Lecture 7: Probabilistic Graphical Models

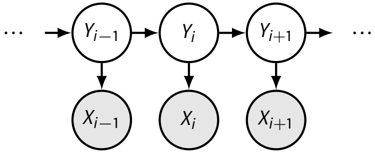
Vlad Niculae & André Martins

Deep Structured Learning Course, Fall 2019

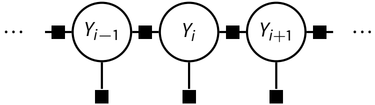
Graphical Models

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

Directed
(today)

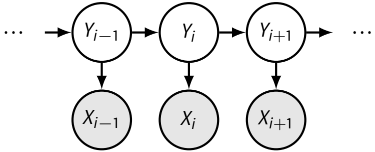


Undirected
(last time)

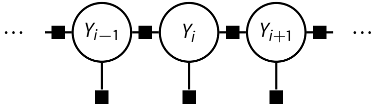


In this unit, we will formalize & extend these graphical representations encountered in previous lectures.

Directed
(last time)



Undirected
(today)



Outline

1 Directed Models

Bayes networks

Conditional independence and D-separation

Causal graphs & the *do* operator

2 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

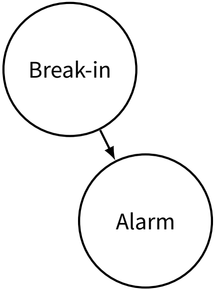
Factor graphs

Bayes (belief) networks

- 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?

Bayes (belief) networks

- 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?



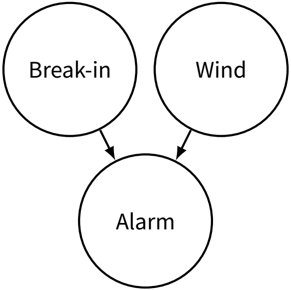
P(B)	B=yes	B=no
	.05	.95

P(A | B)	A=on	A=off
B=yes	.99	.01
B=no	.10	.90

- $\mathrm{P}(B \mid A) = ?$

Bayes (belief) networks

- 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?



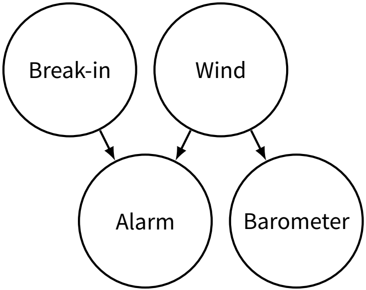
P(B)	B=yes	B=no
	.05	.95

P(A | B, W)		A=on	A=off
B=yes	W=lo	.99	.01
B=yes	W=med	.99	.01
B=yes	W=hi	.999	.001
B=no	W=lo	.01	.99
B=no	W=med	.05	.95
B=no	W=hi	.25	.75

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

Bayes (belief) networks

- 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?



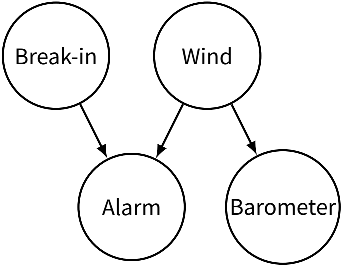
P(B)	B=yes	B=no
	.05	.95

P(A | B, W)		A=on	A=off
B=yes	W=lo	.99	.01
B=yes	W=med	.99	.01
B=yes	W=hi	.999	.001
B=no	W=lo	.01	.99
B=no	W=med	.05	.95
B=no	W=hi	.25	.75

- $\mathrm{P}(B \mid A) = ?$
- Can we observe wind? $\mathrm{P}(B \mid A, W) = ?$
  Maybe we're in the basement, but have a barometer.

Bayes networks

Toolkit for encoding knowledge about
interaction structures between random variables.



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

$$\boxed{\text{In general:} \quad \mathrm{P}(X_1, \ldots, X_n) = \prod_i \mathrm{P}\left(X_i \mid \text{parents}(X_i)\right)}$$

For example: $\quad \mathrm{P}(\text{Break-in}, \text{Wind}, \text{Alarm}, \text{Barometer})$

$$= \mathrm{P}(\text{Break-in})\,\mathrm{P}(\text{Wind})\,\mathrm{P}(\text{Alarm} \mid \text{Break-in}, \text{Wind})\,\mathrm{P}(\text{Barometer} \mid \text{Wind})$$

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

Brk.	Wind	Alarm	Bar.	P
yes	lo	on	lo	0.0243
yes	lo	on	med	0.0002
yes	lo	on	hi	0.0002
yes	lo	off	lo	0.0002
yes	lo	off	med	2.50e-06
yes	lo	off	hi	2.50e-06
yes	med	on	lo	0.0001
yes	med	on	med	0.0146
yes	med	on	hi	0.0001
yes	med	off	lo	1.50e-06
yes	med	off	med	0.0001
yes	med	off	hi	1.50e-06
yes	hi	on	lo	9.99e-05
yes	hi	on	med	9.99e-05
yes	hi	on	hi	0.0098
yes	hi	off	lo	1.00e-07
yes	hi	off	med	1.00e-07
yes	hi	off	hi	9.80e-06

Brk.	Wind	Alarm	Bar.	P
no	lo	on	lo	0.0047
no	lo	on	med	4.75e-05
no	lo	on	hi	4.75e-05
no	lo	off	lo	0.4608
no	lo	off	med	0.0047
no	lo	off	hi	0.0047
no	med	on	lo	0.0001
no	med	on	med	0.0140
no	med	on	hi	0.0001
no	med	off	lo	0.0027
no	med	off	med	0.2653
no	med	off	hi	0.0027
no	hi	on	lo	0.0005
no	hi	on	med	0.0005
no	hi	on	hi	0.0466
no	hi	off	lo	0.0014
no	hi	off	med	0.0014
no	hi	off	hi	0.1397

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

Brk.	Wind	Alarm	Bar.	P
yes	lo	on	lo	0.0243
yes	lo	on	med	0.0002
yes	lo	on	hi	0.0002
yes	lo	off	lo	0.0002
yes	lo	off	med	2.50e-06
yes	lo	off	hi	2.50e-06
yes	med	on	lo	0.0001
yes	med	on	med	0.0146
yes	med	on	hi	0.0001
yes	med	off	lo	1.50e-06
yes	med	off	med	0.0001
yes	med	off	hi	1.50e-06
yes	hi	on	lo	9.99e-05
yes	hi	on	med	9.99e-05
yes	hi	on	hi	0.0098
yes	hi	off	lo	1.00e-07
yes	hi	off	med	1.00e-07
yes	hi	off	hi	9.80e-06

Brk.	Wind	Alarm	Bar.	P
no	lo	on	lo	0.0047
no	lo	on	med	4.75e-05
no	lo	on	hi	4.75e-05
no	lo	off	lo	0.4608
no	lo	off	med	0.0047
no	lo	off	hi	0.0047
no	med	on	lo	0.0001
no	med	on	med	0.0140
no	med	on	hi	0.0001
no	med	off	lo	0.0027
no	med	off	med	0.2653
no	med	off	hi	0.0027
no	hi	on	lo	0.0005
no	hi	on	med	0.0005
no	hi	on	hi	0.0466
no	hi	off	lo	0.0014
no	hi	off	med	0.0014
no	hi	off	hi	0.1397

P(Break-in=yes, Alarm=on) $= 0.0496$

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

Brk.	Wind	Alarm	Bar.	P
yes	lo	on	lo	0.0243
yes	lo	on	med	0.0002
yes	lo	on	hi	0.0002
yes	lo	off	lo	0.0002
yes	lo	off	med	2.50e-06
yes	lo	off	hi	2.50e-06
yes	med	on	lo	0.0001
yes	med	on	med	0.0146
yes	med	on	hi	0.0001
yes	med	off	lo	1.50e-06
yes	med	off	med	0.0001
yes	med	off	hi	1.50e-06
yes	hi	on	lo	9.99e-05
yes	hi	on	med	9.99e-05
yes	hi	on	hi	0.0098
yes	hi	off	lo	1.00e-07
yes	hi	off	med	1.00e-07
yes	hi	off	hi	9.80e-06

Brk.	Wind	Alarm	Bar.	P
no	lo	on	lo	0.0047
no	lo	on	med	4.75e-05
no	lo	on	hi	4.75e-05
no	lo	off	lo	0.4608
no	lo	off	med	0.0047
no	lo	off	hi	0.0047
no	med	on	lo	0.0001
no	med	on	med	0.0140
no	med	on	hi	0.0001
no	med	off	lo	0.0027
no	med	off	med	0.2653
no	med	off	hi	0.0027
no	hi	on	lo	0.0005
no	hi	on	med	0.0005
no	hi	on	hi	0.0466
no	hi	off	lo	0.0014
no	hi	off	med	0.0014
no	hi	off	hi	0.1397

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

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

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

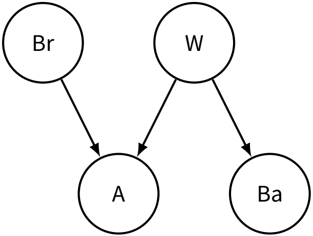
Brk.	Wind	Alarm	Bar.	P
yes	lo	on	lo	0.0243
yes	lo	on	med	0.0002
yes	lo	on	hi	0.0002
yes	lo	off	lo	0.0002
yes	lo	off	med	2.50e-06
yes	lo	off	hi	2.50e-06
yes	med	on	lo	0.0001
yes	med	on	med	0.0146
yes	med	on	hi	0.0001
yes	med	off	lo	1.50e-06
yes	med	off	med	0.0001
yes	med	off	hi	1.50e-06
yes	hi	on	lo	9.99e-05
yes	hi	on	med	9.99e-05
yes	hi	on	hi	0.0098
yes	hi	off	lo	1.00e-07
yes	hi	off	med	1.00e-07
yes	hi	off	hi	9.80e-06

Brk.	Wind	Alarm	Bar.	P
no	lo	on	lo	0.0047
no	lo	on	med	4.75e-05
no	lo	on	hi	4.75e-05
no	lo	off	lo	0.4608
no	lo	off	med	0.0047
no	lo	off	hi	0.0047
no	med	on	lo	0.0001
no	med	on	med	0.0140
no	med	on	hi	0.0001
no	med	off	lo	0.0027
no	med	off	med	0.2653
no	med	off	hi	0.0027
no	hi	on	lo	0.0005
no	hi	on	med	0.0005
no	hi	on	hi	0.0466
no	hi	off	lo	0.0014
no	hi	off	med	0.0014
no	hi	off	hi	0.1397

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

$$\mathrm{P}(\text{Break-in=yes} \mid \text{Alarm=on}) = \frac{\mathrm{P}(\text{Break-in=yes}, \text{Alarm=on})}{\sum_b \mathrm{P}(\text{Break-in=}b, \text{Alarm=on})}$$
$$= .427$$

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



$$P(\mathrm{Br}, \mathrm{W}, \mathrm{A}, \mathrm{Ba})$$
$$= P(\mathrm{Br})\, P(\mathrm{W})\, P(\mathrm{A} \mid \mathrm{Br}, \mathrm{W})\, P(\mathrm{Ba} \mid \mathrm{W})$$

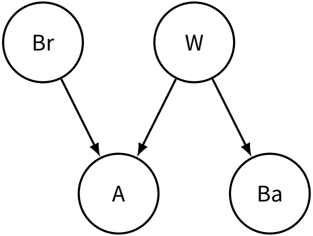
P(Br)	yes	no
	.05	.95

P(W)	lo	mid	hi
	.5	.3	.2

P(A | 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 | 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.



$$P(\mathrm{Br}, \mathrm{W}, \mathrm{A}, \mathrm{Ba})$$
$$= P(\mathrm{Br})\, P(\mathrm{W})\, P(\mathrm{A} \mid \mathrm{Br}, \mathrm{W})\, P(\mathrm{Ba} \mid \mathrm{W})$$

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

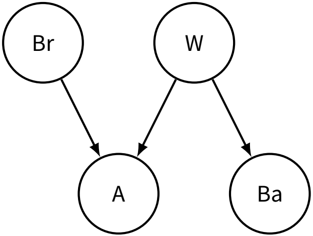
P(Br)	yes	no
	.05	.95

P(W)	lo	mid	hi
	.5	.3	.2

P(A | 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 | 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.



$$P(\text{Br}, \text{W}, \text{A}, \text{Ba})$$
$$= P(\text{Br})\, P(\text{W})\, P(\text{A} \mid \text{Br}, \text{W})\, P(\text{Ba} \mid \text{W})$$

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

P(Br)	yes	no
	.05	.95

P(W)	lo	mid	hi
	.5	.3	.2

P(A | 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 | W)	lo	mid	hi
W=lo	.98	.01	.01
W=mid	.01	.98	.01
W=hi	.01	.01	.98

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

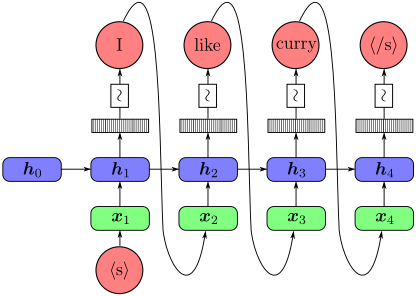
encode structure, not parametrization

are non-unique for a distribution

encode independence **requirements**, not necessarily all

BN are not neural net diagrams

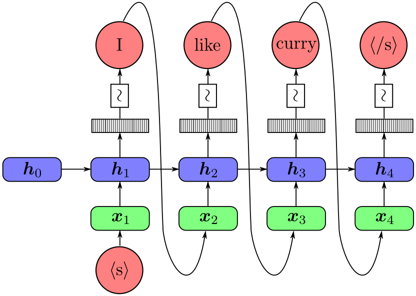
Recall the RNN language model:



- In statistical terms, what are we modeling?

BN are not neural net diagrams

Recall the RNN language model:

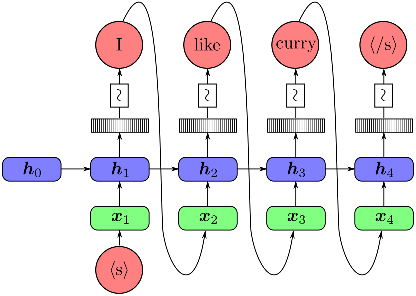


- In statistical terms, what are we modeling?

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

BN are not neural net diagrams

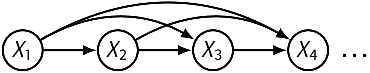
Recall the RNN language model:



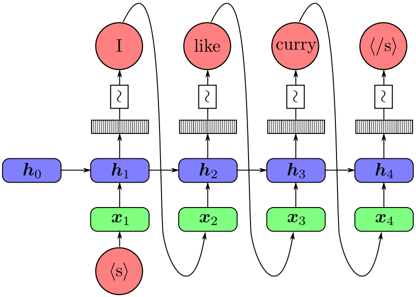
- In statistical terms, what are we modeling?

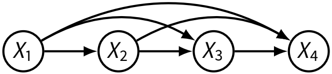
$$P(X_1, \ldots, X_n) = P(X_1)\, P(X_2 \mid X_1)\, P(X_3 \mid X_1, X_2) \ldots$$

- 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)

BN encode structure, not parametrization

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, $Y \in \{\texttt{sports}, \texttt{music}, \ldots\}$.

BN for a generative model: $Y \rightarrow X$

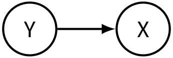
We must posit what are $\mathrm{P}(Y)$ and $\mathrm{P}(X \mid Y)$. Many possible options!

BN encode structure, not parametrization

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, $Y \in \{\mathtt{sports}, \mathtt{music}, \ldots\}$.

BN for a generative model:



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

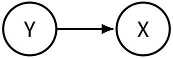
$\mathrm{P}(Y)$: uniform: $\mathrm{P}(Y = \mathtt{sports}) = \mathrm{P}(Y = \mathtt{music}) = \frac{1}{|\mathcal{Y}|}$,

BN encode structure, not parametrization

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, $Y \in \{\texttt{sports}, \texttt{music}, \ldots\}$.

BN for a generative model:



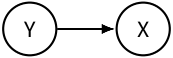
We must posit what are $\mathrm{P}(Y)$ and $\mathrm{P}(X \mid Y)$. Many possible options!

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

BN encode structure, not parametrization

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, $Y \in \{\texttt{sports}, \texttt{music}, \dots\}$.

