# Probabilistic (Graphical) Models

## and inference

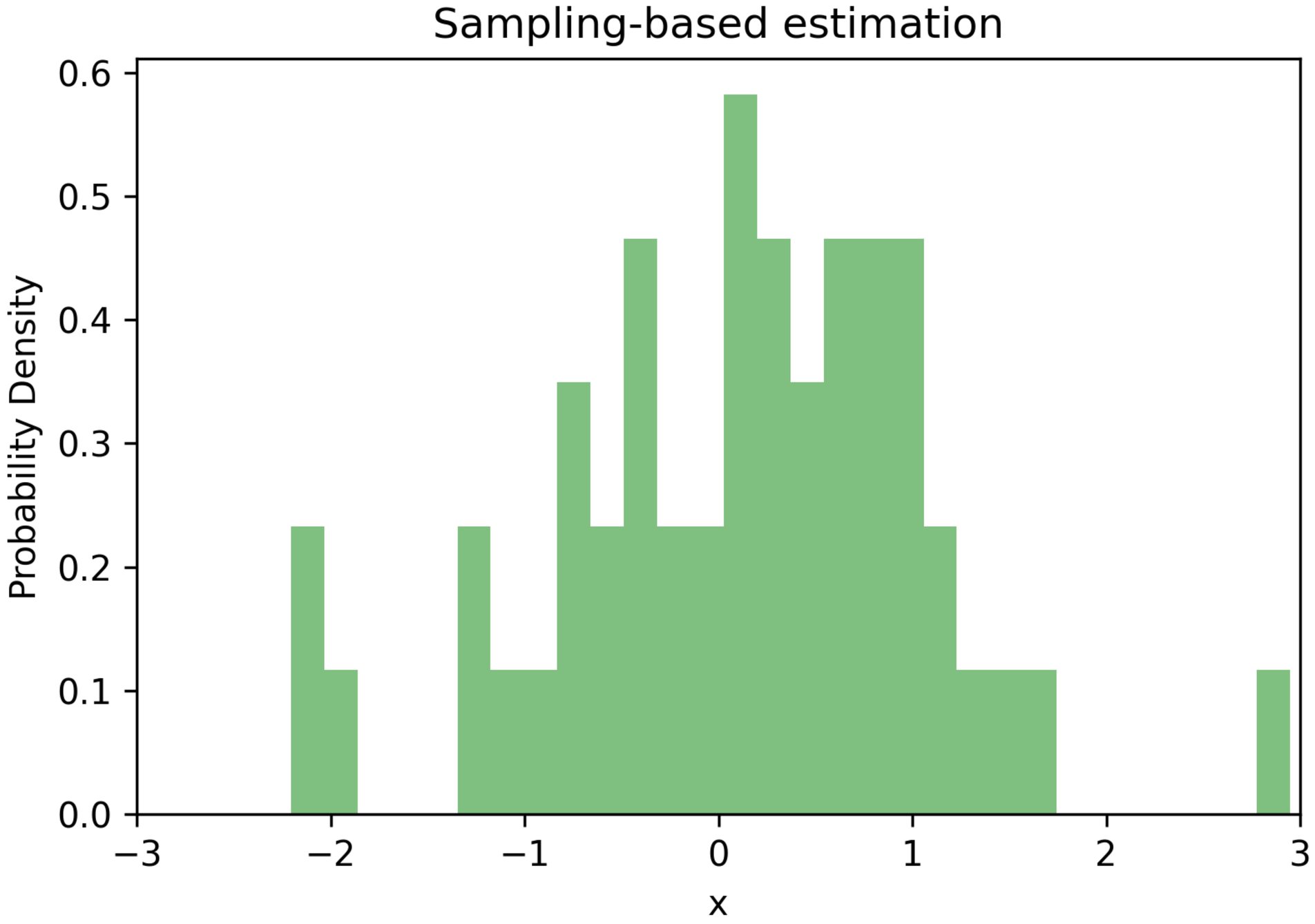
Oliver Obst · Autumn 2024

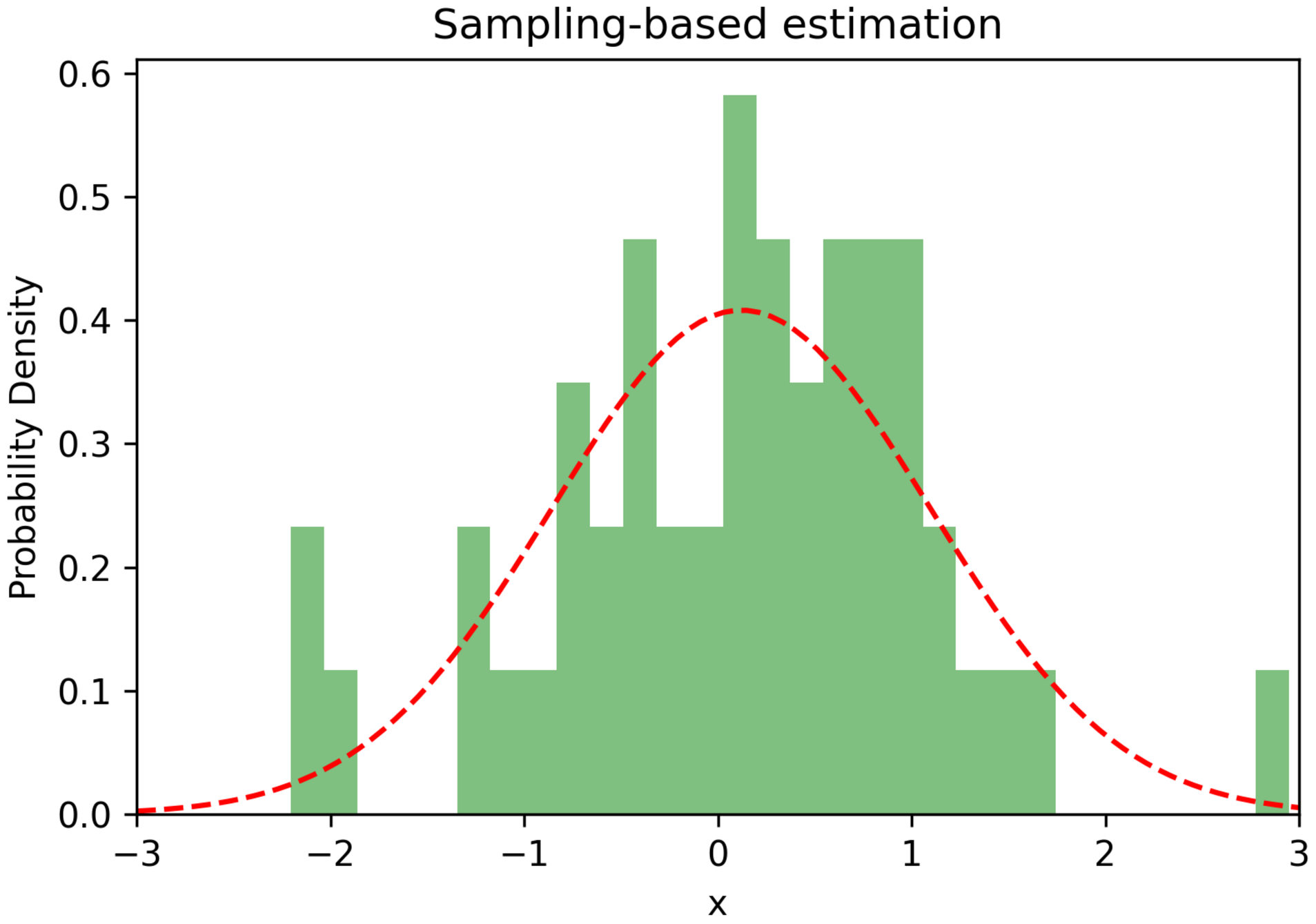
WESTERN SYDNEY
UNIVERSITY

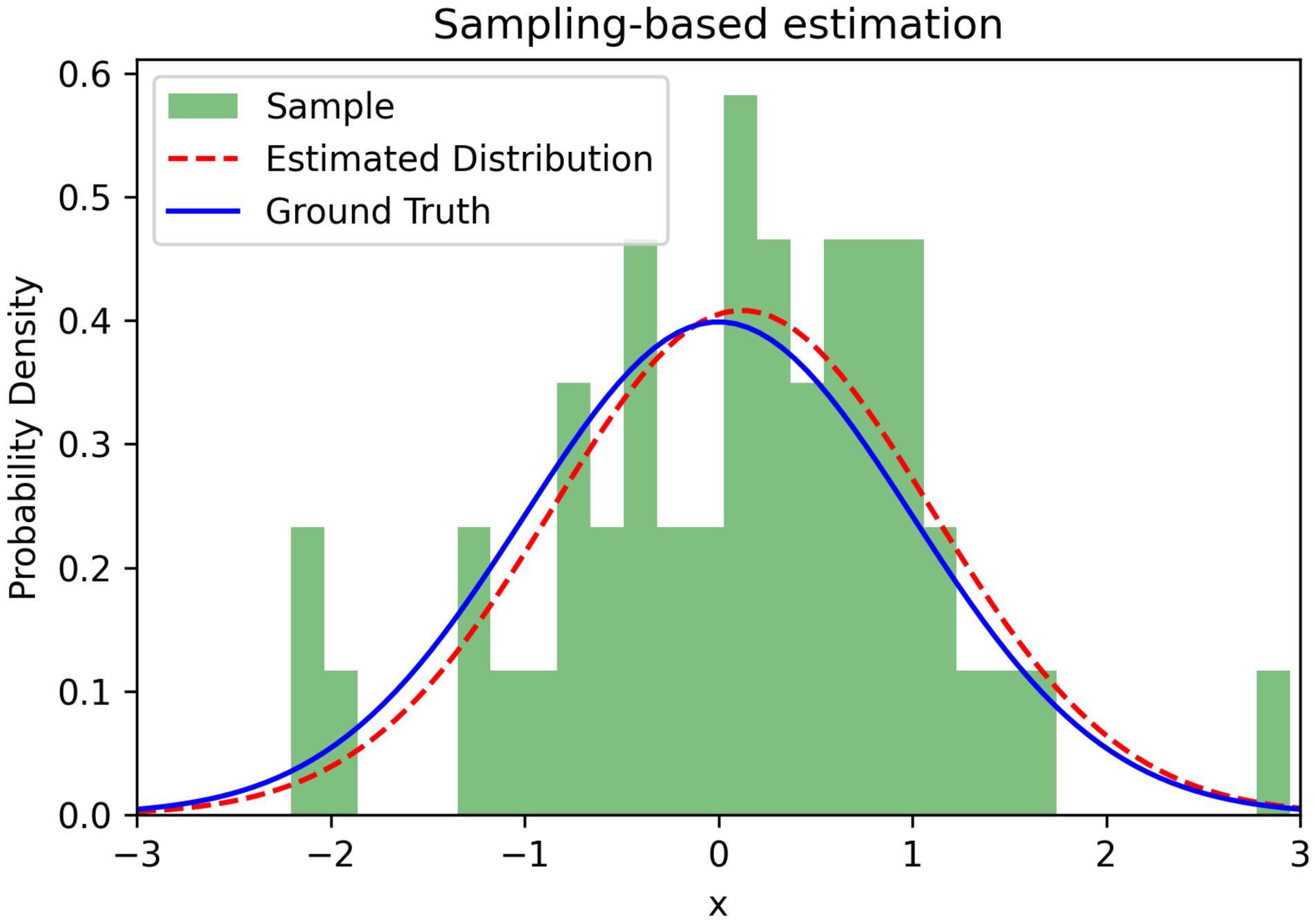
