h_{CD} \rightarrow g_C)$
0 0 1	C g _C (c)	1 1 0	18	5 Multiply all <i>C</i> factors
0 1 3	$\frac{c}{0}$ $\frac{g(c)}{2}$	111	54	. 6 Multiply all <i>B</i> factors
1 0 1 1 1 2	1 3			7 Sum over <i>B</i> .
$C D f_{CD}(c, d)$		ВC	$h_{BC}(b,c)$	$C D h_{CD}(c, d)$
0 0 4	D $v_D(d)$	0 0 0 1	2	0 0 0
	0 0	1 0	2	0 1 2 1 0 0
1 1 3				11 민 · · · · · · · · · · · · · · · · · ·

André Martins (IST)

Lecture 7: Probabilistic Graphical Models

IST, Fall 2020 56 / 66



Query: $\mathbb{P}(a, c \mid D = 1) = ?$

ΑB	$f_{AB}(a, b)$	A C g _{AC} (a, c) A B C	$h_{ABC}(a, b, c)$	1 Introduce evidence!
$\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}$	10 2	0 0 2 0 1 10		20 90	2 Pick order: D, C, B, A
$\begin{array}{ccc} 1 & 0 \\ 1 & 1 \end{array}$	3 9	1 0 2 1 1 7		4 12	3 Multiply all <i>D</i> factors
ВC	$f_{BC}(b,c)$	_		6 18 18	4 Sum over $D (h_{CD} \rightarrow g_C)$ 5 Multiply all C factors
0 0 0 1	1 3	$C g_C(c)$		54	6 Multiply all B factors
$\begin{array}{c} 1 & 0 \\ 1 & 1 \end{array}$	1	$ \begin{array}{ccc} 0 & 2\\ 1 & 3 \end{array} $			7 Sum over <i>B</i> .
СD	$f_{CD}(c,d)$		ВC	$h_{BC}(b,c)$	$C D h_{CD}(c, d)$
0 0 0 1 1 0 1 1	4 2 1 3	$ \begin{array}{c c} D & v_D(d) \\ \hline 0 & 0 \\ 1 & 1 \end{array} $	- 0 0 - 0 1 1 0 1 1	2 9 2 6	0 0 0 0 1 2 1 0 0 1 ⊕ + < ≥3 < ≥ + ≥ - ∞ <
	Andrá Martin		Laster 7. Duck	abilistic Graphical Mo	<u> </u>

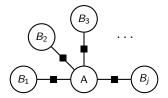
André Martins (IST)

Lecture 7: Probabilistic Graphical Models

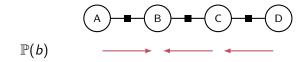
IST, Fall 2020 5

Variable elimination

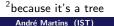
- Answer any query involving max, marginalization, evidence!
- Complexity depends on elimination order: $O(nk^M)$
 - where *n*=n. variables, *k*=dimension, *M*=size of largest intermediate factor.
 - Example: In chain, intuitive order has M = 2. eliminating from middle of chain gives M = 3.
 - Extreme example is a star graph. Best case M = 2, worst M = N!



- In chains and trees: optimal order is easy. Not in general.
- When given a new query, need to restart algorithm from scratch!

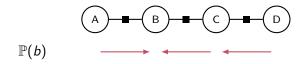


• Optimal order: A, D, C (or D, C, A)



Lecture 7: Probabilistic Graphical Models

IST, Fall 2020 58 / 66

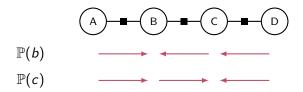


- Optimal order: A, D, C (or D, C, A)
- At each step, we eliminate a variable Y by multiplying (at most²) two factors and summing over Y:

$$g_{Y \to X}(x) = \sum_{y} f_{XY}(x, y) g_Y(y)$$

²because it's a tree

André Martins (IST)

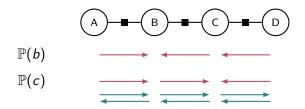


- Optimal order: A, D, C (or D, C, A)
- At each step, we eliminate a variable Y by multiplying (at most²) two factors and summing over Y:

$$g_{Y \to X}(x) = \sum_{y} f_{XY}(x, y) g_Y(y)$$

• These intermediate operations ("messages") are shared for all queries,

²because it's a tree



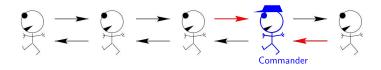
- Optimal order: A, D, C (or D, C, A)
- At each step, we eliminate a variable Y by multiplying (at most²) two factors and summing over Y:

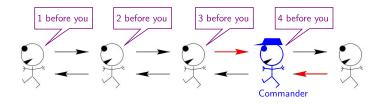
$$g_{Y \to X}(x) = \sum_{y} f_{XY}(x, y) g_Y(y)$$

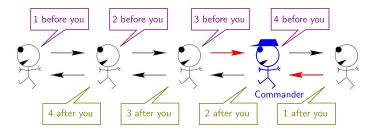
 These intermediate operations ("messages") are shared for all queries, so let's compute all messages up front!