BN for a generative model: 

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

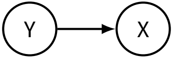
$\mathrm{P}(Y)$: uniform: $\mathrm{P}(Y = \texttt{sports}) = \mathrm{P}(Y = \texttt{music}) = \frac{1}{|\mathcal{Y}|}$, or estimated from data.

$\mathrm{P}(X \mid Y)$ (remember: values of X are sentences)

BN encode structure, not parametrization

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, $Y \in \{\texttt{sports}, \texttt{music}, \ldots\}$.

BN for a generative model: 

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

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

$\mathrm{P}(X \mid Y)$ (remember: values of X are sentences)

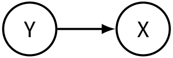
Naive Bayes $\mathrm{P}(X \mid Y) = \prod_{j=1}^{L} \mathrm{P}(X_j \mid Y)$

BN encode structure, not parametrization

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, $Y \in \{\texttt{sports}, \texttt{music}, \ldots\}$.

BN for a generative model:



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

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

$\mathrm{P}(X \mid Y)$ (remember: values of X are sentences)

Naive Bayes	$\mathrm{P}(X \mid Y) = \prod_{j=1}^{L} \mathrm{P}(X_j \mid Y)$
Per-class Markov language model	$\mathrm{P}(X \mid Y) = \prod_{j=1}^{L} \mathrm{P}(X_j \mid X_{j-1}, Y)$

BN encode structure, not parametrization

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, $Y \in \{\texttt{sports}, \texttt{music}, \ldots\}$.

BN for a generative model: $\textsf{Y} \rightarrow \textsf{X}$

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

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

$\mathrm{P}(X \mid Y)$ (remember: values of X are sentences)

Naive Bayes	$\mathrm{P}(X \mid Y) = \prod_{j=1}^{L} \mathrm{P}(X_j \mid Y)$
Per-class Markov language model	$\mathrm{P}(X \mid Y) = \prod_{j=1}^{L} \mathrm{P}(X_j \mid X_{j-1}, Y)$
Per-class recurrent NN language model	$\mathrm{P}(X \mid Y) = \mathrm{LSTM}(x_1, \ldots, x_L; \boldsymbol{w}_y)$

BN encode structure, not parametrization

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, $Y \in \{\texttt{sports}, \texttt{music}, \ldots\}$.

BN for a generative model: $Y \rightarrow X$

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

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

$\mathrm{P}(X \mid Y)$ (remember: values of X are sentences)

Naive Bayes	$\mathrm{P}(X \mid Y) = \prod_{j=1}^{L} \mathrm{P}(X_j \mid Y)$
Per-class Markov language model	$\mathrm{P}(X \mid Y) = \prod_{j=1}^{L} \mathrm{P}(X_j \mid X_{j-1}, Y)$
Per-class recurrent NN language model	$\mathrm{P}(X \mid Y) = \mathrm{LSTM}(x_1, \ldots, x_L; \boldsymbol{w}_y)$

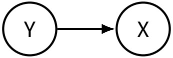
$\mathrm{P}(X \mid Y)$ need not be parametrized as a table.

BN encode structure, not parametrization

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, $Y \in \{\texttt{sports}, \texttt{music}, \dots\}$.

BN for a generative model:



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

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

$\mathrm{P}(X \mid Y)$ (remember: values of X are sentences)

Naive Bayes	$\mathrm{P}(X \mid Y) = \prod_{j=1}^{L} \mathrm{P}(X_j \mid Y)$
Per-class Markov language model	$\mathrm{P}(X \mid Y) = \prod_{j=1}^{L} \mathrm{P}(X_j \mid X_{j-1}, Y)$
Per-class recurrent NN language model	$\mathrm{P}(X \mid Y) = \mathrm{LSTM}(x_1, \dots, x_L; \boldsymbol{w}_y)$

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

Variables need not be discrete! mixture of Gaussians: $\mathrm{P}(X \mid Y = y) \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$.

Equivalent factorizations

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

Equivalent factorizations

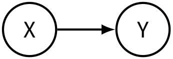
There are many possible factorizations! $\mathrm{P}(X, Y) =$

$$\begin{tikzcd} X \arrow[r] & Y \end{tikzcd}$$

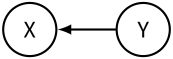
$\mathrm{P}(X)\,\mathrm{P}(Y \mid X)$

Equivalent factorizations

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



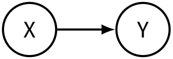
$\mathrm{P}(X)\,\mathrm{P}(Y \mid X)$



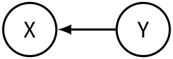
$\mathrm{P}(Y)\,\mathrm{P}(X \mid Y)$

Equivalent factorizations

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



$\mathrm{P}(X)\,\mathrm{P}(Y \mid X)$

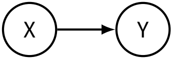


$\mathrm{P}(Y)\,\mathrm{P}(X \mid Y)$

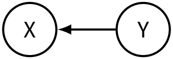
$\mathrm{P}(X)\,\mathrm{P}(Y)$

Equivalent factorizations

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



$\mathrm{P}(X)\,\mathrm{P}(Y \mid X)$



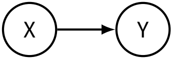
$\mathrm{P}(Y)\,\mathrm{P}(X \mid Y)$

$\mathrm{P}(X)\,\mathrm{P}(Y)$

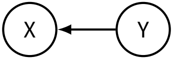
The first two are valid Bayes nets for **any** $\mathrm{P}(X, Y)$!

Equivalent factorizations

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



$\mathrm{P}(X)\,\mathrm{P}(Y \mid X)$



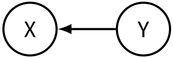
$\mathrm{P}(Y)\,\mathrm{P}(X \mid Y)$

$\mathrm{P}(X)\,\mathrm{P}(Y)$

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

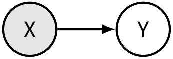
In fact, recall generative vs discriminative classifiers!

- Generative (e.g. naïve Bayes):



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

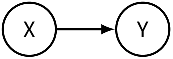
- Discriminative (e.g. logistic regression)



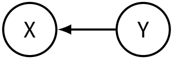
  in LR, we don't model $\mathrm{P}(X)$, we assume X is always observed (gray).

Equivalent factorizations

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



$\mathrm{P}(X)\,\mathrm{P}(Y \mid X)$



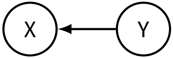
$\mathrm{P}(Y)\,\mathrm{P}(X \mid Y)$

$\mathrm{P}(X)\,\mathrm{P}(Y)$

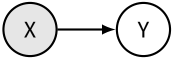
The first two are valid Bayes nets for **any** $\mathrm{P}(X, Y)$!

In fact, recall generative vs discriminative classifiers!

- Generative (e.g. naïve Bayes):

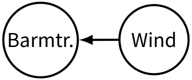
  To classify, we would compute $\mathrm{P}(Y \mid X)$ via Bayes' rule.
- Discriminative (e.g. logistic regression)

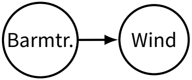
  in LR, we don't model $\mathrm{P}(X)$, we assume X is always observed (gray).

Some arrow direction choices are harder to estimate.

Some make more sense (why?):



vs.



Minimal independence assumptions

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

Bayes net (1): 

Minimal independence assumptions

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

Bayes net (1): 

Example parametrization:

P(X)	A+	A	B	...
	.01	.02	.04	

P(Y)	Jan	Feb	Mar	...
	.10	.12	.09	

Minimal independence assumptions

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

Bayes net (1):

Example parametrization:

P(X)	A+	A	B	...
	.01	.02	.04	

P(Y)	Jan	Feb	Mar	...
	.10	.12	.09	

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

Minimal independence assumptions

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

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

Bayes net (2): $X \leftarrow Y$

Example parametrization:

P(X)	A+	A	B	...
	.01	.02	.04	

P(Y)	Jan	Feb	Mar	...
	.10	.12	.09	

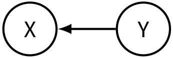
BN (1) imposes $X \perp\!\!\!\perp Y$
in **any parametrization**.

Does it mean we *must* have $X \not\perp\!\!\!\perp Y$?

Minimal independence assumptions

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

Bayes net (1): 

Bayes net (2): 

Example parametrization:

P(X)	A+	A	B	...
	.01	.02	.04	

P(Y)	Jan	Feb	Mar	...
	.10	.12	.09	

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

Does it mean we *must* have $X \not\perp\!\!\!\perp Y$? **NO!**

P(Y)	Jan	Feb	Mar	...
	.10	.12	.09	

$P(X \mid Y)$	A+	A	B	...
Y=Jan	.01	.02	.04	
Y=Feb	.01	.02	.04	
Y=Mar	.01	.02	.04	
...				

Minimal independence assumptions

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

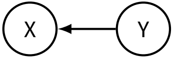
Bayes net (1): 

Example parametrization:

P(X)	A+	A	B	...
	.01	.02	.04	

P(Y)	Jan	Feb	Mar	...
	.10	.12	.09	

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

Bayes net (2): 

Does it mean we *must* have $X \not\perp\!\!\!\perp Y$? **NO!**

P(Y)	Jan	Feb	Mar	...
	.10	.12	.09	

$P(X \mid Y)$	A+	A	B	...
Y=Jan	.01	.02	.04	
Y=Feb	.01	.02	.04	
Y=Mar	.01	.02	.04	
...				

A BN constraints what independences **must be** in the model **as a minimum**.

Outline

1. **Directed Models**
   - Bayes networks
   - Conditional independence and D-separation
   - Causal graphs & the *do* operator
2. **Undirected Models**
   - 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)

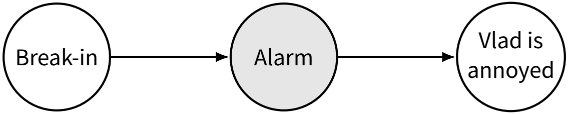
A is independent of B given a set of variables $C = \{C_1, \ldots, C_n\}$, denoted as

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

if and only if

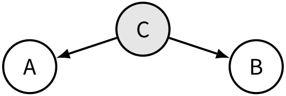
$$\mathrm{P}(A, B \mid C_1, \ldots, C_n) = \mathrm{P}(A \mid C_1, \ldots, C_n)\,\mathrm{P}(B \mid C_1, \ldots, C_n).$$

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

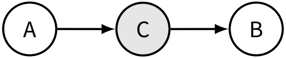


Three fundamental relationships in BN

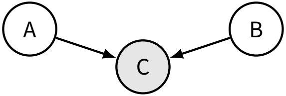
The Fork



The Chain

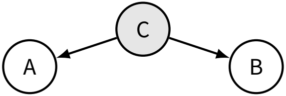


The Collider

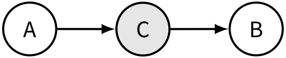


Three fundamental relationships in BN

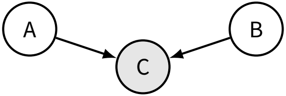
The Fork



The Chain



The Collider



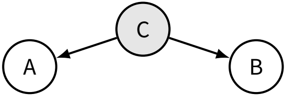
$A \perp\!\!\!\perp B \mid C$

Given C, A and B are independent.

Example: Alarm $\leftarrow$ Wind $\rightarrow$ Barometer

Three fundamental relationships in BN

The Fork

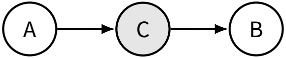


$A \perp\!\!\!\perp B \mid C$

Given C, A and B are independent.

Example: Alarm $\leftarrow$ Wind $\rightarrow$ Barometer

The Chain

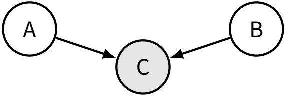


$A \perp\!\!\!\perp B \mid C$

After observing C,
further observing A would not tell us about B.

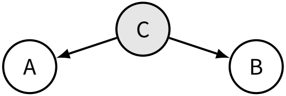
Example: Burglary $\rightarrow$ Alarm $\rightarrow$ Vlad distracted

The Collider



Three fundamental relationships in BN

The Fork

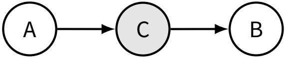


$A \perp\!\!\!\perp B \mid C$

Given C, A and B are independent.

Example: Alarm $\leftarrow$ Wind $\rightarrow$ Barometer

The Chain



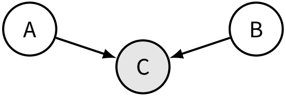
$A \perp\!\!\!\perp B \mid C$

After observing C,

further observing A would not tell us about B.

Example: Burglary $\rightarrow$ Alarm $\rightarrow$ Vlad distracted

The Collider



Surprisingly, $A \perp\!\!\!\perp B$

but **not** $A \perp\!\!\!\perp B \mid C$!

Example: Burglary $\rightarrow$ Alarm $\leftarrow$ Wind

Burglaries occur regardless how windy it is.

If alarm rings, hearing wind makes burglary **less likely!**

Burglary is “explained away” by wind.

Detecting independence: d-separation

Algorithm for deciding if A and B are **d-separated** given set C, implying:

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

For all paths P from A to B in the **skeleton**[1] of the BN, at least one holds:

[1] skeleton = the graph with undirected edges replacing the directed arcs

Detecting independence: d-separation

Algorithm for deciding if A and B are **d-separated** given set C, implying:

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

For all paths P from A to B in the **skeleton**[1] of the BN, at least one holds:

1. P includes a fork with observed parent:

$$X \leftarrow C \rightarrow Y \qquad \text{(with } C \in C\text{)}$$

[1] skeleton = the graph with undirected edges replacing the directed arcs

Detecting independence: d-separation

Algorithm for deciding if A and B are **d-separated** given set $\mathcal{C}$, implying:

$$A \perp\!\!\!\perp B \mid \mathcal{C}.$$

For all paths P from A to B in the **skeleton**[1] of the BN, at least one holds:

1. P includes a fork with observed parent:

$$X \leftarrow C \rightarrow Y \qquad \text{(with } C \in \mathcal{C}\text{)}$$

2. P includes a chain with observed middle:

$$X \rightarrow C \rightarrow Y \quad \text{or} \quad X \leftarrow C \leftarrow Y \qquad \text{(with } C \in \mathcal{C}\text{)}$$

[1] skeleton = the graph with undirected edges replacing the directed arcs

Detecting independence: d-separation

Algorithm for deciding if A and B are **d-separated** given set C, implying:

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

For all paths P from A to B in the **skeleton**[1] of the BN, at least one holds:

1. P includes a fork with observed parent:

$$X \leftarrow C \rightarrow Y \qquad \text{(with } C \in C\text{)}$$

2. P includes a chain with observed middle:

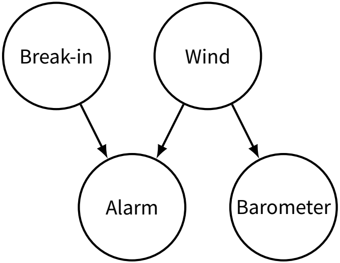
$$X \rightarrow C \rightarrow Y \quad \text{or} \quad X \leftarrow C \leftarrow Y \qquad \text{(with } C \in C\text{)}$$

3. P includes a collider

$$X \rightarrow U \leftarrow Y \qquad \text{(with } U \notin C\text{)}$$

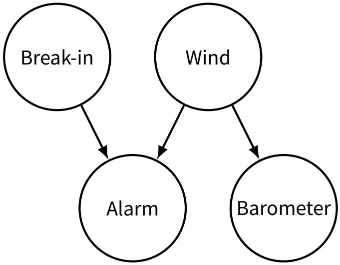
[1] skeleton = the graph with undirected edges replacing the directed arcs

Examples



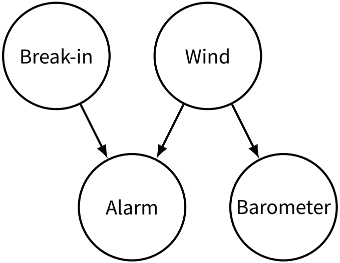
Wind $\perp\!\!\!\perp$ Barometer?

Examples



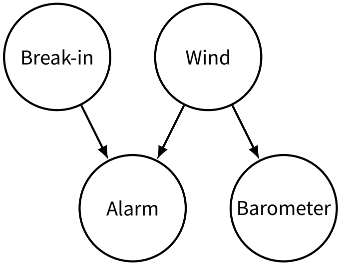
Wind $\perp\!\!\!\perp$ Barometer? **No**

Examples



Wind $\perp\!\!\!\perp$ Barometer? **No**
Break-in $\perp\!\!\!\perp$ Wind?