# Probabilistic (Graphical) Models and Inference

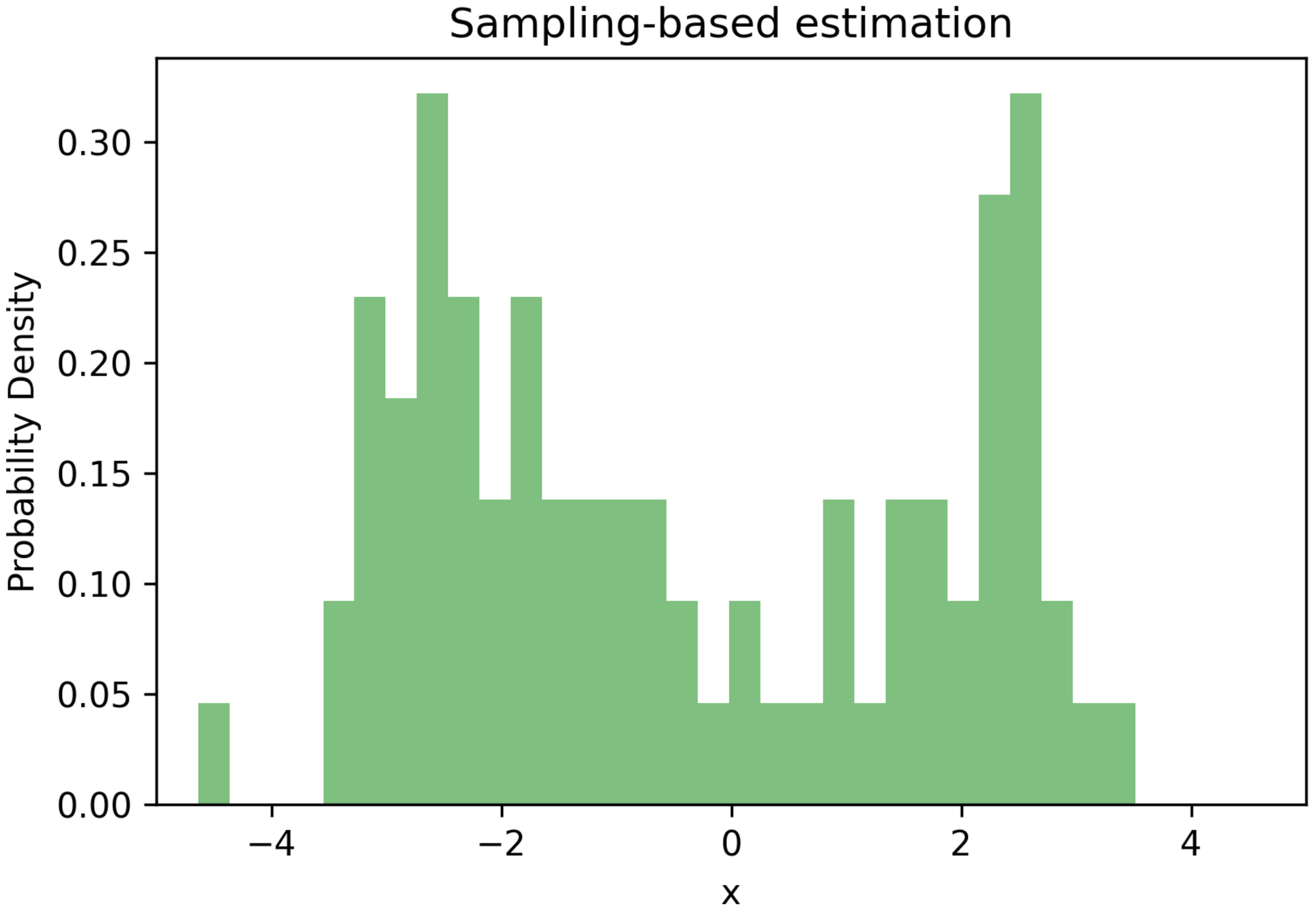
(PGM: Probabilistic Graphical Models: Principles and Techniques by Daphne Koller and Nir Friedman. MIT Press)
(PMLI: Probabilistic Machine Learning: An introduction by Kevin Murphy. MIT Press)