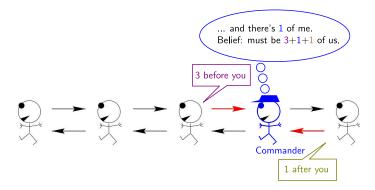
²because it's a tree

André Martins (IST)

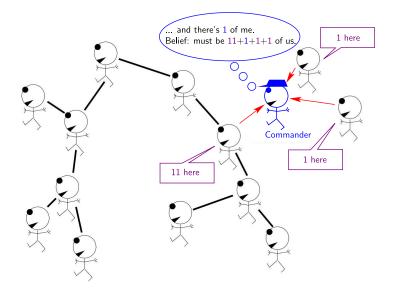








Motivating Example: Counting Soldiers



(Adapted from MacKay 2003 and Gormley & Eisner ACL'14 tutorial.) 📑 🕤

André Martins (IST)

Lecture 7: Probabilistic Graphical Models

Message passing in a tree FG

• Messages from variable X to factor α: aggregate variable beliefs from any other factors. (For leaves, this message is 1).

$$u_{\mathbf{X}\to\alpha}(\mathbf{x}) = \prod_{\beta\in\mathbb{N}(\mathbf{X})-\alpha} \mu_{\beta\to\mathbf{X}}(\mathbf{x})$$

Messages from factor α to variable X: marginalizes over all assignments y₁,..., y_k for Y₁,..., Y_k neighboring α

$$\mu_{\alpha \to X}(x) = \sum_{\substack{y_1, \dots, y_k \\ \{Y_1, \dots, Y_k\} = \mathcal{N}(\alpha) - X}} f_\alpha(x, y_1, \dots, y_k) \prod_{\substack{Y_i \in \mathcal{N}(\alpha) - X}} \nu_{Y_i \to \alpha}(y_i)$$

- A message is sent once all messages it depends on have been received.
- For chain: forward-backward! For tree: leaves-to-root and back.
- If new evidence is added, many messages don't change.
- Replace sum with max for maximization.

From messages to beliefs

- Once we collected all the messages, we can compute local beliefs.
- Variable beliefs:

$$p_X(x) \propto \prod_{lpha \in \mathcal{N}(X)} \mu_{lpha o X}(x)$$

Factor beliefs:

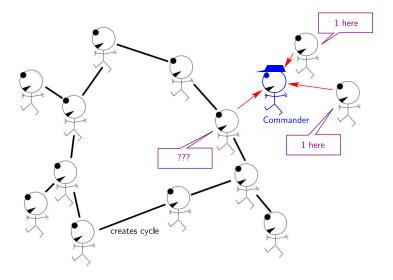
$$p_{\alpha}(x_1,\ldots x_k) \propto f_{\alpha}(x_1,\ldots,x_k) \prod_{X_i \in \mathcal{N}(\alpha)} \nu_{X_i \to \alpha}(x_i)$$

If no cycles, once all messages are passed, beliefs are true marginals:

$$p_X(x) = \mathbb{P}(x), \qquad p_\alpha(x_1,\ldots,x_k) = \mathbb{P}(x_1,\ldots,x_k).$$

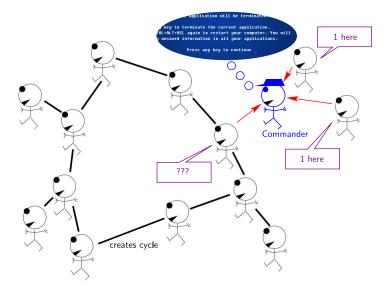
What to do if there are cycles?