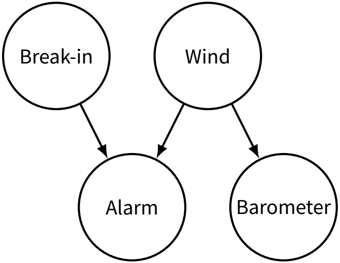
Examples



Wind $\perp\!\!\!\perp$ Barometer? **No**

Break-in $\perp\!\!\!\perp$ Wind? **Yes**

Examples

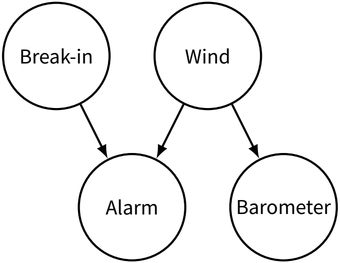


Wind $\perp\!\!\!\perp$ Barometer? **No**

Break-in $\perp\!\!\!\perp$ Wind? **Yes**

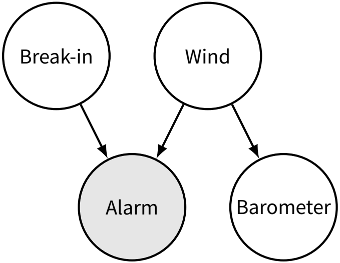
Break-in $\perp\!\!\!\perp$ Barometer?

Examples



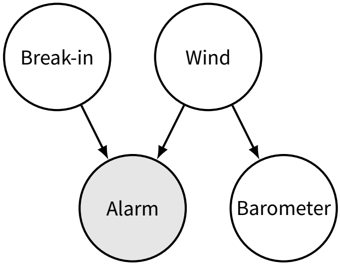
Wind $\perp\!\!\!\perp$ Barometer? **No**
Break-in $\perp\!\!\!\perp$ Wind? **Yes**
Break-in $\perp\!\!\!\perp$ Barometer? **Yes**

Examples



$$\text{Wind} \perp\!\!\!\perp \text{Barometer}? \textbf{ No}$$
$$\text{Break-in} \perp\!\!\!\perp \text{Wind}? \textbf{ Yes}$$
$$\text{Break-in} \perp\!\!\!\perp \text{Barometer}? \textbf{ Yes}$$
$$\text{Break-in} \perp\!\!\!\perp \text{Barometer} \mid \text{Alarm}?$$

Examples



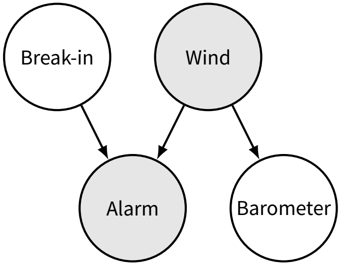
Wind $\perp\!\!\!\perp$ Barometer? **No**

Break-in $\perp\!\!\!\perp$ Wind? **Yes**

Break-in $\perp\!\!\!\perp$ Barometer? **Yes**

Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm? **No**

Examples



Wind $\perp\!\!\!\perp$ Barometer? **No**

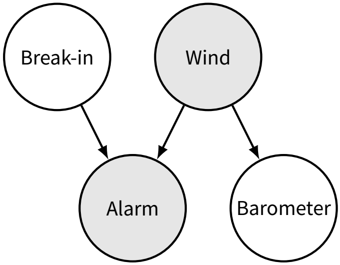
Break-in $\perp\!\!\!\perp$ Wind? **Yes**

Break-in $\perp\!\!\!\perp$ Barometer? **Yes**

Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm? **No**

Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm, Wind?

Examples



Wind $\perp\!\!\!\perp$ Barometer? **No**

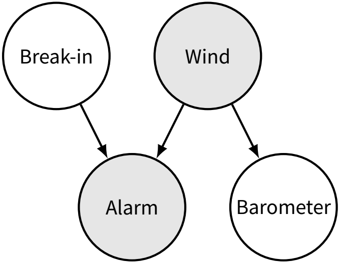
Break-in $\perp\!\!\!\perp$ Wind? **Yes**

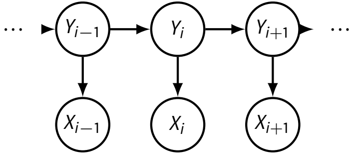
Break-in $\perp\!\!\!\perp$ Barometer? **Yes**

Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm? **No**

Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm, Wind? **Yes**

Examples





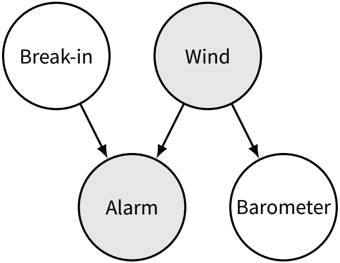
Wind $\perp\!\!\!\perp$ Barometer? **No**

Break-in $\perp\!\!\!\perp$ Wind? **Yes**

Break-in $\perp\!\!\!\perp$ Barometer? **Yes**

Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm? **No**

Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm, Wind? **Yes**



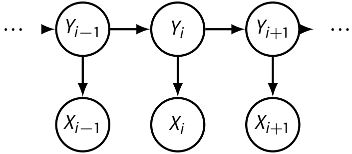
Wind $\perp\!\!\!\perp$ Barometer? **No**

Break-in $\perp\!\!\!\perp$ Wind? **Yes**

Break-in $\perp\!\!\!\perp$ Barometer? **Yes**

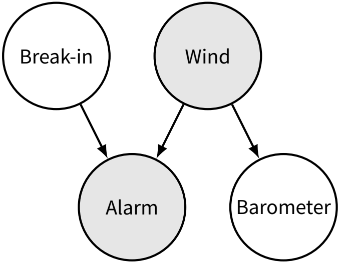
Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm? **No**

Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm, Wind? **Yes**



$$Y_{i+1} \perp\!\!\!\perp Y_{i-1}?$$

Examples



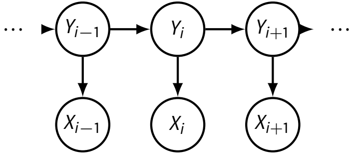
Wind $\perp\!\!\!\perp$ Barometer? **No**

Break-in $\perp\!\!\!\perp$ Wind? **Yes**

Break-in $\perp\!\!\!\perp$ Barometer? **Yes**

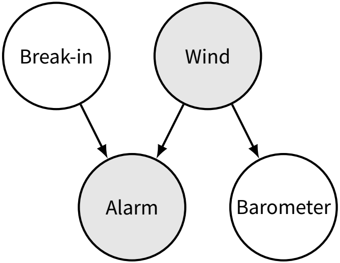
Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm? **No**

Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm, Wind? **Yes**

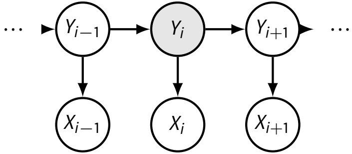


$Y_{i+1} \perp\!\!\!\perp Y_{i-1}$? **No**

Examples

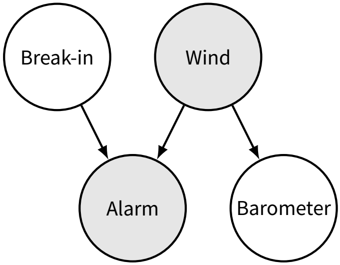


Wind $\perp\!\!\!\perp$ Barometer? **No**
Break-in $\perp\!\!\!\perp$ Wind? **Yes**
Break-in $\perp\!\!\!\perp$ Barometer? **Yes**
Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm? **No**
Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm, Wind? **Yes**



$Y_{i+1} \perp\!\!\!\perp Y_{i-1}$? **No**
$Y_{i+1} \perp\!\!\!\perp Y_{i-1} \mid Y_i$?

Examples



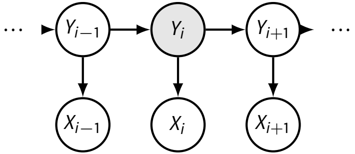
Wind $\perp\!\!\!\perp$ Barometer? **No**

Break-in $\perp\!\!\!\perp$ Wind? **Yes**

Break-in $\perp\!\!\!\perp$ Barometer? **Yes**

Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm? **No**

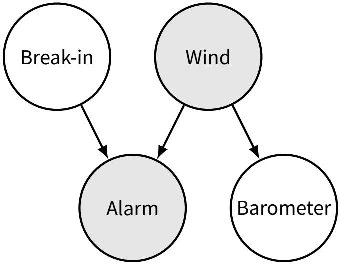
Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm, Wind? **Yes**



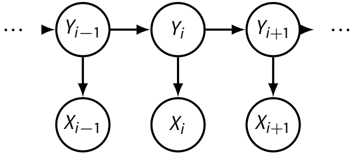
$Y_{i+1} \perp\!\!\!\perp Y_{i-1}$? **No**

$Y_{i+1} \perp\!\!\!\perp Y_{i-1} \mid Y_i$? **Yes**

Examples

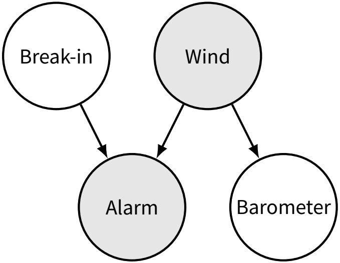


Wind $\perp\!\!\!\perp$ Barometer? **No**
Break-in $\perp\!\!\!\perp$ Wind? **Yes**
Break-in $\perp\!\!\!\perp$ Barometer? **Yes**
Break-in $\perp\!\!\!\perp$ Barometer | Alarm? **No**
Break-in $\perp\!\!\!\perp$ Barometer | Alarm, Wind? **Yes**

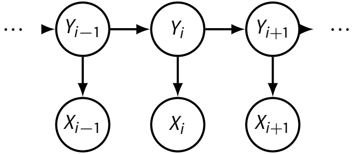


$Y_{i+1} \perp\!\!\!\perp Y_{i-1}$? **No**
$Y_{i+1} \perp\!\!\!\perp Y_{i-1} \mid Y_i$? **Yes**
$Y_{i+1} \perp\!\!\!\perp X_i$?

Examples

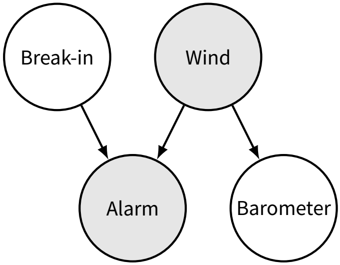


Wind $\perp\!\!\!\perp$ Barometer? **No**
Break-in $\perp\!\!\!\perp$ Wind? **Yes**
Break-in $\perp\!\!\!\perp$ Barometer? **Yes**
Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm? **No**
Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm, Wind? **Yes**



$Y_{i+1} \perp\!\!\!\perp Y_{i-1}$? **No**
$Y_{i+1} \perp\!\!\!\perp Y_{i-1} \mid Y_i$? **Yes**
$Y_{i+1} \perp\!\!\!\perp X_i$? **No**

Examples



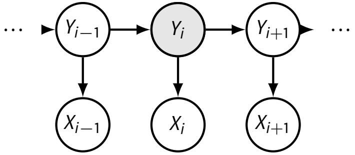
Wind $\perp\!\!\!\perp$ Barometer? **No**

Break-in $\perp\!\!\!\perp$ Wind? **Yes**

Break-in $\perp\!\!\!\perp$ Barometer? **Yes**

Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm? **No**

Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm, Wind? **Yes**



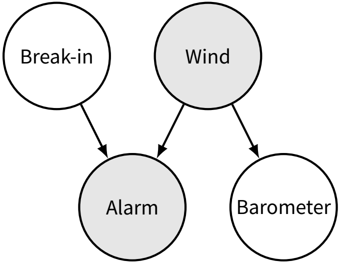
$Y_{i+1} \perp\!\!\!\perp Y_{i-1}$? **No**

$Y_{i+1} \perp\!\!\!\perp Y_{i-1} \mid Y_i$? **Yes**

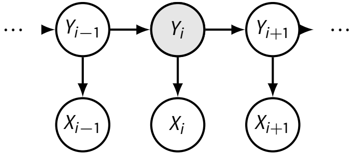
$Y_{i+1} \perp\!\!\!\perp X_i$? **No**

$Y_{i+1} \perp\!\!\!\perp X_i \mid Y_i$?

Examples



Wind $\perp\!\!\!\perp$ Barometer? **No**
Break-in $\perp\!\!\!\perp$ Wind? **Yes**
Break-in $\perp\!\!\!\perp$ Barometer? **Yes**
Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm? **No**
Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm, Wind? **Yes**



$Y_{i+1} \perp\!\!\!\perp Y_{i-1}$? **No**
$Y_{i+1} \perp\!\!\!\perp Y_{i-1} \mid Y_i$? **Yes**
$Y_{i+1} \perp\!\!\!\perp X_i$? **No**
$Y_{i+1} \perp\!\!\!\perp X_i \mid Y_i$? **Yes**

Generative stories and plate notation

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

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

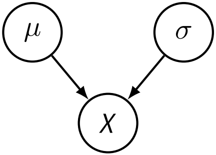
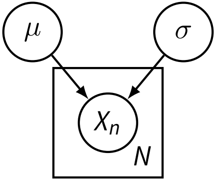


Plate notation is a way to denote **repetition of templates**:

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



Outline

1 Directed Models

Bayes networks

Conditional independence and D-separation

Causal graphs & the *do* operator

2 Undirected Models

Markov random fields

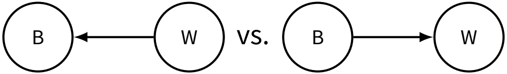
Factor graphs

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

Seeing versus doing

Bayes nets only model independence assumptions.

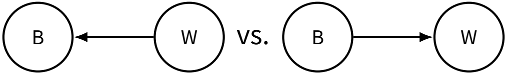
The correlation between the a barometer reading B and wind strength W can be represented either way:



Seeing versus doing

Bayes nets only model independence assumptions.

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



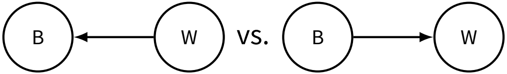
Seeing that the barometer reading is high, we can forecast wind.

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

Seeing versus doing

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.

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

But **setting** the barometer needle to high manually **won't cause wind!**

We write: $\mathrm{P}\left(W \mid \mathrm{do}(B = \mathrm{hi})\right) = ?$

Seeing versus doing

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

Seeing versus doing

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

Two reasons why doing $\neq$ seeing:

- we got the direction wrong
- we missed some confounding factor

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

Seeing versus doing

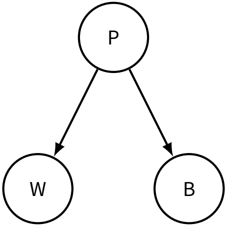
Setting the barometer needle to high manually **won't cause wind!**

Two reasons why doing $\neq$ seeing:

- we got the direction wrong
- 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**.

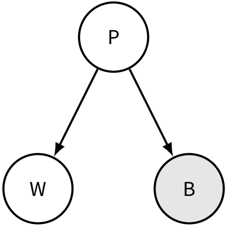
- A causes B only if A is an ancestor of B in $\mathcal{G}$.
- $A \to B$ means A is a direct cause of B.