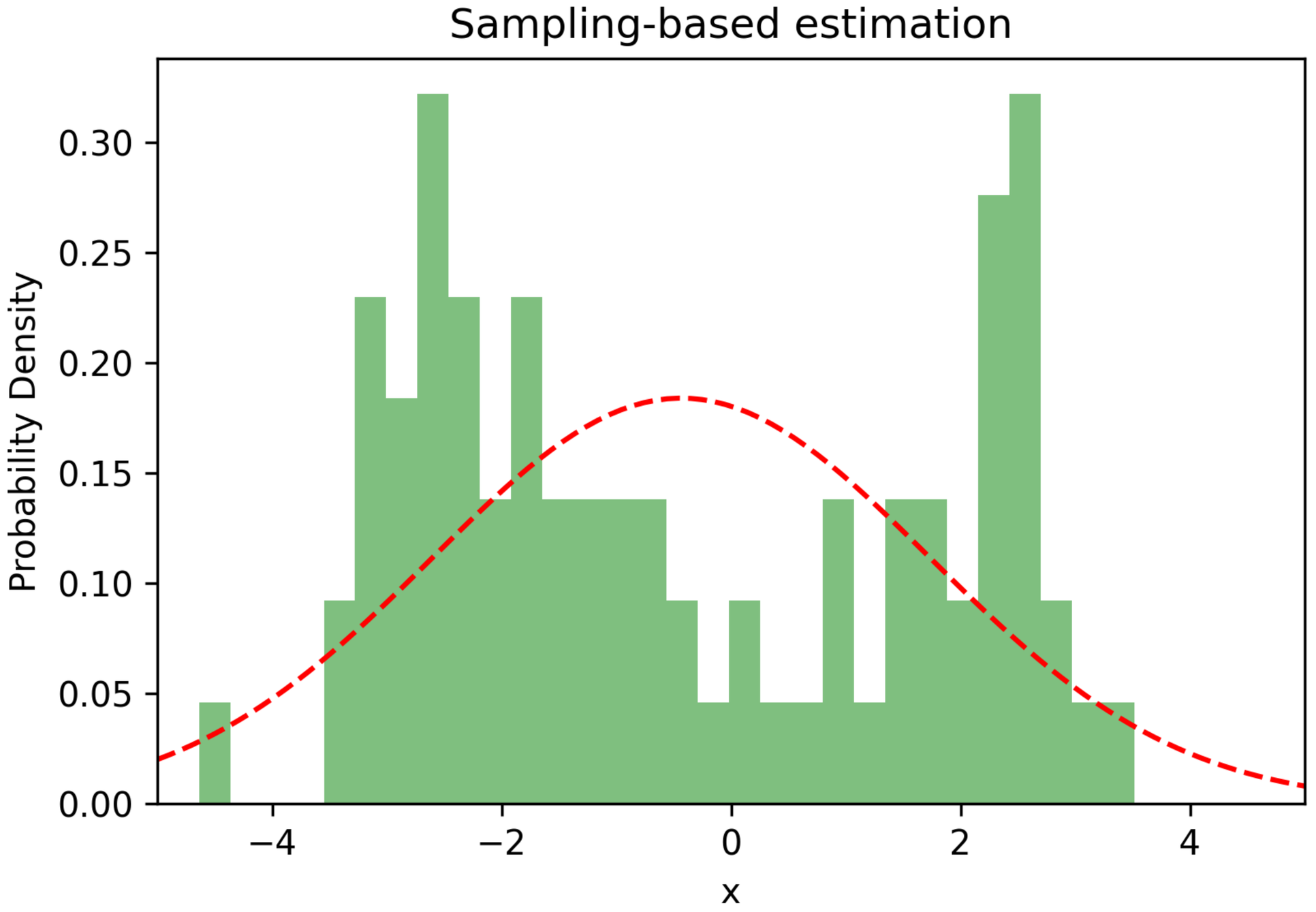
| Week | Lecture | Required reading | Assessment |
|---|---|---|---|
| 1<br>Monday, 4 March 2024 | Introduction, Probability Theory | PGM Chapter 2, PMLI Chapter 6.1 | |
| 2<br>Monday, 11 March 2024 | Directed and undirected networks introduction | PGM Chapter 3 & 4 | Quiz 1 |
| 3<br>Monday, 18 March 2024 | Variable elimination | PGM Chapter 9 | |
| 4<br>Monday, 25 March 2024 | Belief propagation | PGM Chapter 10/11 | Quiz 2 |
| 5<br>Monday, 1 April 2024 | public holiday | | 5 April 2024: census date |
| 6<br>Monday, 8 April 2024 | Message passing / Graph neural networks | https://distill.pub/2021/gnn-intro/ | |
| 7<br>Monday, 15 April 2024 | Sampling | PGM Chapter 12 | Quiz 3 |
| 8<br>Monday, 22 April 2024 | Mid-term break | | |
| 9<br>Monday, 29 April 2024 | Variational inference | https://leimao.github.io/article/Introduction-to-Variational-Inference/ | Intra-session exam |
| 10<br>Monday, 6 May 2024 | Autoregressive models | | Quiz 4 |
| 11<br>Monday, 13 May 2024 | Variational Auto-Encoders | | |
| 12<br>Monday, 20 May 2024 | GANs | | Quiz 5 |
| 13<br>Monday, 27 May 2024 | Energy-based models | | |
| 14<br>Monday, 3 June 2024 | Evaluating generative models | | Quiz 6 |
| Monday, 17 June 2024 | | | Project due |