Counting Soldiers with Loops



Lecture 7: Probabilistic Graphical Models

Counting Soldiers with Loops



Lecture 7: Probabilistic Graphical Models

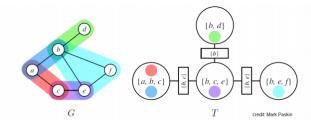
-

< □ > < 同 > < 回 >

62 / 66

Inference in loopy graphs

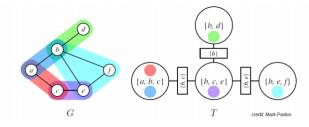
- Exact solution: **Junction Tree** algorithm:
 - convert the graph into a tree, by merging cliques!



- Complexity: like variable elimination. Finding the best tree is NP-hard. (corresponds to finding an ordering for variable elimination.)
- Better than VE because we get all marginals at once.

Inference in loopy graphs

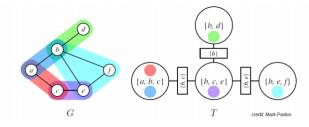
- Exact solution: Junction Tree algorithm:
 - convert the graph into a tree, by merging cliques!



- Complexity: like variable elimination. Finding the best tree is NP-hard. (corresponds to finding an ordering for variable elimination.)
- Better than VE because we get all marginals at once.
- Approximate solution: Loopy Belief Propagation:
 - initialize all messages;
 - pass messages in some order until convergence.
 - (may not terminate, result not guaranteed correct, but works ok.)

Inference in loopy graphs

- Exact solution: **Junction Tree** algorithm:
 - convert the graph into a tree, by merging cliques!



- Complexity: like variable elimination. Finding the best tree is NP-hard. (corresponds to finding an ordering for variable elimination.)
- Better than VE because we get all marginals at once.
- Approximate solution: Loopy Belief Propagation:
 - initialize all messages;
 - pass messages in some order until convergence.
 - (may not terminate, result not guaranteed correct, but works ok.)
 - Many recent algorithms (early 2010s).

IST, Fall 2020 63 / 66

Above, we took the factor scores for granted. We can learn to model them:

Above, we took the factor scores for granted. We can learn to model them: Use some model (neural or feature-based) to produce **unary scores**:

 $f_{\mathcal{A}}(y) = \exp s_{\mathcal{A},y} = (\text{for example}) \exp w_{\mathcal{A},y} \cdot \phi_{\mathcal{A}}(x)$

and pairwise scores:

$$f_{AB}(y, y') = \exp s_{AB,y,y'} = (\text{for example}) \exp w_{A,B,y,y'} \cdot \phi_{A,B}(x)$$

Above, we took the factor scores for granted. We can learn to model them: Use some model (neural or feature-based) to produce **unary scores**:

 $f_{\mathcal{A}}(y) = \exp s_{\mathcal{A},y} = (\text{for example}) \exp w_{\mathcal{A},y} \cdot \phi_{\mathcal{A}}(x)$

and pairwise scores:

 $f_{AB}(y, y') = \exp s_{AB,y,y'} = (\text{for example}) \exp w_{A,B,y,y'} \cdot \phi_{A,B}(x)$ (In general, factor scores $f_{\alpha}(y_{\alpha}) = \exp s_{\alpha,y_{\alpha}}$)

Above, we took the factor scores for granted. We can learn to model them: Use some model (neural or feature-based) to produce **unary scores**:

 $f_{\mathcal{A}}(y) = \exp s_{\mathcal{A},y} = (\text{for example}) \exp w_{\mathcal{A},y} \cdot \phi_{\mathcal{A}}(x)$

and pairwise scores:

 $f_{AB}(y, y') = \exp s_{AB,y,y'} = (\text{for example}) \exp w_{A,B,y,y'} \cdot \phi_{A,B}(x)$ (In general, factor scores $f_{\alpha}(y_{\alpha}) = \exp s_{\alpha,y_{\alpha}}$) The probability of an entire labeling y is then $\prod_{\alpha} f_{\alpha}(y_{\alpha}) = \exp s_{\alpha,y_{\alpha}}$

$$\mathbb{P}(y \mid x) = \frac{\prod_{\alpha} I_{\alpha}(y_{\alpha})}{Z}$$
 meaning $\log \mathbb{P}(y \mid x) = \sum_{\alpha} s_{\alpha, y_{\alpha}} - \log Z$

Above, we took the factor scores for granted. We can learn to model them: Use some model (neural or feature-based) to produce **unary scores**:

 $f_{\mathcal{A}}(y) = \exp s_{\mathcal{A},y} = (\text{for example}) \exp w_{\mathcal{A},y} \cdot \phi_{\mathcal{A}}(x)$

and pairwise scores:

 $f_{AB}(y, y') = \exp s_{AB,y,y'} = (\text{for example}) \exp w_{A,B,y,y'} \cdot \phi_{A,B}(x)$ (In general, factor scores $f_{\alpha}(y_{\alpha}) = \exp s_{\alpha,y_{\alpha}}$) The probability of an entire labeling y is then $\mathbb{T}(x + y) = \prod_{\alpha} f_{\alpha}(y_{\alpha}) \exp s_{\alpha,y_{\alpha}} = \exp s_{\alpha,y_{\alpha}}$

$$\mathbb{P}(y \mid x) = \frac{\prod_{\alpha} r_{\alpha}(y_{\alpha})}{Z}$$
 meaning $\log \mathbb{P}(y \mid x) = \sum_{\alpha} s_{\alpha, y_{\alpha}} - \log Z$

Gradient updates wrt a factor's scores:

$$\frac{\partial \log \mathbb{P}(y \mid x)}{\partial s_{\alpha, y_{\alpha}}} = [[y_{\alpha} = y_{\alpha}^{\mathsf{true}}]] - \mathbb{P}(y_{\alpha} \mid x)$$

Above, we took the factor scores for granted. We can learn to model them: Use some model (neural or feature-based) to produce **unary scores**:

 $f_{\mathcal{A}}(y) = \exp s_{\mathcal{A},y} = (\text{for example}) \exp w_{\mathcal{A},y} \cdot \phi_{\mathcal{A}}(x)$

and pairwise scores:

 $f_{AB}(y, y') = \exp s_{AB,y,y'} = (\text{for example}) \exp w_{A,B,y,y'} \cdot \phi_{A,B}(x)$ (In general, factor scores $f_{\alpha}(y_{\alpha}) = \exp s_{\alpha,y_{\alpha}}$) The probability of an entire labeling y is then $\prod f(y_{\alpha})$

$$\mathbb{P}(y \mid x) = \frac{\prod_{\alpha} t_{\alpha}(y_{\alpha})}{Z} \quad \text{meaning} \quad \log \mathbb{P}(y \mid x) = \sum_{\alpha} s_{\alpha, y_{\alpha}} - \log Z$$

Gradient updates wrt a factor's scores:

$$\frac{\partial \log \mathbb{P}(y \mid x)}{\partial s_{\alpha, y_{\alpha}}} = [[y_{\alpha} = y_{\alpha}^{\mathsf{true}}]] - \mathbb{P}(y_{\alpha} \mid x)$$

The updates use the factor beliefs $\mathbb{P}(y_{\alpha} \mid x) = p_{\alpha}(y_{\alpha})$ for each factor!

- MRFs and pairwise MRFs, both special cases of FGs.
- Powerful, expressive, widely used for discriminative modelling.
- Exact inference when not loopy.
 - We've seen some ideas of what to do when loopy
 - We did not cover more advanced approaches, relating message passing and dual decomposition: (Martins et al., 2015; Kolmogorov, 2006; Komodakis et al., 2007; Globerson and Jaakkola, 2007)
- For learning: a generalization of linear-chain CRFs

References I

Dawid, A. P. (2010). Beware of the DAG! In Causality: objectives and assessment, pages 59-86.

- Globerson, A. and Jaakkola, T. (2007). Fixing Max-Product: Convergent message passing algorithms for MAP LP-relaxations. In Proc. of NeurIPS.
- Koller, D. and Friedman, N. (2009). Probabilistic Graphical Models: Principles and Techniques. The MIT Press.
- Kolmogorov, V. (2006). Convergent Tree-Reweighted Message Passing for energy minimization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(10):1568–1583.
- Komodakis, N., Paragios, N., and Tziritas, G. (2007). MRF optimization via dual decomposition: Message-Passing revisited. In *Proc. of ICCV*.
- MacKay, D. (2003). Information Theory, Inference, and Learning Algorithms, volume 7. Cambridge University Press.
- Martins, A. F., Figueiredo, M. A., Aguiar, P. M., Smith, N. A., and Xing, E. P. (2015). AD3: Alternating directions dual decomposition for MAP inference in graphical models. *JMLR*, 16(1):495–545.
- Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann.
- Pearl, J. (2000). Causality: models, reasoning and inference, volume 29. Springer.
- Pearl, J. (2012). The do-calculus revisited. arXiv preprint arXiv:1210.4852.

< ロ > < 同 > < 三 > < 三 >