A good model is essential. Wrong causal assumptions $\to$ 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 versus doing, more rigorously

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



Seeing versus doing, more rigorously

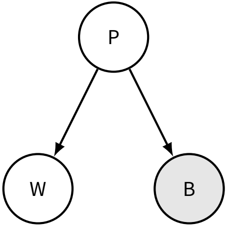
Seeing (*observational*): $\mathrm{P}(W \mid B = \text{hi})$

Measure the world for a while (or call IPMA)

Date	Pressure	Wind	Barometer
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:

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



Seeing versus doing, more rigorously

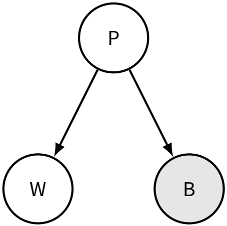
Seeing (*observational*): $\mathrm{P}(W \mid B = \mathrm{hi})$

Measure the world for a while (or call IPMA)

Date	Pressure	Wind	Barometer
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:

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



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

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

Seeing versus doing, more rigorously

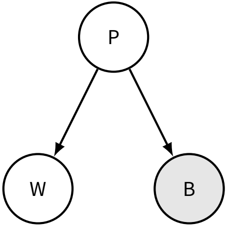
Seeing (*observational*): $\mathrm{P}(W \mid B = \mathrm{hi})$

Measure the world for a while (or call IPMA)

Date	Pressure	Wind	Barometer
1977-01-01	hi	hi	hi
1977-01-02	hi	mid	hi
1977-01-02	mid	mid	mid
...			
2019-11-03	hi	hi	hi

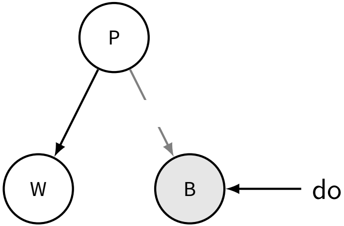
gives:

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



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

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



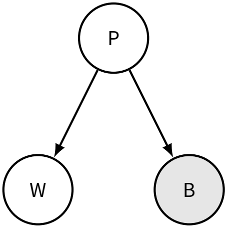
Seeing (*observational*): $\mathrm{P}(W \mid B = \mathrm{hi})$

Measure the world for a while (or call IPMA)

Date	Pressure	Wind	Barometer
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:

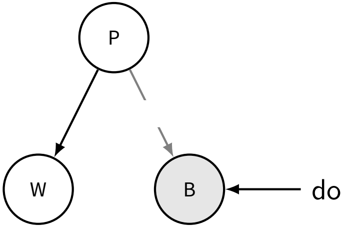
$\mathrm{P}(W \mid B)$	lo	mid	hi
$B = \mathrm{hi}$	.01	.01	.98



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

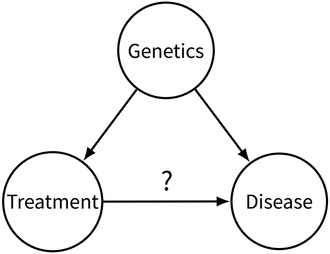
Set the needle to high breaking inbound arrows;
re-generate **new** data in this **new** DAG
(or estimate what that would give.)

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



Randomized controlled trials

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

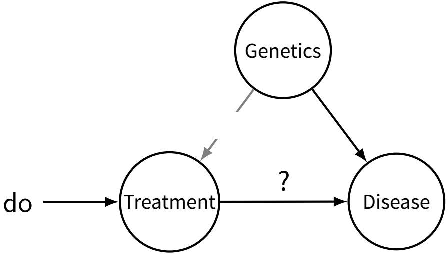


Patient	Treatment	Genetics	Disease
#42	real	?	cured
#68	placebo	?	not cured
...			

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

Randomized controlled trials

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



Patient	Treatment	Genetics	Disease
#42	real	?	cured
#68	placebo	?	not cured
...			

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

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

Pearl's 3 rules

Notation:

X, Y, Z, W	disjoint sets of events (sets of nodes); may be empty
$\mathcal{G}_{\bar{X}}$	the graph with all edges **into** X removed.
$\mathcal{G}_{\underline{X}}$	the graph with all edges **out of** X removed.
$Z(X)$	subset of nodes in Z which are not ancestors of X.
y; $\mathrm{do}(x)$	shorthand for $Y = y$; respectively $\mathrm{do}(X = x)$.

1. Ignoring observations:

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

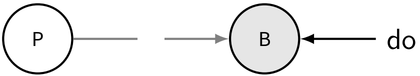
2. Action/observation exchange: the back-door criterion

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

3. Ignoring actions

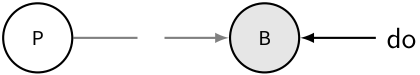
$$\mathrm{P}(y \mid \mathrm{do}(x), \mathrm{do}(z), w) = \mathrm{P}(y \mid \mathrm{do}(x), w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z \mid X, W)_{\mathcal{G}_{\bar{X}, Z(\bar{W})}}$$

Examples 1,2: Pressure and barometer

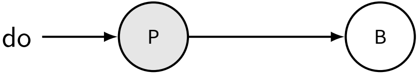


Rule 3: $\mathrm{P}(P = \mathrm{hi} \mid \mathrm{do}(B = hi)) = \mathrm{P}(P = \mathrm{hi})$ since $(P \perp\!\!\!\perp B)_{\mathcal{G}_{\bar{B}}}$

Examples 1,2: Pressure and barometer

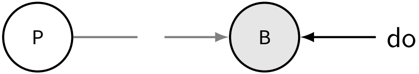


Rule 3: $\mathrm{P}(P = \mathrm{hi} \mid \mathrm{do}(B = hi)) = \mathrm{P}(P = \mathrm{hi})$ since $(P \perp\!\!\!\perp B)_{\mathcal{G}_{\bar{B}}}$

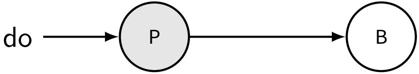


Rule 2: $\mathrm{P}(B = \mathrm{hi} \mid \mathrm{do}(P = \mathrm{lo})) = \mathrm{P}(B = \mathrm{hi} \mid P = \mathrm{lo})$ since $(B \perp\!\!\!\perp P)_{\mathcal{G}_{\underline{P}}}$

Examples 1,2: Pressure and barometer



$$\text{Rule 3:} \qquad \mathrm{P}(P = \mathrm{hi} \mid \mathrm{do}(B = hi)) = \mathrm{P}(P = \mathrm{hi}) \quad \text{since} \quad (P \perp\!\!\!\perp B)_{\mathcal{G}_{\overline{B}}}$$

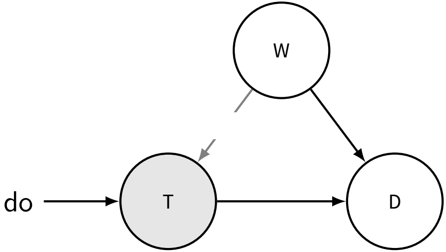


$$\text{Rule 2:} \qquad \mathrm{P}(B = \mathrm{hi} \mid \mathrm{do}(P = \mathrm{lo})) = \mathrm{P}(B = \mathrm{hi} \mid P = \mathrm{lo}) \quad \text{since} \quad (B \perp\!\!\!\perp P)_{\mathcal{G}_{\underline{P}}}$$

Good check: we get the intuitively correct results.

Example 3: Measurable confounder

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



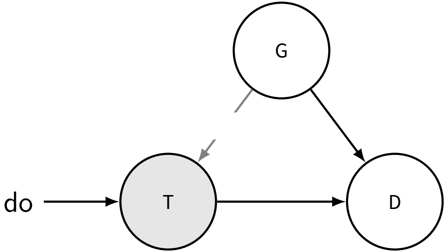
Condition on wealth (which thus needs to be measurable)

$$
\begin{aligned}
\mathrm{P}(D = \text{cured} \mid \mathrm{do}(T = \mathrm{y})) &= \mathrm{P}(D = \text{cured} \mid \mathrm{do}(T = \mathrm{y}), W = \mathrm{y})\, \mathrm{P}(W = \mathrm{y} \mid \mathrm{do}(T = \mathrm{y})) \\
&\quad + \mathrm{P}(D = \text{cured} \mid \mathrm{do}(T = \mathrm{y}), W = \mathrm{n})\, \mathrm{P}(W = \mathrm{n} \mid \mathrm{do}(T = \mathrm{y})) \\
&= \mathrm{P}(D = \text{cured} \mid \mathrm{do}(T = \mathrm{y}), W = \mathrm{y})\, \mathrm{P}(W = \mathrm{y}) \\
&\quad + \mathrm{P}(D = \text{cured} \mid \mathrm{do}(T = \mathrm{y}), W = \mathrm{n})\, \mathrm{P}(W = \mathrm{n}) \qquad \text{(R3)} \\
&= \mathrm{P}(D = \text{cured} \mid T = \mathrm{y}, W = \mathrm{y})\, \mathrm{P}(W = \mathrm{y}) \\
&\quad + \mathrm{P}(D = \text{cured} \mid T = \mathrm{y}, W = \mathrm{n})\, \mathrm{P}(W = \mathrm{n}) \qquad \text{(R2)}
\end{aligned}
$$

Example 3: an impossible one

T: treatment, D: disease.

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

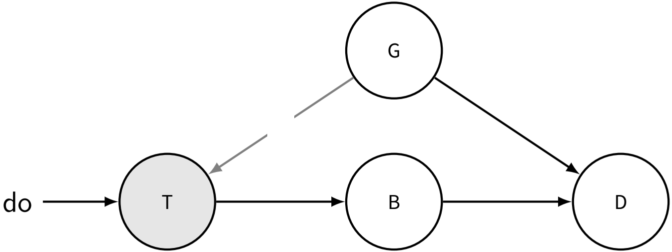


Without more info or more assumptions, we're stuck!

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.)

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

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.

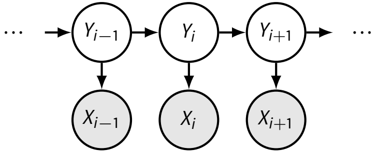
Further reading: (Pearl, 1988; Koller and Friedman, 2009; Pearl, 2000, 2012; Dawid, 2010)

Slides on causal inference and learning causal structure (links):

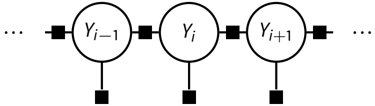
- Sanna Tyrväinen, Introduction to Causal Calculus
- Ricardo Silva, Causality
- Dominik Janzing & Bernhard Schölkopf, Causality

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

Directed
(last time)



Undirected
(today)



Outline

1. **Directed Models**
   - Bayes networks
   - Conditional independence and D-separation
   - Causal graphs & the *do* operator
2. **Undirected Models**
   - Markov random fields
   - Factor graphs

Outline

1. **Directed Models**
   - Bayes networks
   - Conditional independence and D-separation
   - Causal graphs & the *do* operator

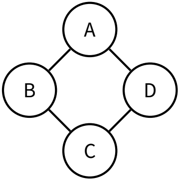
2. **Undirected Models**
   - Markov random fields
   - Factor graphs

Modelling 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.

Modelling 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.

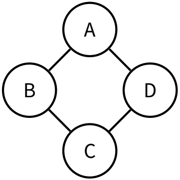
$$\mathrm{P}(a, b, c, d) \propto f(a, b) \cdot f(b, c) \cdot f(c, d) \cdot f(d, a)$$

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

X	Y	f(x,y)
Y	Y	100
Y	N	1
N	Y	1
N	N	100

Modelling 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.

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

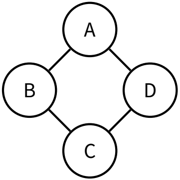
For any $X, Y \in \{A, B, C, D\}$, f is the compatibility function

X	Y	f(x,y)
Y	Y	100
Y	N	1
N	Y	1
N	N	100

- Can we represent this exact factorization in a Bayes net?

Modelling 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.

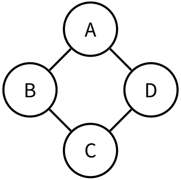
$$\mathrm{P}(a, b, c, d) \propto f(a, b) \cdot f(b, c) \cdot f(c, d) \cdot f(d, a)$$

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

X	Y	f(x,y)
Y	Y	100
Y	N	1
N	Y	1
N	N	100

- Can we represent this exact factorization in a Bayes net? **no!**

Markov random fields



Definition

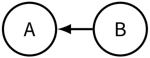
Let $\mathcal{G}$ be an *undirected* graph with nodes corresponding to random variables $X_1, \ldots, X_N$. Let $C(\mathcal{G})$ denote the set of *cliques* (fully connected subgraphs) of $\mathcal{G}$. A MRF is a distribution of the form

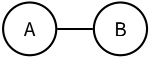
$$\mathrm{P}(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{c \in C} f_c(\boldsymbol{x}_c)$$

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

Any BN can be encoded in a MRF

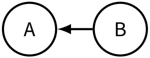
2. Convert all arcs $A \to B$ into undirected edges $A - B$.





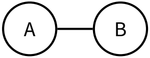
Any BN can be encoded in a MRF

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



A	B	$\mathrm{P}(a \mid b)$
Y	Y	.9
N	Y	.1
Y	N	.1
N	N	.9

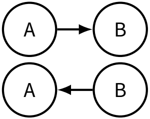
B	$\mathrm{P}(b)$
Y	.75
N	.25

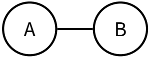


A	B	$f(a, b)$
Y	Y	$.9 \cdot .75$
N	Y	$.1 \cdot .75$
Y	N	$.1 \cdot .25$
N	N	$.9 \cdot .75$

Any BN can be encoded in a MRF

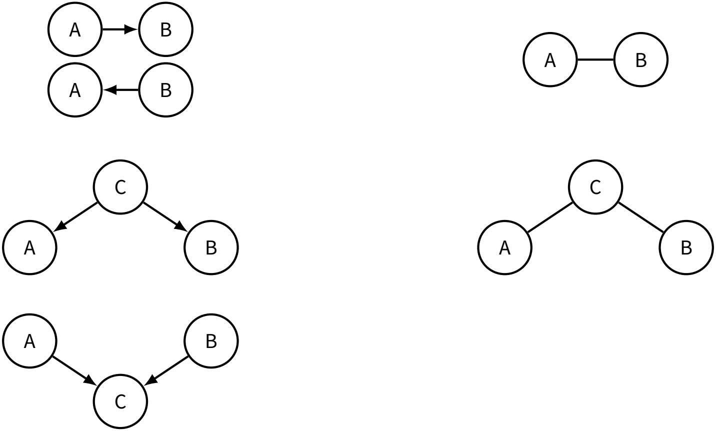
2. Convert all arcs $A \rightarrow B$ into undirected edges $A - B$.





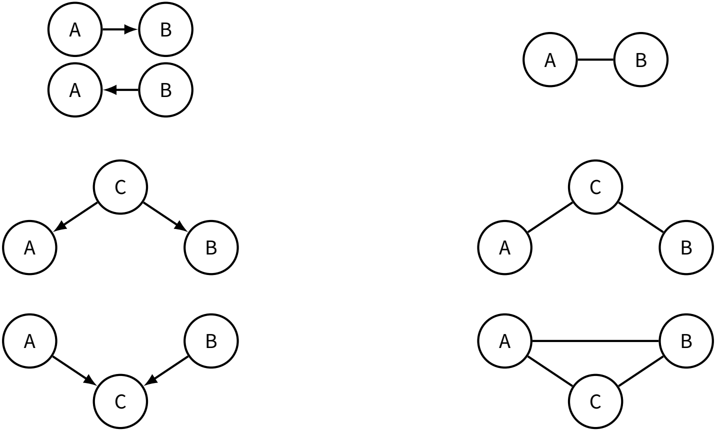
Any BN can be encoded in a MRF

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



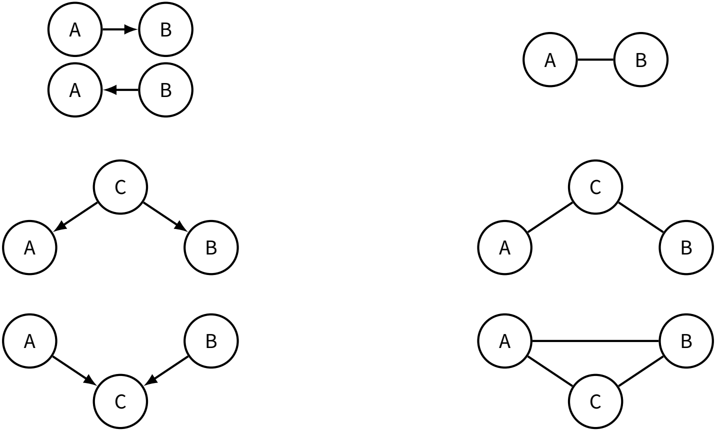
Any BN can be encoded in a MRF

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



Any BN can be encoded in a MRF

1. First, add edge $A - C$ for any collider structure $A \to B \leftarrow C$;
2. Convert all arcs $A \to B$ into undirected edges $A - B$.

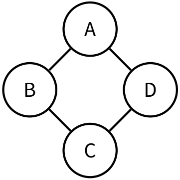


Loose conversion

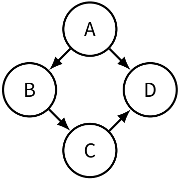
Similarly, we can convert a MRF to a BN (we won't cover it.)