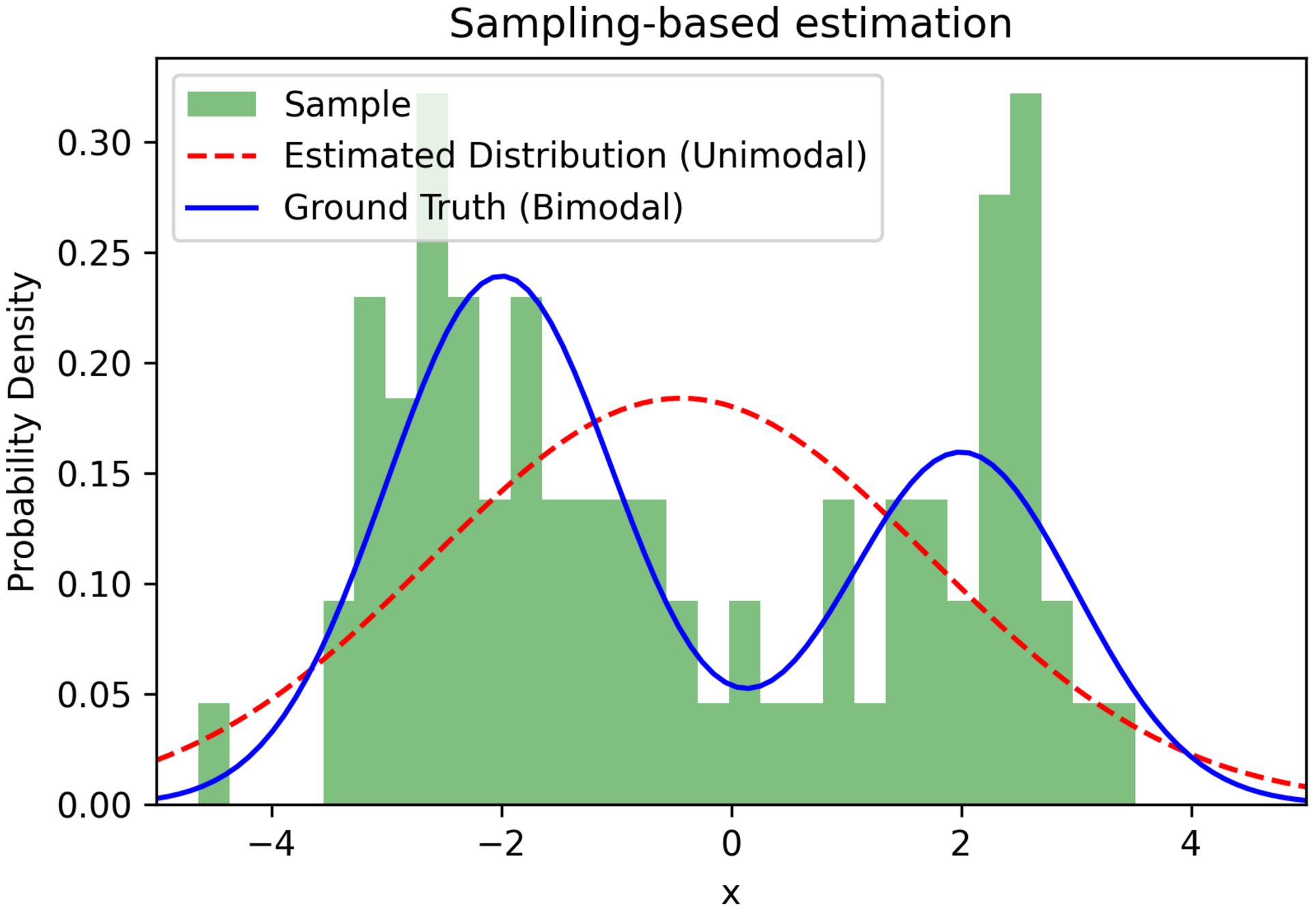

Sampling-based estimation
Probability Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
−3
−2
−1
0
1
2
3
x


Sampling-based estimation
Probability Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
−3
−2
−1
0
1
2
3
x


Sampling-based estimation
Sample
Estimated Distribution
Ground Truth
Probability Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
−3
−2
−1
0
1
2
3
x


Sampling-based estimation
Probability Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
−4
−2
0
2
4
x


Sampling-based estimation
Probability Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
−4
−2
0
2
4
x


Sampling-based estimation
Sample
Estimated Distribution (Unimodal)
Ground Truth (Bimodal)
Probability Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
−4
−2
0
2
4
x

# Simple sampling

# Sampling-based estimation

$\mathscr{D} = \{x[1], \ldots, x[M]\}$ sampled IID from $P$

# Sampling-based estimation

$\mathscr{D} = \{x[1], \ldots, x[M]\}$ sampled IID from $P$

- Independent: The independence aspect means that the random variables in the collection do not influence or depend on each other.
- Identically distributed: This aspect means that all random variables share the same probability distribution.

# Sampling-based estimation

$\mathscr{D} = \{x[1], \ldots, x[M]\}$ sampled IID from $P$

If $P(X = 1) = p$

# Sampling-based estimation

$\mathscr{D} = \{x[1], \ldots, x[M]\}$ sampled IID from $P$

If $P(X = 1) = p$

Estimator for $p$:

# Sampling-based estimation

$\mathscr{D} = \{x[1], \ldots, x[M]\}$ sampled IID from $P$

If $P(X = 1) = p$

Estimator for $p$:

$$T_{\mathscr{D}} = \frac{1}{M} \sum_{m=1}^{M} x[m]$$

# Sampling-based estimation

$\mathscr{D} = \{x[1], \ldots, x[M]\}$ sampled IID from $P$

If $P(X = 1) = p$

Estimator for $p$:

$$T_{\mathscr{D}} = \frac{1}{M} \sum_{m=1}^{M} x[m]$$

More generally, for any distribution $P$, function $f$:

# Sampling-based estimation

$\mathscr{D} = \{x[1], \ldots, x[M]\}$ sampled IID from $P$

If $P(X = 1) = p$

Estimator for $p$:

$$T_{\mathscr{D}} = \frac{1}{M} \sum_{m=1}^{M} x[m]$$

More generally, for any distribution $P$, function $f$:

$$\boldsymbol{E}_P[f] \approx \frac{1}{M} \sum_{m=1}^{M} f(x[m])$$

# Sampling-based estimation

$\mathscr{D} = \{x[1], \ldots, x[M]\}$ sampled IID from $P$

If $P(X = 1) = p$

Estimator for $p$:

$$T_{\mathscr{D}} = \frac{1}{M} \sum_{m=1}^{M} x[m]$$

More generally, for any distribution $P$, function $f$:

$$\boldsymbol{E}_P[f] \approx \frac{1}{M} \sum_{m=1}^{M} f(x[m])$$

$f$ on samples

# Sampling-based estimation

$\mathscr{D} = \{x[1], \ldots, x[M]\}$ sampled IID from $P$

If $P(X = 1) = p$

Estimator for $p$:

$$T_{\mathscr{D}} = \frac{1}{M} \sum_{m=1}^{M} x[m]$$

More generally, for any distribution $P$, function $f$:

$$\boldsymbol{E}_P[f] \approx \frac{1}{M} \sum_{m=1}^{M} f(x[m])$$

$f$ on samples

Empirical expectation

# Sampling from discrete distribution

$\text{Val}(X) = \{x^1, \ldots, x^k\}$ $\quad$ $P(x^i) = \theta^i$

# Sampling from discrete distribution

$$\theta^1, \ldots, \theta^k$$

$$\text{Val}(X) = \{x^1, \ldots, x^k\} \qquad P(x^i) = \theta^i$$

# Sampling from discrete distribution

$$\theta^1, \ldots, \theta^k$$

$$\text{Val}(X) = \{x^1, \ldots, x^k\} \qquad P(x^i) = \theta^i$$

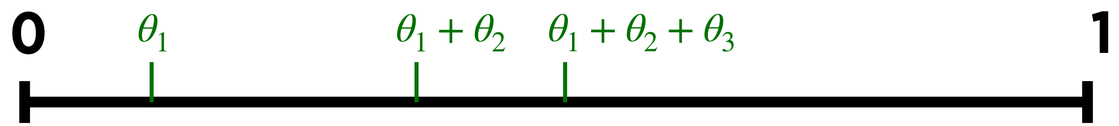
# Sampling from discrete distribution

$$\text{Val}(X) = \{x^1, \ldots, x^k\} \qquad \overset{\theta^1, \ldots, \theta^k}{P(x^i) = \theta^i}$$


**Uniformly in [0...1]**

# Sampling from discrete distribution

$$\text{Val}(X) = \{x^1, \ldots, x^k\} \qquad \overset{\theta^1, \ldots, \theta^k}{P(x^i) = \theta^i}$$

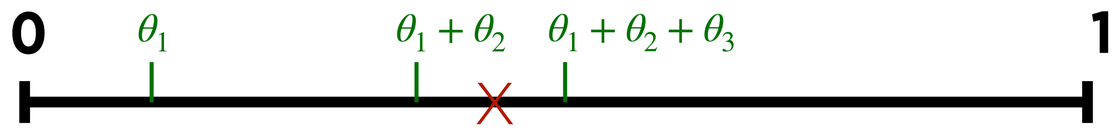



**Uniformly in [0...1]**

# Sampling from discrete distribution

$$\theta^1, \ldots, \theta^k$$

$$\text{Val}(X) = \{x^1, \ldots, x^k\} \qquad P(x^i) = \theta^i$$

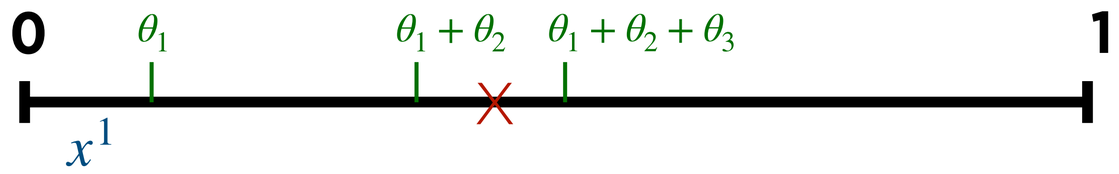



**Uniformly in [0...1]**

# Sampling from discrete distribution

$$\theta^1, \ldots, \theta^k$$

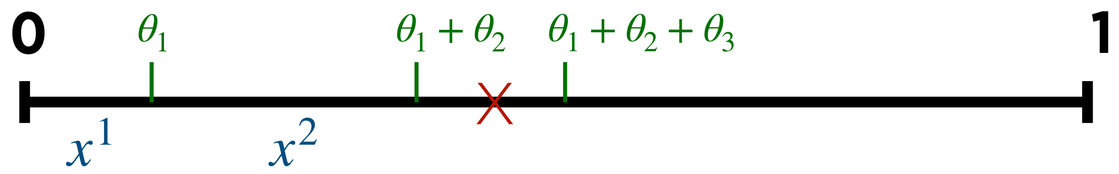
$$\mathrm{Val}(X) = \{x^1, \ldots, x^k\} \qquad P(x^i) = \theta^i$$




**Uniformly in [0...1]**

# Sampling from discrete distribution

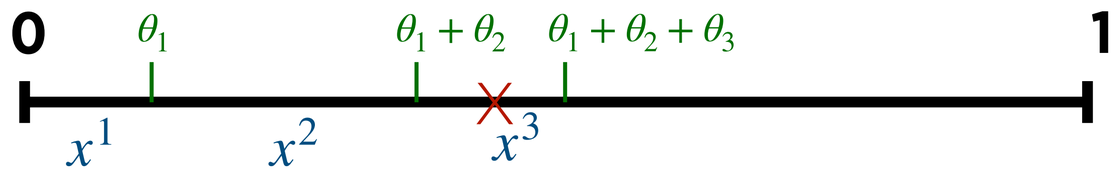
$$\text{Val}(X) = \{x^1, \ldots, x^k\} \qquad \overset{\theta^1, \ldots, \theta^k}{P(x^i) = \theta^i}$$




**Uniformly in [0...1]**

# Sampling from discrete distribution

$$\text{Val}(X) = \{x^1, \ldots, x^k\} \qquad \overset{\theta^1, \ldots, \theta^k}{P(x^i) = \theta^i}$$

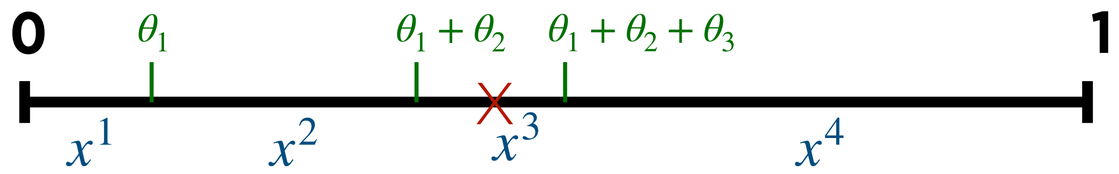



**Uniformly in [0...1]**

# Sampling from discrete distribution

$$\theta^1, \ldots, \theta^k$$

$$\text{Val}(X) = \{x^1, \ldots, x^k\} \qquad P(x^i) = \theta^i$$

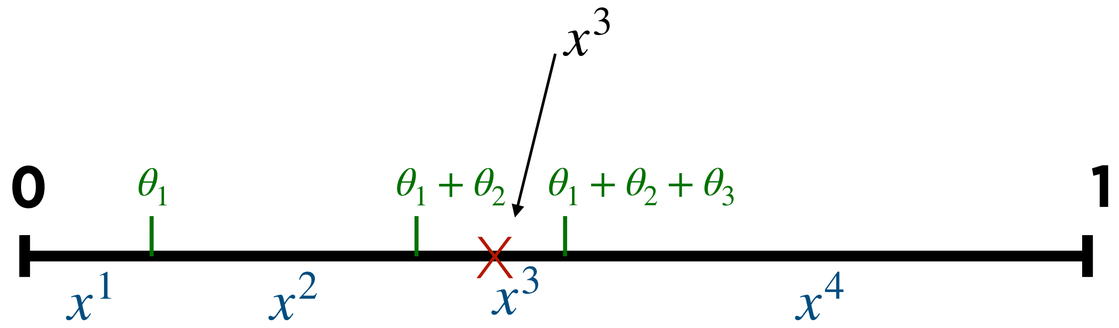



**Uniformly in [0...1]**

# Sampling from discrete distribution

$$\theta^1, \ldots, \theta^k$$

$$\text{Val}(X) = \{x^1, \ldots, x^k\} \qquad P(x^i) = \theta^i$$

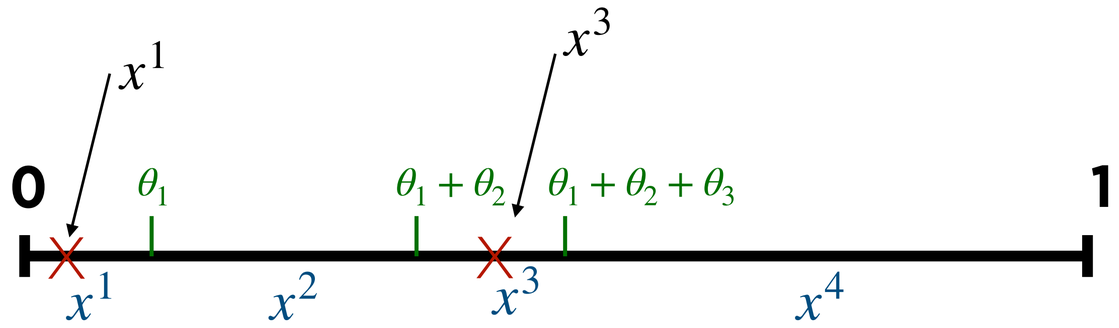



**Uniformly in [0...1]**

# Sampling from discrete distribution

$$\theta^1, \ldots, \theta^k$$

$$\text{Val}(X) = \{x^1, \ldots, x^k\} \qquad P(x^i) = \theta^i$$




**Uniformly in [0...1]**

# How good is our estimator?

# Sampling-based estimation

**Properties — how far off is our estimate from the true value?**

$$T_{\mathscr{D}} = \frac{1}{M} \sum_{m=1}^{M} x[m]$$

# Sampling-based estimation

**Properties — how far off is our estimate from the true value?**

$$T_{\mathscr{D}} = \frac{1}{M} \sum_{m=1}^{M} x[m]$$

$$P_{\mathscr{D}}(T_{\mathscr{D}} \notin [p - \epsilon, p + \epsilon]) \leq 2e^{-2M\epsilon^2}$$

# Sampling-based estimation

## Properties — how far off is our estimate from the true value?

$$T_{\mathscr{D}} = \frac{1}{M}\sum_{m=1}^{M} x[m]$$

$$P_{\mathscr{D}}(T_{\mathscr{D}} \notin [p-\epsilon, p+\epsilon]) \leq 2e^{-2M\epsilon^2}$$

is $\epsilon$ away from $p$

# Sampling-based estimation

## Properties — how far off is our estimate from the true value?

Hoeffding Bound

$$T_{\mathscr{D}} = \frac{1}{M}\sum_{m=1}^{M} x[m]$$

$$P_{\mathscr{D}}(T_{\mathscr{D}} \notin [p-\epsilon, p+\epsilon]) \leq 2e^{-2M\epsilon^2}$$

is $\epsilon$ away from $p$

# Sampling-based estimation

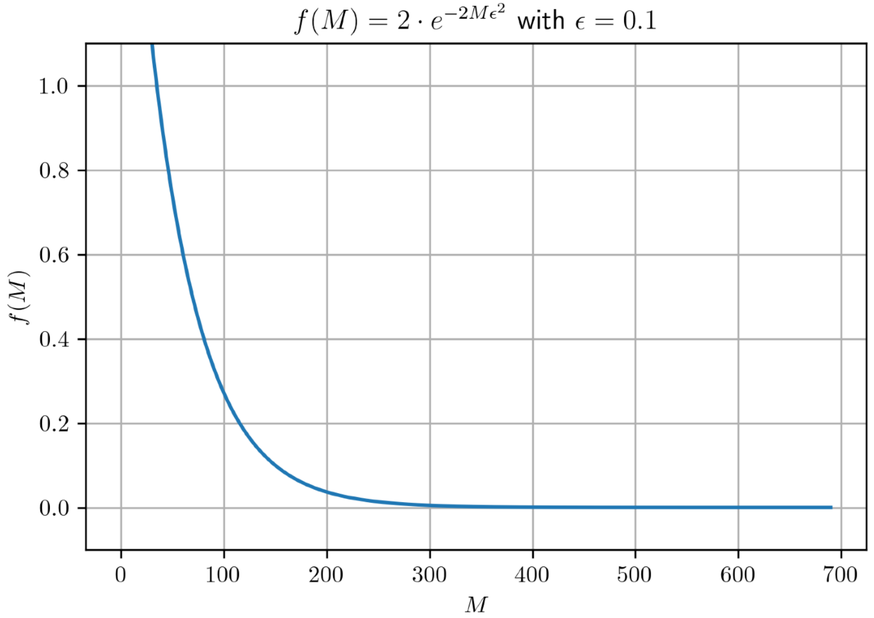
## Properties — how far off is our estimate from the true value?