However, **independences may be lost** in either direction.

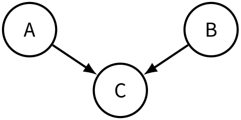
From **To**



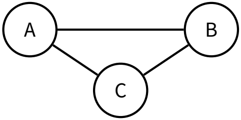
$$A \perp\!\!\!\perp C \mid B, D$$
$$B \perp\!\!\!\perp D \mid A, C$$



$$\cancel{A \perp\!\!\!\perp C \mid B, D}$$
$$B \perp\!\!\!\perp D \mid A, C$$



$$A \perp\!\!\!\perp B$$

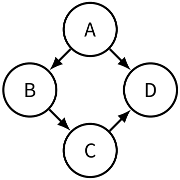


$$\cancel{A \perp\!\!\!\perp B}$$

Bayes vs Markov

Bayes network

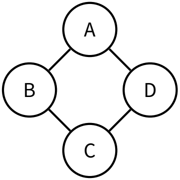
- Factors are conditionals (normalized)
- Easy to sample
- Can be made causal
- Can easily find $\mathrm{P}(x_1, \ldots, x_n)$.



$$\mathrm{P}(a, b, c, d) = \mathrm{P}(a)\,\mathrm{P}(b \mid a)\,\mathrm{P}(c \mid b)\,\mathrm{P}(d \mid a, c)$$

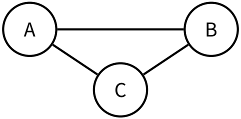
Markov networks

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

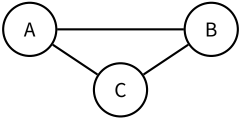


$$\mathrm{P}(a, b, c, d) = {}^{1}\!/_{Z}\, f_1(a, b) f_2(b, c) f_3(c, d) f_4(d, a)$$

What are the factors in a MRF?

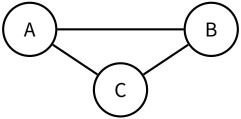


What are the factors in a MRF?



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

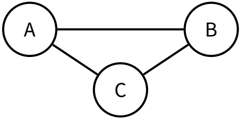
What are the factors in a MRF?



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

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

What are the factors in a MRF?

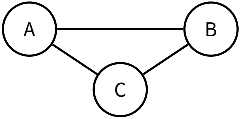


Single clique: $\{A, B, C\}$, so $\mathrm{P}(a, b, c) = \frac{1}{Z} f(a, b, c)$.

No way to represent $\mathrm{P}(a, b, c) = {}^1\!/_Z\, f_1(a, b) f_2(b, c) f_3(c, a)$.

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

What are the factors in a MRF?

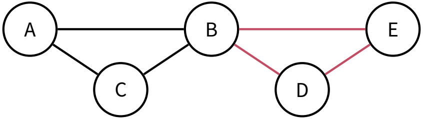


Single clique: $\{A, B, C\}$, so $\mathrm{P}(a, b, c) = \frac{1}{Z} f(a, b, c)$.

No way to represent $\mathrm{P}(a, b, c) = {}^1\!/_Z\, f_1(a, b) f_2(b, c) f_3(c, a)$.

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

But what if we want to mix them?



$$\mathrm{P}(a, b, c, d, e) = {}^1\!/_Z\, f_1(a, b) f_2(b, c) f_3(c, a) f_4(b, d, e)$$

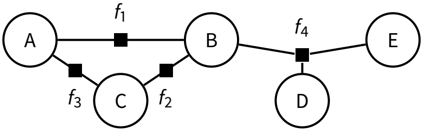
Outline

1. **Directed Models**
   - Bayes networks
   - Conditional independence and D-separation
   - Causal graphs & the *do* operator
2. **Undirected Models**
   - Markov random fields
   - Factor graphs

Factor graphs

Explicitly represent factors in the graph to remove ambiguity.

$$\mathrm{P}(a, b, c, d, e) = {}^1\!/_Z\, f_1(a, b) f_2(b, c) f_3(c, a) f_4(b, d, e)$$



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

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

where $f_\alpha \geq 0$, and $\boldsymbol{X}_\alpha$ 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 fixed compatibility tables, answer **queries**
  - Maximization: Find most likely assignment $x_1, \ldots, x_N$ (possibly given evidence $x_i : i \in \mathcal{E}$).

$$\arg\max_{x_1,\ldots,x_M} \mathrm{P}(x_1, \ldots, x_N \mid \boldsymbol{x}_{\mathcal{E}})$$

  - Marginalization: Find the marginal probability of some partial assignment over $x_j : j \in \mathcal{M}$ (possibly given evidence $x_i : i \in \mathcal{E}$)

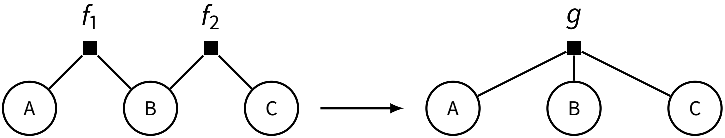
$$\mathrm{P}(\boldsymbol{x}_{\mathcal{M}} \mid \boldsymbol{x}_{\mathcal{E}})$$

  - **NP-hard** / **#P-hard** in general!
- **Learning:** Given a dataset, estimate the compatibility tables (or, in general a model that produces them.)
- Since BN $\rightarrow$ MRF $\rightarrow$ FG, it suffices to study inference algorithms for FG.[2]

[2] But not learning, since we cannot map back to BN losslessly!

Multiplying factors

A core operation: combining factors by multipliying them.



A	B	$f_1(a,b)$
0	0	3
0	1	1
1	0	2
1	1	8

B	C	$f_2(a,b)$
0	0	5
0	1	4
1	0	1
1	1	1

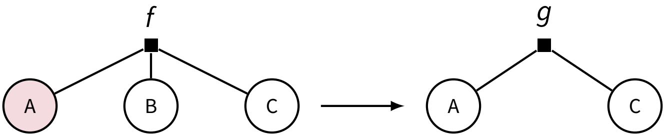
$\rightarrow$

A	B	C	$g(a,b,c)$
0	0	0	$3 \cdot 5 = 15$
0	0	1	$3 \cdot 4 = 12$
0	1	0	$1 \cdot 1 = 1$
0	1	1	$1 \cdot 1 = 1$
1	0	0	$2 \cdot 5 = 10$
1	0	1	$2 \cdot 4 = 8$
1	1	0	$8 \cdot 1 = 8$
1	1	1	$8 \cdot 1 = 8$

Distribution is preserved:

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

Maximizing over a variable



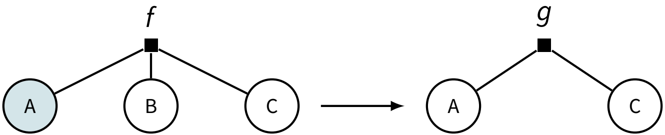
A	B	C	$f(a,b,c)$
0	0	0	15
0	0	1	12
0	1	0	1
0	1	1	1
1	0	0	10
1	0	1	8
1	1	0	8
1	1	1	8

— maximizing over A $\rightarrow$

B	C	$g(b,c)$
0	0	15
0	1	12
1	0	8
1	1	8

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

Marginalizing over a variable



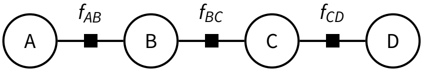
A	B	C	$f(a,b,c)$
0	0	0	15
0	0	1	12
0	1	0	1
0	1	1	1
1	0	0	10
1	0	1	8
1	1	0	8
1	1	1	8

— summing over A $\rightarrow$

B	C	$g(b,c)$
0	0	25
0	1	20
1	0	9
1	1	9

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

Variable elimination



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

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
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

Variable elimination

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

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

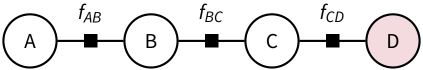
A B	$f_{AB}(a, b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b, c)$
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

Variable elimination



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

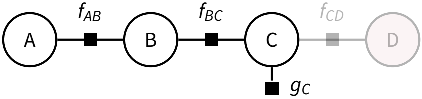
A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
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
2. Maximize over D ($f_{CD} \to g_C$)

Variable elimination



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

A B	$f_{AB}(a, b)$
0 0	10
0 1	2
1 0	3
1 1	9

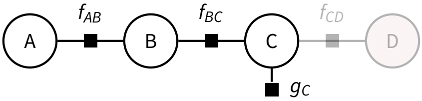
B C	$f_{BC}(b, c)$
0 0	1
0 1	3
1 0	1
1 1	2

C D	$f_{CD}(c, d)$
0 0	4
0 1	2
1 0	1
1 1	3

C	$g_C(c)$
0	$4^{D=0}$
1	$3^{D=1}$

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

Variable elimination



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

A B	$f_{AB}(a, b)$
0 0	10
0 1	2
1 0	3
1 1	9

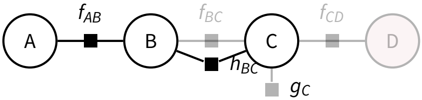
B C	$f_{BC}(b, c)$
0 0	1
0 1	3
1 0	1
1 1	2

C D	$f_{CD}(c, d)$
0 0	4
0 1	2
1 0	1
1 1	3

C	$g_C(c)$
0	$4^{D=0}$
1	$3^{D=1}$

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}

Variable elimination



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

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
0 0	1
0 1	3
1 0	1
1 1	2

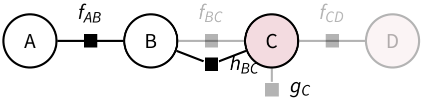
C D	$f_{CD}(c,d)$
0 0	4
0 1	2
1 0	1
1 1	3

C	$g_C(c)$
0	$4^{D=0}$
1	$3^{D=1}$

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

1. Pick order: D, C, B, A
2. Maximize over D ($f_{CD} \to g_C$)
3. Multiply f_{BC} with g_C giving h_{BC}

Variable elimination



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

A B	$f_{AB}(a, b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b, c)$
0 0	1
0 1	3
1 0	1
1 1	2

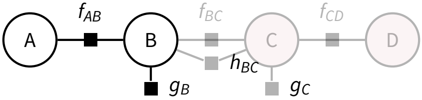
C D	$f_{CD}(c, d)$
0 0	4
0 1	2
1 0	1
1 1	3

C	$g_C(c)$
0	$4^{D=0}$
1	$3^{D=1}$

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

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

Variable elimination



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

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
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

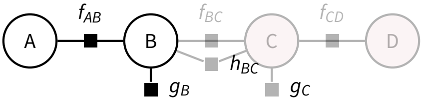
B	$g_B(b)$
0	$9^{C=1}$
1	$6^{C=1}$

C	$g_C(c)$
0	$4^{D=0}$
1	$3^{D=1}$

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

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

Variable elimination



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

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
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

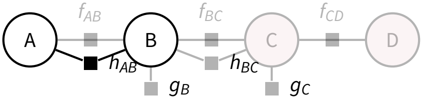
B	$g_B(b)$
0	$9^{C=1}$
1	$6^{C=1}$

C	$g_C(c)$
0	$4^{D=0}$
1	$3^{D=1}$

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

Variable elimination



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

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
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

B	$g_B(b)$
0	$9^{C=1}$
1	$6^{C=1}$

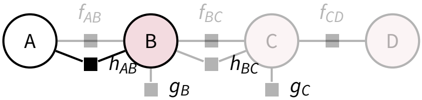
C	$g_C(c)$
0	$4^{D=0}$
1	$3^{D=1}$

A B	$h_{AB}(a,b)$
0 0	$10 \cdot 9 = 90^{C=1}$
0 1	$2 \cdot 6 = 12^{C=1}$
1 0	$3 \cdot 9 = 27^{C=1}$
1 1	$9 \cdot 6 = 54^{C=1}$

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

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

Variable elimination



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

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
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

B	$g_B(b)$
0	$9^{C=1}$
1	$6^{C=1}$

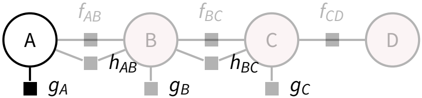
C	$g_C(c)$
0	$4^{D=0}$
1	$3^{D=1}$

A B	$h_{AB}(a,b)$
0 0	$10 \cdot 9 = 90^{C=1}$
0 1	$2 \cdot 6 = 12^{C=1}$
1 0	$3 \cdot 9 = 27^{C=1}$
1 1	$9 \cdot 6 = 54^{C=1}$

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

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

Variable elimination



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

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
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

A	$g_A(a)$
0	$90^{B=0}$
1	$54^{B=1}$

B	$g_B(b)$
0	$9^{C=1}$
1	$6^{C=1}$

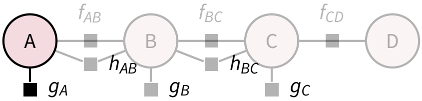
C	$g_C(c)$
0	$4^{D=0}$
1	$3^{D=1}$

A B	$h_{AB}(a,b)$
0 0	$10 \cdot 9 = 90^{C=1}$
0 1	$2 \cdot 6 = 12^{C=1}$
1 0	$3 \cdot 9 = 27^{C=1}$
1 1	$9 \cdot 6 = 54^{C=1}$

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

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

Variable elimination



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

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
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

A	$g_A(a)$
0	$90^{B=0}$
1	$54^{B=1}$

B	$g_B(b)$
0	$9^{C=1}$
1	$6^{C=1}$

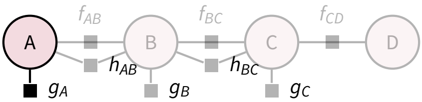
C	$g_C(c)$
0	$4^{D=0}$
1	$3^{D=1}$

A B	$h_{AB}(a,b)$
0 0	$10 \cdot 9 = 90^{C=1}$
0 1	$2 \cdot 6 = 12^{C=1}$
1 0	$3 \cdot 9 = 27^{C=1}$
1 1	$9 \cdot 6 = 54^{C=1}$

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

1. Pick order: D, C, B, A
2. Maximize over D ($f_{CD} \to g_C$)
3. Multiply f_{BC} with g_C giving h_{BC}
4. Maximize over C ($h_{BC} \to g_B$)
5. Multiply f_{AB} with g_B giving h_{AB}
6. Maximize over B ($h_{AB} \to g_A$)
7. Maximize over A ($g_A \to \emptyset$)

Variable elimination



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

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
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

A	$g_A(a)$
0	$90^{B=0}$
1	$54^{B=1}$

B	$g_B(b)$
0	$9^{C=1}$
1	$6^{C=1}$

C	$g_C(c)$
0	$4^{D=0}$
1	$3^{D=1}$

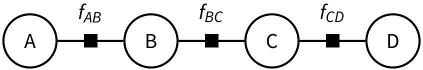
A B	$h_{AB}(a,b)$
0 0	$10 \cdot 9 = 90^{C=1}$
0 1	$2 \cdot 6 = 12^{C=1}$
1 0	$3 \cdot 9 = 27^{C=1}$
1 1	$9 \cdot 6 = 54^{C=1}$

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

1. Pick order: D, C, B, A
2. Maximize over D ($f_{CD} \to g_C$)
3. Multiply f_{BC} with g_C giving h_{BC}
4. Maximize over C ($h_{BC} \to g_B$)
5. Multiply f_{AB} with g_B giving h_{AB}
6. Maximize over B ($h_{AB} \to g_A$)
7. Maximize over A ($g_A \to \emptyset$)
8. Just like Viterbi!
   The max is $^{90}/_{Z}$.

   Backtrace to get
   $\arg\max : (0, 0, 1, 1)$.