Hoeffding Bound

$$T_{\mathscr{D}} = \frac{1}{M}\sum_{m=1}^{M} x[m]$$

$$P_{\mathscr{D}}(T_{\mathscr{D}} \notin [p-\epsilon, p+\epsilon]) \leq 2e^{-2M\epsilon^2}$$

is $\epsilon$ away from p



# Sampling-based estimation

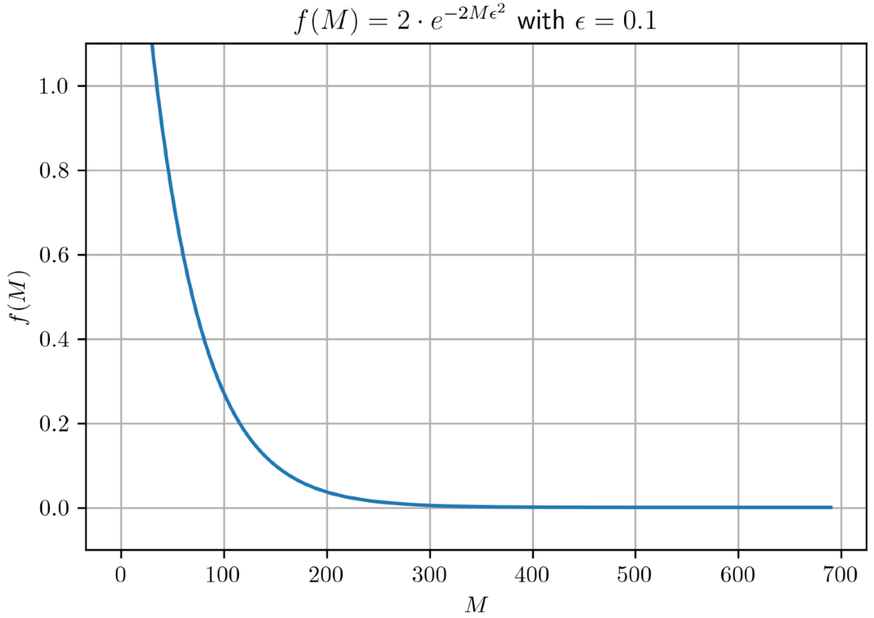
## Properties — how far off is our estimate from the true value?

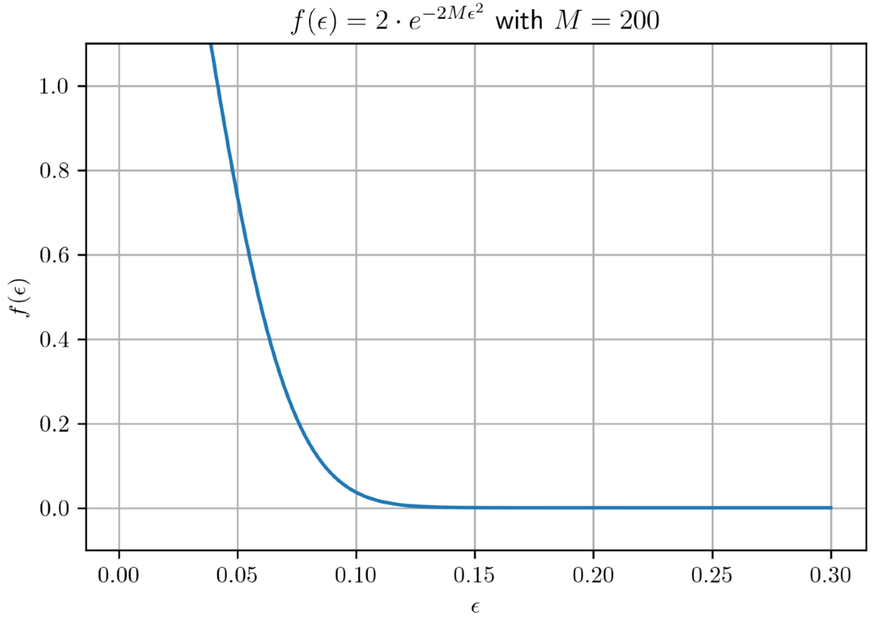
Hoeffding Bound

$$T_{\mathscr{D}} = \frac{1}{M}\sum_{m=1}^{M} x[m]$$

$$P_{\mathscr{D}}(T_{\mathscr{D}} \notin [p-\epsilon, p+\epsilon]) \leq 2e^{-2M\epsilon^2}$$

is $\epsilon$ away from p



# Sampling-based estimation

## Properties — how far off is our estimate from the true value?

$$T_{\mathscr{D}} = \frac{1}{M}\sum_{m=1}^{M} x[m]$$

Hoeffding Bound:

$$P_{\mathscr{D}}(T_{\mathscr{D}} \notin [p-\epsilon, p+\epsilon]) \leq 2e^{-2M\epsilon^2}$$

is $\epsilon$ away from $p$

# Sampling-based estimation

## Properties — how far off is our estimate from the true value?

$$T_{\mathscr{D}} = \frac{1}{M}\sum_{m=1}^{M} x[m]$$

Hoeffding Bound:

$$P_{\mathscr{D}}(T_{\mathscr{D}} \notin [p-\epsilon, p+\epsilon]) \leq 2e^{-2M\epsilon^2}$$

is $\epsilon$ away from p

Chernoff Bound:

$$P_{\mathscr{D}}(T_{\mathscr{D}} \notin [p(1-\epsilon), p(1+\epsilon)]) \leq 2e^{-2Mp\epsilon^2/3}$$

# Sampling-based estimation

## Properties — how far off is our estimate from the true value?

$$T_{\mathscr{D}} = \frac{1}{M}\sum_{m=1}^{M} x[m]$$

Hoeffding Bound:

$$P_{\mathscr{D}}(T_{\mathscr{D}} \notin [p-\epsilon, p+\epsilon]) \leq 2e^{-2M\epsilon^2}$$

is $\epsilon$ away from $p$

Chernoff Bound:

$$P_{\mathscr{D}}(T_{\mathscr{D}} \notin [p(1-\epsilon), p(1+\epsilon)]) \leq 2e^{-2Mp\epsilon^2/3}$$

# Sampling-based estimation

## Properties — how far off is our estimate from the true value?

$$T_{\mathscr{D}} = \frac{1}{M}\sum_{m=1}^{M} x[m]$$

Hoeffding Bound:

$$P_{\mathscr{D}}(T_{\mathscr{D}} \notin [p-\epsilon, p+\epsilon]) \leq 2e^{-2M\epsilon^2}$$

is $\epsilon$ away from $p$

Chernoff Bound:

$$P_{\mathscr{D}}(T_{\mathscr{D}} \notin [p(1-\epsilon), p(1+\epsilon)]) \leq 2e^{-2Mp\epsilon^2/3}$$