Variable elimination: sum



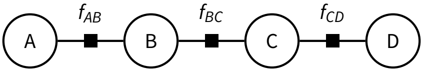
Query: $Z = \sum_{a,b,c,d} f(a,b,c,d) = ?$

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
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

Variable elimination: sum



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

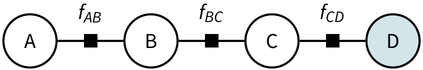
A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
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

Variable elimination: sum



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

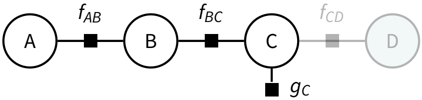
A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
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
2. **Sum** over D ($f_{CD} \to g_C$)

Variable elimination: sum



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

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

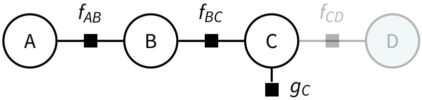
B C	$f_{BC}(b,c)$
0 0	1
0 1	3
1 0	1
1 1	2

C D	$f_{CD}(c,d)$
0 0	4
0 1	2
1 0	1
1 1	3

C	$g_C(c)$
0	6
1	4

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

Variable elimination: sum



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

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

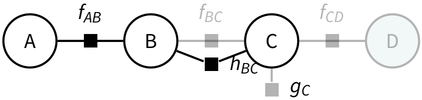
B C	$f_{BC}(b,c)$
0 0	1
0 1	3
1 0	1
1 1	2

C D	$f_{CD}(c,d)$
0 0	4
0 1	2
1 0	1
1 1	3

C	$g_C(c)$
0	6
1	4

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}

Variable elimination: sum



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

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

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

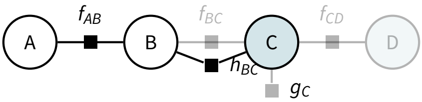
B C	$f_{BC}(b,c)$
0 0	1
0 1	3
1 0	1
1 1	2

C D	$f_{CD}(c,d)$
0 0	4
0 1	2
1 0	1
1 1	3

C	$g_C(c)$
0	6
1	4

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

Variable elimination: sum



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

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
0 0	1
0 1	3
1 0	1
1 1	2

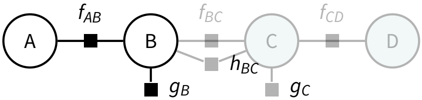
C D	$f_{CD}(c,d)$
0 0	4
0 1	2
1 0	1
1 1	3

C	$g_C(c)$
0	6
1	4

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

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}
4. **Sum** over C ($h_{BC} \rightarrow g_B$)

Variable elimination: sum



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

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

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
0 0	1
0 1	3
1 0	1
1 1	2

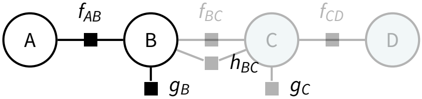
C D	$f_{CD}(c,d)$
0 0	4
0 1	2
1 0	1
1 1	3

B	$g_B(b)$
0	18
1	14

C	$g_C(c)$
0	6
1	4

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

Variable elimination: sum



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

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
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

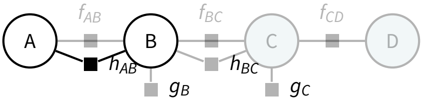
B	$g_B(b)$
0	18
1	14

C	$g_C(c)$
0	6
1	4

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

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

Variable elimination: sum



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

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
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

B	$g_B(b)$
0	18
1	14

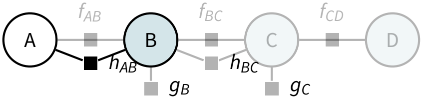
C	$g_C(c)$
0	6
1	4

A B	$h_{AB}(a,b)$
0 0	$10 \cdot 18 = 180$
0 1	$2 \cdot 14 = 28$
1 0	$3 \cdot 18 = 54$
1 1	$9 \cdot 14 = 126$

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

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

Variable elimination: sum



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

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
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

B	$g_B(b)$
0	18
1	14

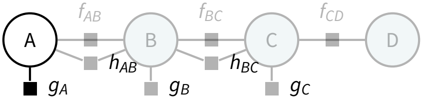
C	$g_C(c)$
0	6
1	4

A B	$h_{AB}(a,b)$
0 0	$10 \cdot 18 = 180$
0 1	$2 \cdot 14 = 28$
1 0	$3 \cdot 18 = 54$
1 1	$9 \cdot 14 = 126$

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

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

Variable elimination: sum



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

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
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

A	$g_A(a)$
0	208
1	180

B	$g_B(b)$
0	18
1	14

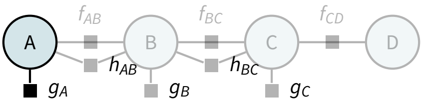
C	$g_C(c)$
0	6
1	4

A B	$h_{AB}(a,b)$
0 0	$10 \cdot 18 = 180$
0 1	$2 \cdot 14 = 28$
1 0	$3 \cdot 18 = 54$
1 1	$9 \cdot 14 = 126$

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

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

Variable elimination: sum



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

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
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

A	$g_A(a)$
0	208
1	180

B	$g_B(b)$
0	18
1	14

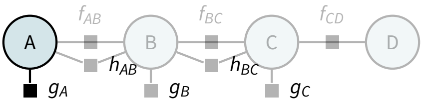
C	$g_C(c)$
0	6
1	4

A B	$h_{AB}(a,b)$
0 0	$10 \cdot 18 = 180$
0 1	$2 \cdot 14 = 28$
1 0	$3 \cdot 18 = 54$
1 1	$9 \cdot 14 = 126$

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

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

Variable elimination: sum



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

A B	$f_{AB}(a, b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b, c)$
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

A	$g_A(a)$
0	208
1	180

B	$g_B(b)$
0	18
1	14

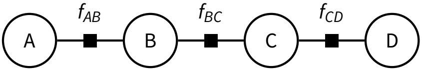
C	$g_C(c)$
0	6
1	4

A B	$h_{AB}(a, b)$
0 0	$10 \cdot 18 = 180$
0 1	$2 \cdot 14 = 28$
1 0	$3 \cdot 18 = 54$
1 1	$9 \cdot 14 = 126$

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

1. Pick order: D, C, B, A
2. **Sum** over D ($f_{CD} \to g_C$)
3. Multiply f_{BC} with g_C giving h_{BC}
4. **Sum** over C ($h_{BC} \to g_B$)
5. Multiply f_{AB} with g_B giving h_{AB}
6. **Sum** over B ($h_{AB} \to g_A$)
7. **Sum** over A ($g_A \to \emptyset$)
8. Just like the Forward algorithm!
   $Z = 388$.
   so $\mathrm{P}(0, 0, 1, 1) = {}^{90}/_{Z} \approx .23$
   Note: we obtained for free
   $\mathrm{P}(A = 0) = {}^{208}/_{388} \approx .54$.

Variable elimination: more complicated example



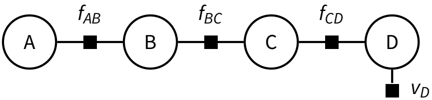
Query: $\mathrm{P}(a, c \mid D = 1) = ?$

A B	$f_{AB}(a, b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b, c)$
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

Variable elimination: more complicated example



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

A B	$f_{AB}(a, b)$
0 0	10
0 1	2
1 0	3
1 1	9

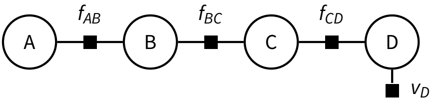
B C	$f_{BC}(b, c)$
0 0	1
0 1	3
1 0	1
1 1	2

C D	$f_{CD}(c, d)$
0 0	4
0 1	2
1 0	1
1 1	3

D	$v_D(d)$
0	0
1	1

1. Introduce evidence!

Variable elimination: more complicated example



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

A B	$f_{AB}(a, b)$
0 0	10
0 1	2
1 0	3
1 1	9

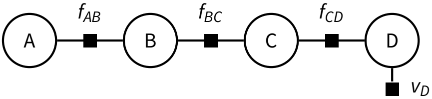
B C	$f_{BC}(b, c)$
0 0	1
0 1	3
1 0	1
1 1	2

C D	$f_{CD}(c, d)$
0 0	4
0 1	2
1 0	1
1 1	3

D	$v_D(d)$
0	0
1	1

1. Introduce evidence!
2. Pick order: D, C, B, A

Variable elimination: more complicated example



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

A B	$f_{AB}(a, b)$
0 0	10
0 1	2
1 0	3
1 1	9

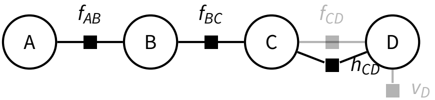
B C	$f_{BC}(b, c)$
0 0	1
0 1	3
1 0	1
1 1	2

C D	$f_{CD}(c, d)$
0 0	4
0 1	2
1 0	1
1 1	3

D	$v_D(d)$
0	0
1	1

1. Introduce evidence!
2. Pick order: D, C, B, A
3. Multiply all D factors

Variable elimination: more complicated example



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

A B	$f_{AB}(a, b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b, c)$
0 0	1
0 1	3
1 0	1
1 1	2

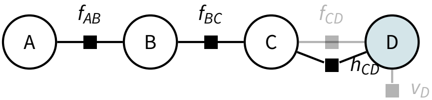
C D	$f_{CD}(c, d)$
0 0	4
0 1	2
1 0	1
1 1	3

D	$v_D(d)$
0	0
1	1

1. Introduce evidence!
2. Pick order: D, C, B, A
3. Multiply all D factors

C D	$h_{CD}(c, d)$
0 0	0
0 1	2
1 0	0
1 1	3

Variable elimination: more complicated example



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

1. Introduce evidence!
2. Pick order: D, C, B, A
3. Multiply all D factors
4. Sum over D ($h_{CD} \to g_C$)

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

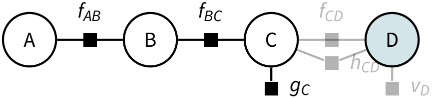
B C	$f_{BC}(b,c)$
0 0	1
0 1	3
1 0	1
1 1	2

C D	$f_{CD}(c,d)$
0 0	4
0 1	2
1 0	1
1 1	3

D	$v_D(d)$
0	0
1	1

C D	$h_{CD}(c,d)$
0 0	0
0 1	2
1 0	0
1 1	3

Variable elimination: more complicated example



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

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
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

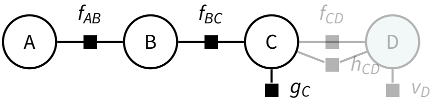
C	$g_C(c)$
0	2
1	3

D	$v_D(d)$
0	0
1	1

1. Introduce evidence!
2. Pick order: D, C, B, A
3. Multiply all D factors
4. Sum over D ($h_{CD} \to g_C$)

C D	$h_{CD}(c,d)$
0 0	0
0 1	2
1 0	0
1 1	3

Variable elimination: more complicated example



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

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
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

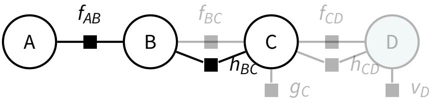
C	$g_C(c)$
0	2
1	3

D	$v_D(d)$
0	0
1	1

1. Introduce evidence!
2. Pick order: D, C, B, A
3. Multiply all D factors
4. Sum over D ($h_{CD} \to g_C$)
5. Multiply all C factors

C D	$h_{CD}(c,d)$
0 0	0
0 1	2
1 0	0
1 1	3

Variable elimination: more complicated example



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

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
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

C	$g_C(c)$
0	2
1	3

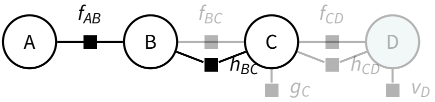
D	$v_D(d)$
0	0
1	1

B C	$h_{BC}(b,c)$
0 0	2
0 1	9
1 0	2
1 1	6

C D	$h_{CD}(c,d)$
0 0	0
0 1	2
1 0	0
1 1	3

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

Variable elimination: more complicated example



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

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
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

C	$g_C(c)$
0	2
1	3

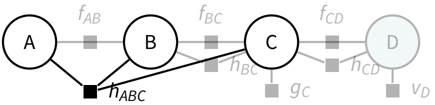
D	$v_D(d)$
0	0
1	1

B C	$h_{BC}(b,c)$
0 0	2
0 1	9
1 0	2
1 1	6

C D	$h_{CD}(c,d)$
0 0	0
0 1	2
1 0	0
1 1	3

1. Introduce evidence!
2. Pick order: D, C, B, A
3. Multiply all D factors
4. Sum over D ($h_{CD} \to g_C$)
5. Multiply all C factors
6. Multiply all B factors

Variable elimination: more complicated example



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

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
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

C	$g_C(c)$
0	2
1	3

D	$v_D(d)$
0	0
1	1

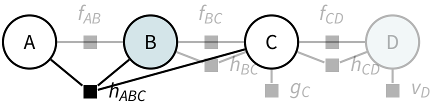
A B C	$h_{ABC}(a,b,c)$
0 0 0	20
0 0 1	90
0 1 0	4
0 1 1	12
1 0 0	6
1 0 1	18
1 1 0	18
1 1 1	54

B C	$h_{BC}(b,c)$
0 0	2
0 1	9
1 0	2
1 1	6

C D	$h_{CD}(c,d)$
0 0	0
0 1	2
1 0	0
1 1	3

1. Introduce evidence!
2. Pick order: D, C, B, A
3. Multiply all D factors
4. Sum over D ($h_{CD} \to g_C$)
5. Multiply all C factors
6. Multiply all B factors

Variable elimination: more complicated example



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

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
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

C	$g_C(c)$
0	2
1	3

D	$v_D(d)$
0	0
1	1

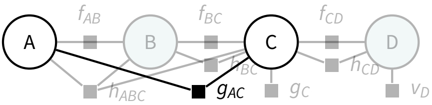
A B C	$h_{ABC}(a,b,c)$
0 0 0	20
0 0 1	90
0 1 0	4
0 1 1	12
1 0 0	6
1 0 1	18
1 1 0	18
1 1 1	54

B C	$h_{BC}(b,c)$
0 0	2
0 1	9
1 0	2
1 1	6

1. Introduce evidence!
2. Pick order: D, C, B, A
3. Multiply all *D* factors
4. Sum over *D* ($h_{CD} \to g_C$)
5. Multiply all *C* factors
6. Multiply all *B* factors
7. Sum over *B*.

C D	$h_{CD}(c,d)$
0 0	0
0 1	2
1 0	0
1 1	3

Variable elimination: more complicated example



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

A B	$f_{AB}(a,b)$
0 0	10
0 1	2
1 0	3
1 1	9

B C	$f_{BC}(b,c)$
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

A C	$g_{AC}(a,c)$
0 0	24
0 1	102
1 0	24
1 1	72

C	$g_C(c)$
0	2
1	3

D	$v_D(d)$
0	0
1	1

A B C	$h_{ABC}(a,b,c)$
0 0 0	20
0 0 1	90
0 1 0	4
0 1 1	12
1 0 0	6
1 0 1	18
1 1 0	18
1 1 1	54

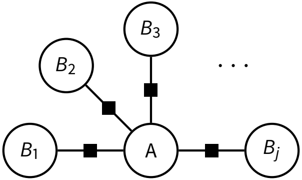
B C	$h_{BC}(b,c)$
0 0	2
0 1	9
1 0	2
1 1	6

1. Introduce evidence!
2. Pick order: D, C, B, A
3. Multiply all *D* factors
4. Sum over *D* ($h_{CD} \to g_C$)
5. Multiply all *C* factors
6. Multiply all *B* factors
7. Sum over *B*.

C D	$h_{CD}(c,d)$
0 0	0
0 1	2
1 0	0
1 1	3

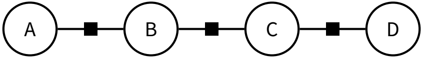
Variable elimination

- Answer any query involving max, marginalization, evidence!
- Complexity depends on **elimination order**: $\mathcal{O}(nk^M)$
  - 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!

Variable elimination as message passing

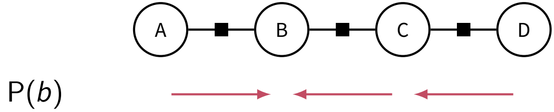


$P(b)$

- Optimal order: A, D, C (or D, C, A)

[3] because it's a tree

Variable elimination as message passing

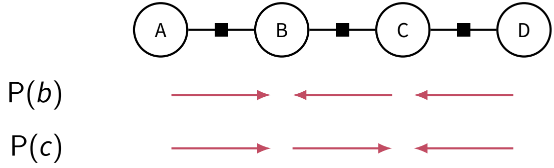


- Optimal order: A, D, C (or D, C, A)
- At each step, we eliminate a variable Y by multiplying (at most[3]) two factors and summing over Y:

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

[3] because it's a tree

Variable elimination as message passing



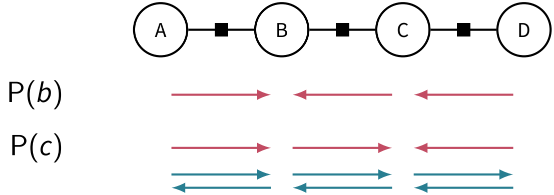
- Optimal order: A, D, C (or D, C, A)
- At each step, we eliminate a variable Y by multiplying (at most[3]) two factors and summing over Y:

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

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

[3] because it's a tree

Variable elimination as message passing



- Optimal order: A, D, C (or D, C, A)
- At each step, we eliminate a variable Y by multiplying (at most[3]) 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!