# Sampling-based estimation

## Properties — how far off is our estimate from the true value?

$$T_{\mathscr{D}} = \frac{1}{M}\sum_{m=1}^{M} x[m]$$

Hoeffding Bound:

$$P_{\mathscr{D}}(T_{\mathscr{D}} \notin [p-\epsilon, p+\epsilon]) \leq 2e^{-2M\epsilon^2}$$

is $\epsilon$ away from p

For additive bound $\epsilon$ on error with probability $> 1-\delta$:

$$M \geq \frac{\ln(2/\delta)}{2\epsilon^2}.$$

Chernoff Bound:

$$P_{\mathscr{D}}(T_{\mathscr{D}} \notin [p(1-\epsilon), p(1+\epsilon)]) \leq 2e^{-2Mp\epsilon^2/3}$$

# Sampling-based estimation

## Properties — how far off is our estimate from the true value?

$$T_{\mathscr{D}} = \frac{1}{M}\sum_{m=1}^{M} x[m]$$

Hoeffding Bound:

$$P_{\mathscr{D}}(T_{\mathscr{D}} \notin [p - \epsilon, p + \epsilon]) \leq 2e^{-2M\epsilon^2}$$

is $\epsilon$ away from p

For additive bound $\epsilon$ on error with probability $> 1 - \delta$:

$$M \geq \frac{\ln(2/\delta)}{2\epsilon^2}.$$

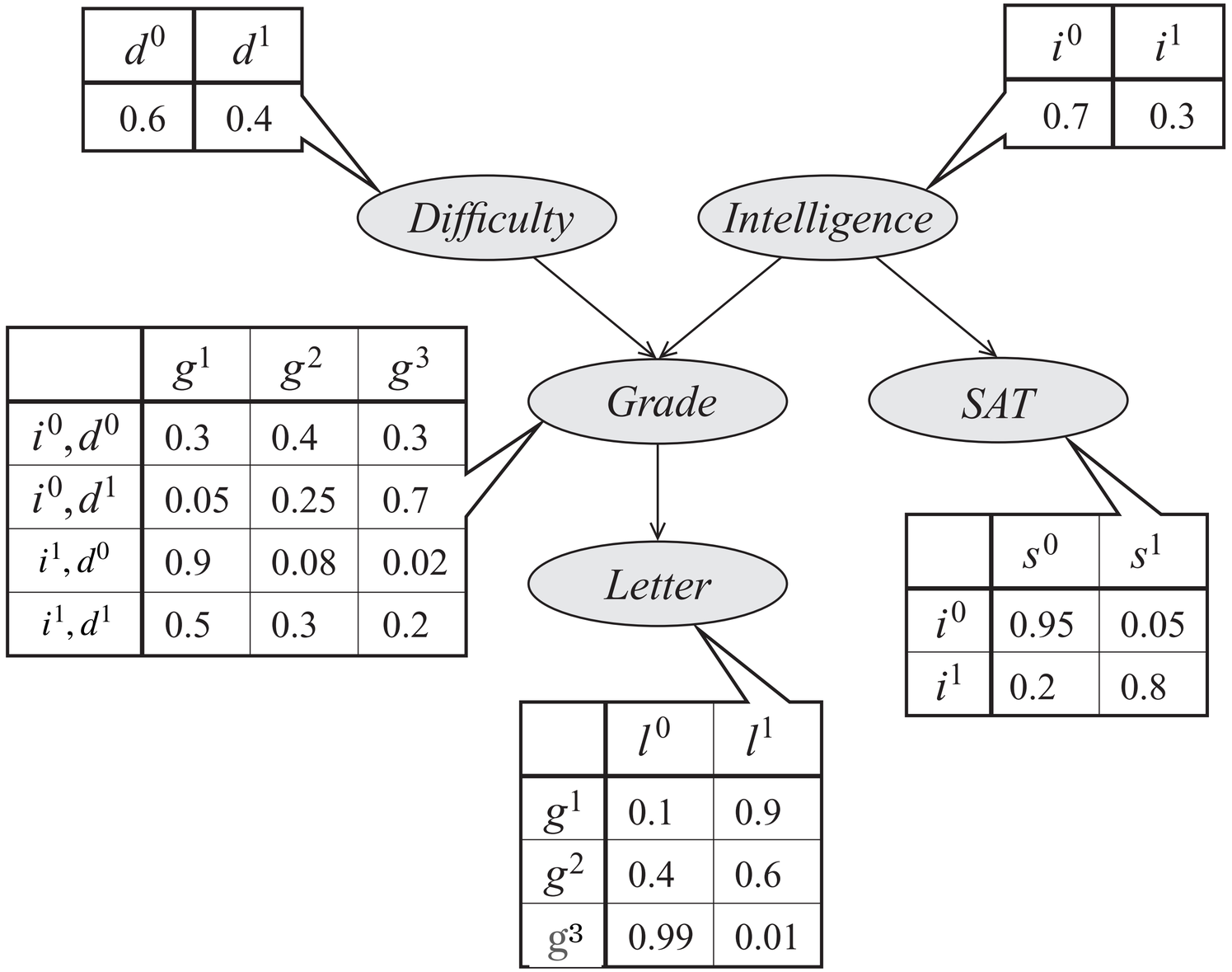
Chernoff Bound:

$$P_{\mathscr{D}}(T_{\mathscr{D}} \notin [p(1 - \epsilon), p(1 + \epsilon)]) \leq 2e^{-2Mp\epsilon^2/3}$$

For multiplicative bound $\epsilon$ on error with probability $> 1 - \delta$:

$$M \geq 3\frac{\ln(2/\delta)}{2p\epsilon^2}.$$

# Forward sampling from a BN




| $d^0$ | $d^1$ |
|---|---|
| 0.6 | 0.4 |

| $i^0$ | $i^1$ |
|---|---|
| 0.7 | 0.3 |

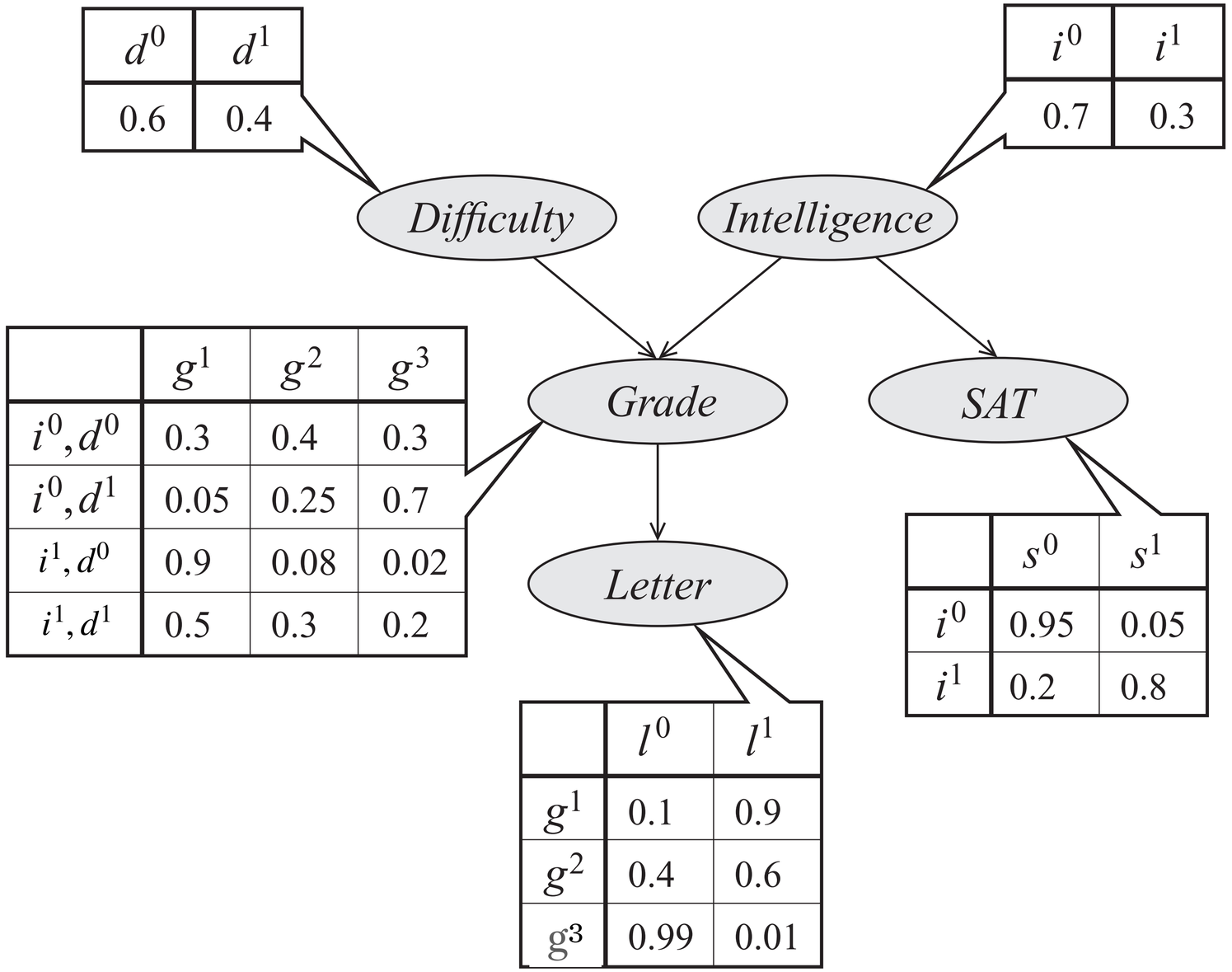
| | $g^1$ | $g^2$ | $g^3$ |
|---|---|---|---|
| $i^0, d^0$ | 0.3 | 0.4 | 0.3 |
| $i^0, d^1$ | 0.05 | 0.25 | 0.7 |
| $i^1, d^0$ | 0.9 | 0.08 | 0.02 |
| $i^1, d^1$ | 0.5 | 0.3 | 0.2 |