[3]because it's a tree

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 $\mathbf{1}$).

$$\nu_{X\to\alpha}(x) = \prod_{\beta\in\mathcal{N}(X)-\alpha} \mu_{\beta\to X}(x)$$

- Messages from factor α to variable X: marginalizes over all assignments $y_1, \ldots, y_k$ for $Y_1, \ldots, Y_k$ neighboring α

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

- 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:

$$p_X(x) \propto \prod_{\alpha \in \mathcal{N}(X)} \mu_{\alpha \to X}(x)$$

- Factor beliefs:

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

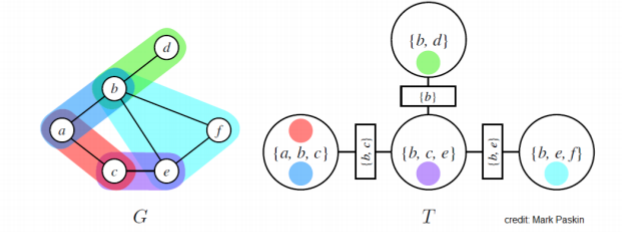
- If no cycles, once all messages are passed, beliefs are true marginals:

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

- What to do if there are cycles?

Inference in loopy graphs

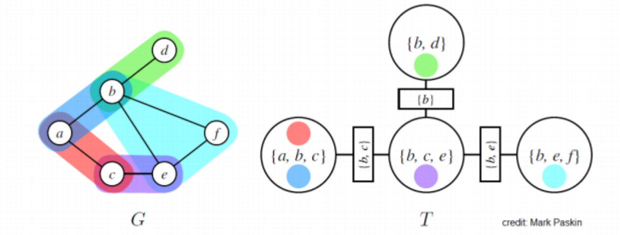
- Exact solution: **Junction Tree** algorithm:
  - convert the graph into a tree, by merging cliques!



  - 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

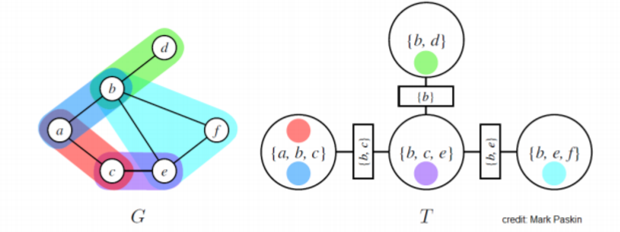
- Exact solution: **Junction Tree** algorithm:
  - convert the graph into a tree, by merging cliques!



  - 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.)

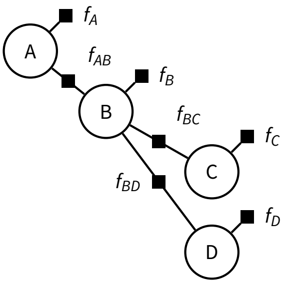
Inference in loopy graphs

- 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).

Example: classifying opinion in a forum



$a, b, c, d \in \{+, -\}$

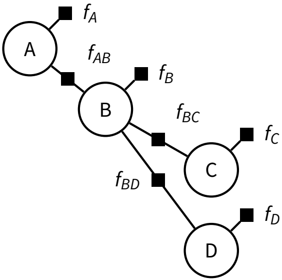
y	$f_A(y)$	$f_B(y)$	$f_C(y)$	$f_D(y)$
-	10	1	1	10
+	1	1	1	1

y	z	f(y, z)
-	-	5
-	+	1
+	-	1
+	+	2

> A: I didn't like the movie.
>
> B: Hmm, strange, why not?
>
> C: It was slow.
>
> D: It was the worst movie this year.

- Unary factors: *soft evidence*. B, C locally ambiguous.
- Pairwise factors, all equal: $f_{AB} = f_{BC} = f_{BD} = f$.

Example: classifying opinion in a forum

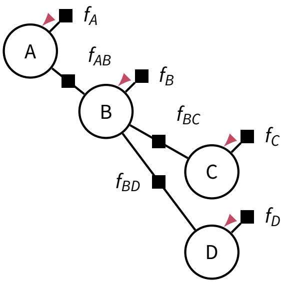


$a, b, c, d \in \{+, -\}$

y	$f_A(y)$	$f_B(y)$	$f_C(y)$	$f_D(y)$
–	10	1	1	10
+	1	1	1	1

y	z	f(y, z)
–	–	5
–	+	1
+	–	1
+	+	2

Example: classifying opinion in a forum



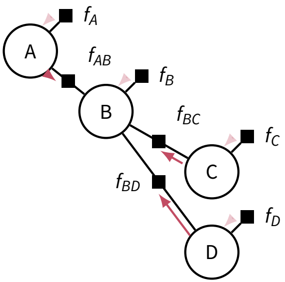
$a, b, c, d \in \{+, -\}$

y	$f_A(y)$	$f_B(y)$	$f_C(y)$	$f_D(y)$
–	10	1	1	10
+	1	1	1	1

y	z	f(y, z)
–	–	5
–	+	1
+	–	1
+	+	2

1. Unary to var: $\mu_{f_Y \to Y} = f_Y$. example: $\mu_{f_D \to D} = \begin{cases} 10 \\ 1 \end{cases}$

Example: classifying opinion in a forum



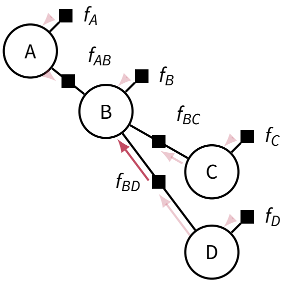
$a, b, c, d \in \{+, -\}$

y	$f_A(y)$	$f_B(y)$	$f_C(y)$	$f_D(y)$
-	10	1	1	10
+	1	1	1	1

y	z	f(y, z)
-	-	5
-	+	1
+	-	1
+	+	2

1. Unary to var: $\mu_{f_Y \to Y} = f_Y$. example: $\mu_{f_D \to D} = \begin{cases} 10 \\ 1 \end{cases}$
2. Pass from leaves to their neighboring pw. factors: $\nu_{D \to f_{BD}} = \mu_{f_D \to D} = f_D$. Similarly, $\nu_{C \to f_{BC}} = f_C$, $\nu_{A \to f_{AB}} = f_A$

Example: classifying opinion in a forum



$a, b, c, d \in \{+, -\}$

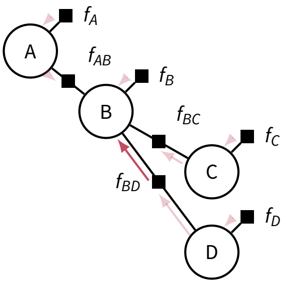
y	$f_A(y)$	$f_B(y)$	$f_C(y)$	$f_D(y)$
–	10	1	1	10
+	1	1	1	1

y	z	f(y, z)
–	–	5
–	+	1
+	–	1
+	+	2

1. Unary to var: $\mu_{f_Y \to Y} = f_Y$. example: $\mu_{f_D \to D} = \begin{cases} 10 \\ 1 \end{cases}$
2. Pass from leaves to their neighboring pw. factors: $\nu_{D \to f_{BD}} = \mu_{f_D \to D} = f_D$. Similarly, $\nu_{C \to f_{BC}} = f_C$, $\nu_{A \to f_{AB}} = f_A$
3. Pass factor messages to B (sum-product!):

$$\mu_{f_{BD} \to B}(b) = \sum_d f(b, d) \nu_{D \to f_{BD}}(d) =$$

Example: classifying opinion in a forum



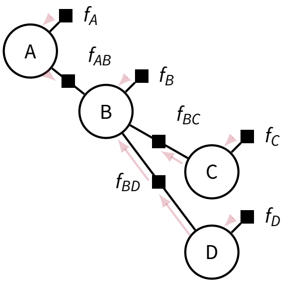
$a, b, c, d \in \{+, -\}$

y	$f_A(y)$	$f_B(y)$	$f_C(y)$	$f_D(y)$
-	10	1		10
+	1	1		1

y	z	f(y, z)
-	-	5
-	+	1
+	-	1
+	+	2

1. Unary to var: $\mu_{f_Y \to Y} = f_Y$. example: $\mu_{f_D \to D} = \begin{cases} 10 \\ 1 \end{cases}$
2. Pass from leaves to their neighboring pw. factors: $\nu_{D \to f_{BD}} = \mu_{f_D \to D} = f_D$. Similarly, $\nu_{C \to f_{BC}} = f_C$, $\nu_{A \to f_{AB}} = f_A$
3. Pass factor messages to B (sum-product!): $$\mu_{f_{BD} \to B}(b) = \sum_d f(b, d) \nu_{D \to f_{BD}}(d) = \begin{cases} 50 + 1 = 51 \\ \end{cases}$$

Example: classifying opinion in a forum



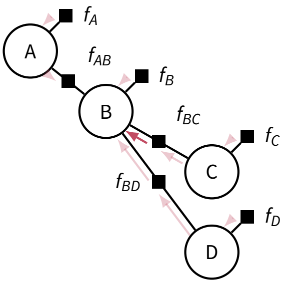
$a, b, c, d \in \{+, -\}$

y	$f_A(y)$	$f_B(y)$	$f_C(y)$	$f_D(y)$
-	10	1		10
+	1	1		1

y	z	f(y, z)
-	-	5
-	+	1
+	-	1
+	+	2

1. Unary to var: $\mu_{f_Y \to Y} = f_Y$. example: $\mu_{f_D \to D} = \begin{cases} 10 \\ 1 \end{cases}$
2. Pass from leaves to their neighboring pw. factors: $\nu_{D \to f_{BD}} = \mu_{f_D \to D} = f_D$. Similarly, $\nu_{C \to f_{BC}} = f_C, \nu_{A \to f_{AB}} = f_A$
3. Pass factor messages to B (sum-product!):
$$\mu_{f_{BD} \to B}(b) = \sum_d f(b, d)\nu_{D \to f_{BD}}(d) = \begin{cases} 50 + 1 = 51 \\ 10 + 2 = 12 \end{cases}$$

Example: classifying opinion in a forum



$a, b, c, d \in \{+, -\}$

y	$f_A(y)$	$f_B(y)$	$f_C(y)$	$f_D(y)$
-	10		1	10
+	1		1	1

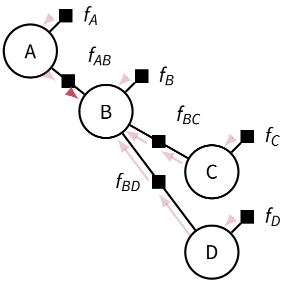
y	z	f(y, z)
-	-	5
-	+	1
+	-	1
+	+	2

1. Unary to var: $\mu_{f_Y \to Y} = f_Y$. example: $\mu_{f_D \to D} = \begin{cases} 10 \\ 1 \end{cases}$
2. Pass from leaves to their neighboring pw. factors: $\nu_{D \to f_{BD}} = \mu_{f_D \to D} = f_D$. Similarly, $\nu_{C \to f_{BC}} = f_C, \nu_{A \to f_{AB}} = f_A$
3. Pass factor messages to B (sum-product!):

$$\mu_{f_{BD} \to B}(b) = \sum_d f(b, d) \nu_{D \to f_{BD}}(d) = \begin{cases} 50 + 1 = 51 \\ 10 + 2 = 12 \end{cases}$$

$$\mu_{f_{BC} \to B}(b) = \sum_c f(b, c) \nu_{C \to f_{BC}}(c) = \begin{cases} 6 \\ 3 \end{cases}$$

Example: classifying opinion in a forum



$a, b, c, d \in \{+, -\}$

y	$f_A(y)$	$f_B(y)$	$f_C(y)$	$f_D(y)$
-	10	1	1	10
+	1	1	1	1

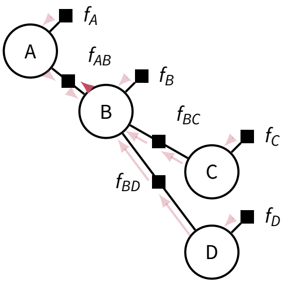
y	z	f(y, z)
-	-	5
-	+	1
+	-	1
+	+	2

1. Unary to var: $\mu_{f_Y \to Y} = f_Y$. example: $\mu_{f_D \to D} = \begin{cases} 10 \\ 1 \end{cases}$
2. Pass from leaves to their neighboring pw. factors: $\nu_{D \to f_{BD}} = \mu_{f_D \to D} = f_D$. Similarly, $\nu_{C \to f_{BC}} = f_C$, $\nu_{A \to f_{AB}} = f_A$
3. Pass factor messages to B (sum-product!):

$$\mu_{f_{BD} \to B}(b) = \sum_d f(b, d) \nu_{D \to f_{BD}}(d) = \begin{cases} 50 + 1 = 51 \\ 10 + 2 = 12 \end{cases}$$

$$\mu_{f_{BC} \to B}(b) = \sum_c f(b, c) \nu_{C \to f_{BC}}(c) = \begin{cases} 6 \\ 3 \end{cases} \quad \mu_{f_{AB} \to B} = \begin{cases} 51 \\ 12 \end{cases}$$

Example: classifying opinion in a forum



$a, b, c, d \in \{+, -\}$

y	$f_A(y)$	$f_B(y)$	$f_C(y)$	$f_D(y)$
-	10	1	1	10
+	1	1	1	1

y	z	f(y, z)
-	-	5
-	+	1
+	-	1
+	+	2

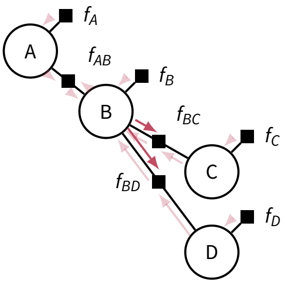
1. Unary to var: $\mu_{f_Y \to Y} = f_Y$. example: $\mu_{f_D \to D} = \begin{cases} 10 \\ 1 \end{cases}$
2. Pass from leaves to their neighboring pw. factors: $\nu_{D \to f_{BD}} = \mu_{f_D \to D} = f_D$. Similarly, $\nu_{C \to f_{BC}} = f_C$, $\nu_{A \to f_{AB}} = f_A$
3. Pass factor messages to B (sum-product!):

$$\mu_{f_{BD} \to B}(b) = \sum_d f(b, d) \nu_{D \to f_{BD}}(d) = \begin{cases} 50 + 1 = 51 \\ 10 + 2 = 12 \end{cases}$$

$$\mu_{f_{BC} \to B}(b) = \sum_c f(b, c) \nu_{C \to f_{BC}}(c) = \begin{cases} 6 \\ 3 \end{cases} \qquad \mu_{f_{AB} \to B} = \begin{cases} 51 \\ 12 \end{cases}$$

4. And back: $\nu_{B \to f_{AB}} = \mu_{f_{BC} \to B} \cdot \mu_{f_{BD} \to B} = \begin{cases} 51 \cdot 6 = 306 \\ 12 \cdot 3 = 36 \end{cases}$

Example: classifying opinion in a forum



$a, b, c, d \in \{+, -\}$

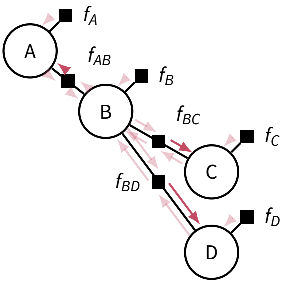
y	$f_A(y)$	$f_B(y)$	$f_C(y)$	$f_D(y)$
-	10	1	1	10
+	1	1	1	1

y	z	f(y, z)
-	-	5
-	+	1
+	-	1
+	+	2

1. Unary to var: $\mu_{f_Y \to Y} = f_Y$. example: $\mu_{f_D \to D} = \begin{cases} 10 \\ 1 \end{cases}$
2. Pass from leaves to their neighboring pw. factors: $\nu_{D \to f_{BD}} = \mu_{f_D \to D} = f_D$. Similarly, $\nu_{C \to f_{BC}} = f_C$, $\nu_{A \to f_{AB}} = f_A$
3. Pass factor messages to B (sum-product!):
$$\mu_{f_{BD} \to B}(b) = \sum_d f(b, d) \nu_{D \to f_{BD}}(d) = \begin{cases} 50 + 1 = 51 \\ 10 + 2 = 12 \end{cases}$$
$$\mu_{f_{BC} \to B}(b) = \sum_c f(b, c) \nu_{C \to f_{BC}}(c) = \begin{cases} 6 \\ 3 \end{cases} \quad \mu_{f_{AB} \to B} = \begin{cases} 51 \\ 12 \end{cases}$$
4. And back: $\nu_{B \to f_{AB}} = \mu_{f_{BC} \to B} \cdot \mu_{f_{BD} \to B} = \begin{cases} 51 \cdot 6 = 306 \\ 12 \cdot 3 = 36 \end{cases}$

   Similarly, $\nu_{B \to f_{BC}} = \begin{cases} 51 \cdot 51 \\ 12 \cdot 12 \end{cases}$ and $\nu_{B \to f_{BD}} = \begin{cases} 51 \cdot 6 \\ 12 \cdot 3 \end{cases}$

Example: classifying opinion in a forum



$a, b, c, d \in \{+, -\}$

y	$f_A(y)$	$f_B(y)$	$f_C(y)$	$f_D(y)$
–	10	1	1	10
+	1	1	1	1

y	z	f(y, z)
–	–	5
–	+	1
+	–	1
+	+	2

1. Unary to var: $\mu_{f_Y \to Y} = f_Y$. example: $\mu_{f_D \to D} = \begin{cases} 10 \\ 1 \end{cases}$
2. Pass from leaves to their neighboring pw. factors: $\nu_{D \to f_{BD}} = \mu_{f_D \to D} = f_D$. Similarly, $\nu_{C \to f_{BC}} = f_C$, $\nu_{A \to f_{AB}} = f_A$
3. Pass factor messages to B (sum-product!):

$$\mu_{f_{BD} \to B}(b) = \sum_d f(b, d) \nu_{D \to f_{BD}}(d) = \begin{cases} 50 + 1 = 51 \\ 10 + 2 = 12 \end{cases}$$

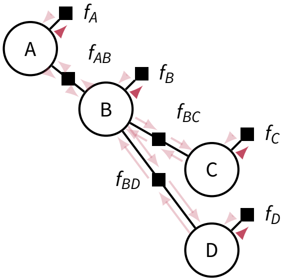
$$\mu_{f_{BC} \to B}(b) = \sum_c f(b, c) \nu_{C \to f_{BC}}(c) = \begin{cases} 6 \\ 3 \end{cases} \quad \mu_{f_{AB} \to B} = \begin{cases} 51 \\ 12 \end{cases}$$

4. And back: $\nu_{B \to f_{AB}} = \mu_{f_{BC} \to B} \cdot \mu_{f_{BD} \to B} = \begin{cases} 51 \cdot 6 = 306 \\ 12 \cdot 3 = 36 \end{cases}$

Similarly, $\nu_{B \to f_{BC}} = \begin{cases} 51 \cdot 51 \\ 12 \cdot 12 \end{cases}$ and $\nu_{B \to f_{BD}} = \begin{cases} 51 \cdot 6 \\ 12 \cdot 3 \end{cases}$

5. Finally $\mu_{f_{AB} \to A}(a) = \sum_b f(a, b) \nu_{B \to f_{AB}}(b) = \begin{cases} 1566 \\ 378 \end{cases}$ etc.

Example: classifying opinion in a forum



$a, b, c, d \in \{+, -\}$

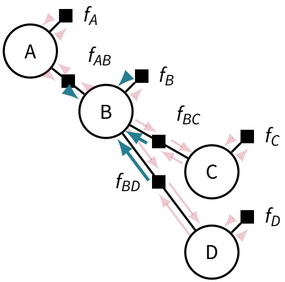
y	$f_A(y)$	$f_B(y)$	$f_C(y)$	$f_D(y)$
-	10	1	1	10
+	1	1	1	1

y	z	f(y, z)
-	-	5
-	+	1
+	-	1
+	+	2

1. Unary to var: $\mu_{f_Y \to Y} = f_Y$. example: $\mu_{f_D \to D} = \begin{cases} 10 \\ 1 \end{cases}$
2. Pass from leaves to their neighboring pw. factors: $\nu_{D \to f_{BD}} = \mu_{f_D \to D} = f_D$. Similarly, $\nu_{C \to f_{BC}} = f_C$, $\nu_{A \to f_{AB}} = f_A$
3. Pass factor messages to B (sum-product!):
$$\mu_{f_{BD} \to B}(b) = \sum_d f(b,d)\nu_{D \to f_{BD}}(d) = \begin{cases} 50 + 1 = 51 \\ 10 + 2 = 12 \end{cases}$$
$$\mu_{f_{BC} \to B}(b) = \sum_c f(b,c)\nu_{C \to f_{BC}}(c) = \begin{cases} 6 \\ 3 \end{cases} \quad \mu_{f_{AB} \to B} = \begin{cases} 51 \\ 12 \end{cases}$$
4. And back: $\nu_{B \to f_{AB}} = \mu_{f_{BC} \to B} \cdot \mu_{f_{BD} \to B} = \begin{cases} 51 \cdot 6 = 306 \\ 12 \cdot 3 = 36 \end{cases}$

   Similarly, $\nu_{B \to f_{BC}} = \begin{cases} 51 \cdot 51 \\ 12 \cdot 12 \end{cases}$ and $\nu_{B \to f_{BD}} = \begin{cases} 51 \cdot 6 \\ 12 \cdot 3 \end{cases}$
5. Finally $\mu_{f_{AB} \to A}(a) = \sum_b f(a,b)\nu_{B \to f_{AB}}(b) = \begin{cases} 1566 \\ 378 \end{cases}$ etc.

Example: classifying opinion in a forum



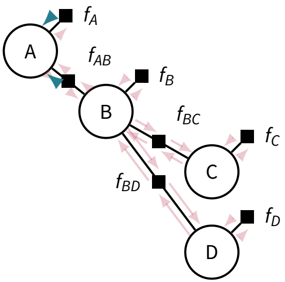
$a, b, c, d \in \{+, -\}$

y	$f_A(y)$	$f_B(y)$	$f_C(y)$	$f_D(y)$
-	10	1	1	10
+	1	1	1	1

y	z	f(y, z)
-	-	5
-	+	1
+	-	1
+	+	2

1. Unary to var: $\mu_{f_Y \to Y} = f_Y$. example: $\mu_{f_D \to D} = \begin{cases} 10 \\ 1 \end{cases}$
2. Pass from leaves to their neighboring pw. factors: $\nu_{D \to f_{BD}} = \mu_{f_D \to D} = f_D$. Similarly, $\nu_{C \to f_{BC}} = f_C$, $\nu_{A \to f_{AB}} = f_A$
3. Pass factor messages to B (sum-product!):
$$\mu_{f_{BD} \to B}(b) = \sum_d f(b, d) \nu_{D \to f_{BD}}(d) = \begin{cases} 50 + 1 = 51 \\ 10 + 2 = 12 \end{cases}$$
$$\mu_{f_{BC} \to B}(b) = \sum_c f(b, c) \nu_{C \to f_{BC}}(c) = \begin{cases} 6 \\ 3 \end{cases} \quad \mu_{f_{AB} \to B} = \begin{cases} 51 \\ 12 \end{cases}$$
4. And back: $\nu_{B \to f_{AB}} = \mu_{f_{BC} \to B} \cdot \mu_{f_{BD} \to B} = \begin{cases} 51 \cdot 6 = 306 \\ 12 \cdot 3 = 36 \end{cases}$
   Similarly, $\nu_{B \to f_{BC}} = \begin{cases} 51 \cdot 51 \\ 12 \cdot 12 \end{cases}$ and $\nu_{B \to f_{BD}} = \begin{cases} 51 \cdot 6 \\ 12 \cdot 3 \end{cases}$
5. Finally $\mu_{f_{AB} \to A}(a) = \sum_b f(a, b) \nu_{B \to f_{AB}}(b) = \begin{cases} 1566 \\ 378 \end{cases}$ etc.
6. $p_B \propto \prod_\alpha \mu_{\alpha \to B} \propto \begin{cases} 51 \cdot 51 \cdot 6 \cdot 1 \\ 12 \cdot 12 \cdot 3 \cdot 1 \end{cases} = \begin{cases} .97 \\ .03 \end{cases}$

Example: classifying opinion in a forum



$a, b, c, d \in \{+, -\}$

y	$f_A(y)$	$f_B(y)$	$f_C(y)$	$f_D(y)$
-	10	1	1	10
+	1	1	1	1

y	z	f(y, z)
-	-	5
-	+	1
+	-	1
+	+	2

1. Unary to var: $\mu_{f_Y \to Y} = f_Y$. example: $\mu_{f_D \to D} = \begin{cases} 10 \\ 1 \end{cases}$
2. Pass from leaves to their neighboring pw. factors: $\nu_{D \to f_{BD}} = \mu_{f_D \to D} = f_D$. Similarly, $\nu_{C \to f_{BC}} = f_C, \nu_{A \to f_{AB}} = f_A$
3. Pass factor messages to B (sum-product!):
   $$\mu_{f_{BD} \to B}(b) = \sum_d f(b, d) \nu_{D \to f_{BD}}(d) = \begin{cases} 50 + 1 = 51 \\ 10 + 2 = 12 \end{cases}$$
   $$\mu_{f_{BC} \to B}(b) = \sum_c f(b, c) \nu_{C \to f_{BC}}(c) = \begin{cases} 6 \\ 3 \end{cases} \quad \mu_{f_{AB} \to B} = \begin{cases} 51 \\ 12 \end{cases}$$
4. And back: $\nu_{B \to f_{AB}} = \mu_{f_{BC} \to B} \cdot \mu_{f_{BD} \to B} = \begin{cases} 51 \cdot 6 = 306 \\ 12 \cdot 3 = 36 \end{cases}$

   Similarly, $\nu_{B \to f_{BC}} = \begin{cases} 51 \cdot 51 \\ 12 \cdot 12 \end{cases}$ and $\nu_{B \to f_{BD}} = \begin{cases} 51 \cdot 6 \\ 12 \cdot 3 \end{cases}$
5. Finally $\mu_{f_{AB} \to A}(a) = \sum_b f(a, b) \nu_{B \to f_{AB}}(b) = \begin{cases} 1566 \\ 378 \end{cases}$ etc.
6. $p_B \propto \prod_\alpha \mu_{\alpha \to B} \propto \begin{cases} 51 \cdot 51 \cdot 6 \cdot 1 \\ 12 \cdot 12 \cdot 3 \cdot 1 \end{cases} = \begin{cases} .97 \\ .03 \end{cases}$ $p_A \propto \begin{cases} 15660 \\ 378 \end{cases} = \begin{cases} .98 \\ .02 \end{cases}$

CRFs for any factor graph

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

CRFs for any factor graph

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**:

$$f_A(y) = \exp s_{A,y} \;=\; \text{(for example)}\;\; \exp \boldsymbol{w} \cdot \phi((\boldsymbol{x}, A), y)$$

and **pairwise scores**:

$$f_{AB}(y, y') = \exp s_{AB,y,y'} \;=\; \text{(for example)}\;\; \exp \boldsymbol{w} \cdot \phi((\boldsymbol{x}, A, B), (y, y'))$$

CRFs for any factor graph

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**:

$$f_A(y) = \exp s_{A,y} \quad = \quad \text{(for example)} \quad \exp \boldsymbol{w} \cdot \phi((\boldsymbol{x}, A), y)$$

and **pairwise scores**:

$$f_{AB}(y, y') = \exp s_{AB,y,y'} \quad = \quad \text{(for example)} \quad \exp \boldsymbol{w} \cdot \phi((\boldsymbol{x}, A, B), (y, y'))$$

(In general, **factor scores** $f_\alpha(\boldsymbol{y}_\alpha) = \exp s_{\alpha,\boldsymbol{y}_\alpha}$)

CRFs for any factor graph

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**:

$$f_A(y) = \exp s_{A,y} \;=\; \text{(for example)} \;\; \exp \boldsymbol{w} \cdot \phi((\boldsymbol{x}, A), y)$$

and **pairwise scores**:

$$f_{AB}(y, y') = \exp s_{AB,y,y'} \;=\; \text{(for example)} \;\; \exp \boldsymbol{w} \cdot \phi((\boldsymbol{x}, A, B), (y, y'))$$

(In general, **factor scores** $f_\alpha(\boldsymbol{y}_\alpha) = \exp s_{\alpha,\boldsymbol{y}_\alpha}$)

The probability of an entire labeling $\boldsymbol{y}$ is then

$$\mathrm{P}(\boldsymbol{y} \mid \boldsymbol{x}) = \frac{\prod_\alpha f_\alpha(\boldsymbol{y}_\alpha)}{Z} \quad \text{meaning} \quad \log \mathrm{P}(\boldsymbol{y} \mid \boldsymbol{x}) = \sum_\alpha s_{\alpha,\boldsymbol{y}_\alpha} - \log Z$$

CRFs for any factor graph

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**:

$$f_A(y) = \exp s_{A,y} \;=\; \text{(for example)} \;\; \exp \boldsymbol{w} \cdot \phi((\boldsymbol{x}, A), y)$$

and **pairwise scores**:

$$f_{AB}(y, y') = \exp s_{AB,y,y'} \;=\; \text{(for example)} \;\; \exp \boldsymbol{w} \cdot \phi((\boldsymbol{x}, A, B), (y, y'))$$

(In general, **factor scores** $f_\alpha(\boldsymbol{y}_\alpha) = \exp s_{\alpha,\boldsymbol{y}_\alpha}$)

The probability of an entire labeling $\boldsymbol{y}$ is then

$$\mathrm{P}(\boldsymbol{y} \mid \boldsymbol{x}) = \frac{\prod_\alpha f_\alpha(\boldsymbol{y}_\alpha)}{Z} \quad \text{meaning} \quad \log \mathrm{P}(\boldsymbol{y} \mid \boldsymbol{x}) = \sum_\alpha s_{\alpha,\boldsymbol{y}_\alpha} - \log Z$$

Gradient updates wrt a factor's scores:

$$\frac{\partial \log \mathrm{P}(\boldsymbol{y} \mid \boldsymbol{x})}{\partial s_{\alpha,\boldsymbol{y}_\alpha}} = [[\boldsymbol{y}_\alpha = \boldsymbol{y}_\alpha^{\text{true}}]] - \mathrm{P}(\boldsymbol{y}_\alpha \mid \boldsymbol{x})$$

CRFs for any factor graph

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**:

$$f_A(y) = \exp s_{A,y} \quad = \quad \text{(for example)} \ \ \exp \boldsymbol{w} \cdot \phi((\boldsymbol{x}, A), y)$$

and **pairwise scores**:

$$f_{AB}(y, y') = \exp s_{AB,y,y'} \quad = \quad \text{(for example)} \ \ \exp \boldsymbol{w} \cdot \phi((\boldsymbol{x}, A, B), (y, y'))$$

(In general, **factor scores** $f_\alpha(\boldsymbol{y}_\alpha) = \exp s_{\alpha,\boldsymbol{y}_\alpha}$)

The probability of an entire labeling $\boldsymbol{y}$ is then

$$\mathrm{P}(\boldsymbol{y} \mid \boldsymbol{x}) = \frac{\prod_\alpha f_\alpha(\boldsymbol{y}_\alpha)}{Z} \quad \text{meaning} \quad \log \mathrm{P}(\boldsymbol{y} \mid \boldsymbol{x}) = \sum_\alpha s_{\alpha,\boldsymbol{y}_\alpha} - \log Z$$

Gradient updates wrt a factor's scores:

$$\frac{\partial \log \mathrm{P}(\boldsymbol{y} \mid \boldsymbol{x})}{\partial s_{\alpha,\boldsymbol{y}_\alpha}} = [[\boldsymbol{y}_\alpha = \boldsymbol{y}_\alpha^{\text{true}}]] - \mathrm{P}(\boldsymbol{y}_\alpha \mid \boldsymbol{x})$$

The updates the factor beliefs $\mathrm{P}(\boldsymbol{y}_\alpha \mid \boldsymbol{x}) = p_\alpha(\boldsymbol{y}_\alpha)$ for each factor!

Undirected models: summary

- 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*.

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*.