| | $s^0$ | $s^1$ |
|---|---|---|
| $i^0$ | 0.95 | 0.05 |
| $i^1$ | 0.2 | 0.8 |

| | $l^0$ | $l^1$ |
|---|---|---|
| $g^1$ | 0.1 | 0.9 |
| $g^2$ | 0.4 | 0.6 |
| $g^3$ | 0.99 | 0.01 |

# Forward sampling from a BN

$P(D, I, G, S, L)$




| $d^0$ | $d^1$ |
|---|---|
| 0.6 | 0.4 |

| $i^0$ | $i^1$ |
|---|---|
| 0.7 | 0.3 |

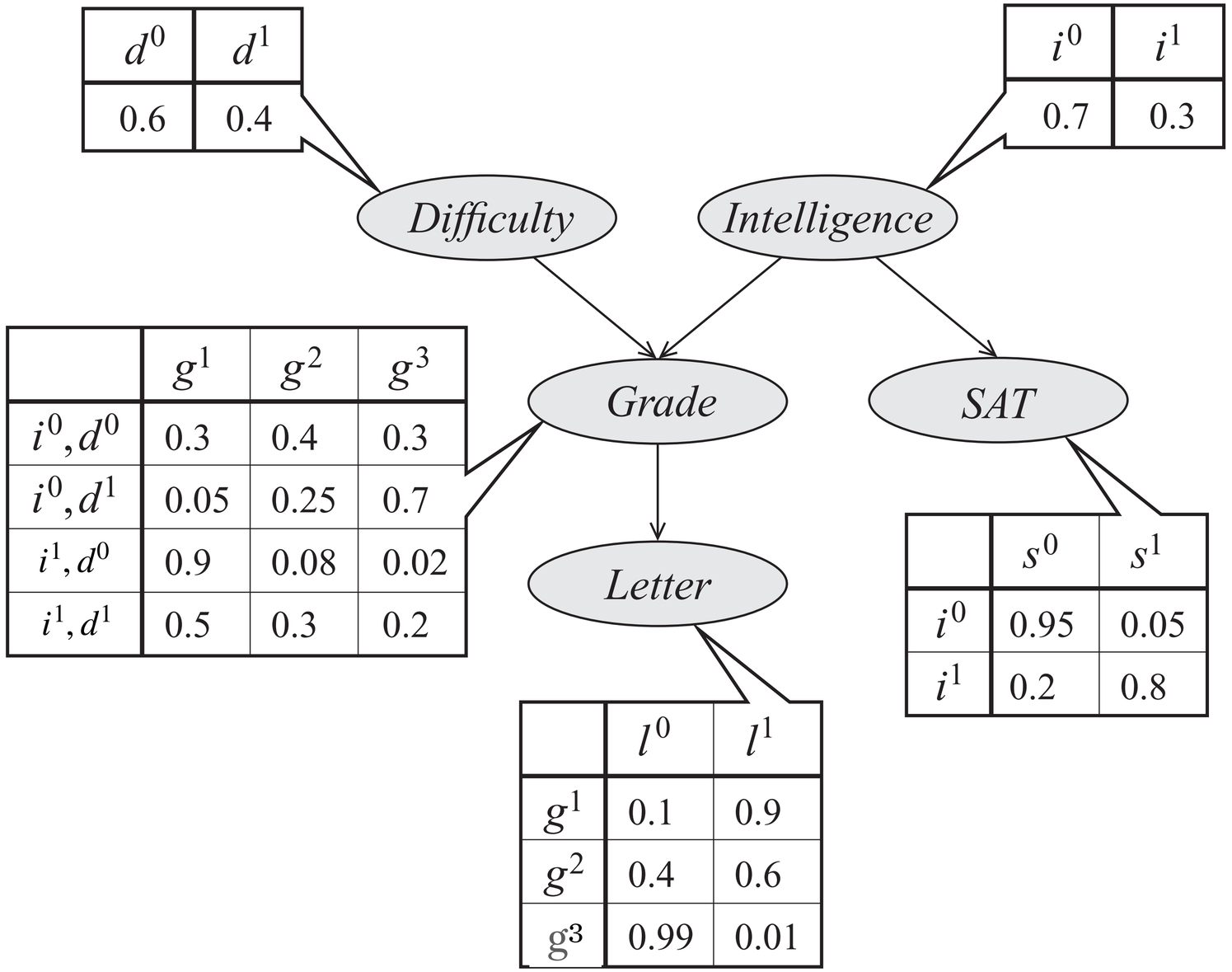
| | $g^1$ | $g^2$ | $g^3$ |
|---|---|---|---|
| $i^0, d^0$ | 0.3 | 0.4 | 0.3 |
| $i^0, d^1$ | 0.05 | 0.25 | 0.7 |
| $i^1, d^0$ | 0.9 | 0.08 | 0.02 |
| $i^1, d^1$ | 0.5 | 0.3 | 0.2 |

| | $s^0$ | $s^1$ |
|---|---|---|
| $i^0$ | 0.95 | 0.05 |
| $i^1$ | 0.2 | 0.8 |

| | $l^0$ | $l^1$ |
|---|---|---|
| $g^1$ | 0.1 | 0.9 |
| $g^2$ | 0.4 | 0.6 |
| $g^3$ | 0.99 | 0.01 |

# Forward sampling from a BN

$P(D, I, G, S, L)$




| $d^0$ | $d^1$ |
|---|---|
| 0.6 | 0.4 |

| $i^0$ | $i^1$ |
|---|---|
| 0.7 | 0.3 |

| | $g^1$ | $g^2$ | $g^3$ |
|---|---|---|---|
| $i^0, d^0$ | 0.3 | 0.4 | 0.3 |
| $i^0, d^1$ | 0.05 | 0.25 | 0.7 |
| $i^1, d^0$ | 0.9 | 0.08 | 0.02 |
| $i^1, d^1$ | 0.5 | 0.3 | 0.2 |

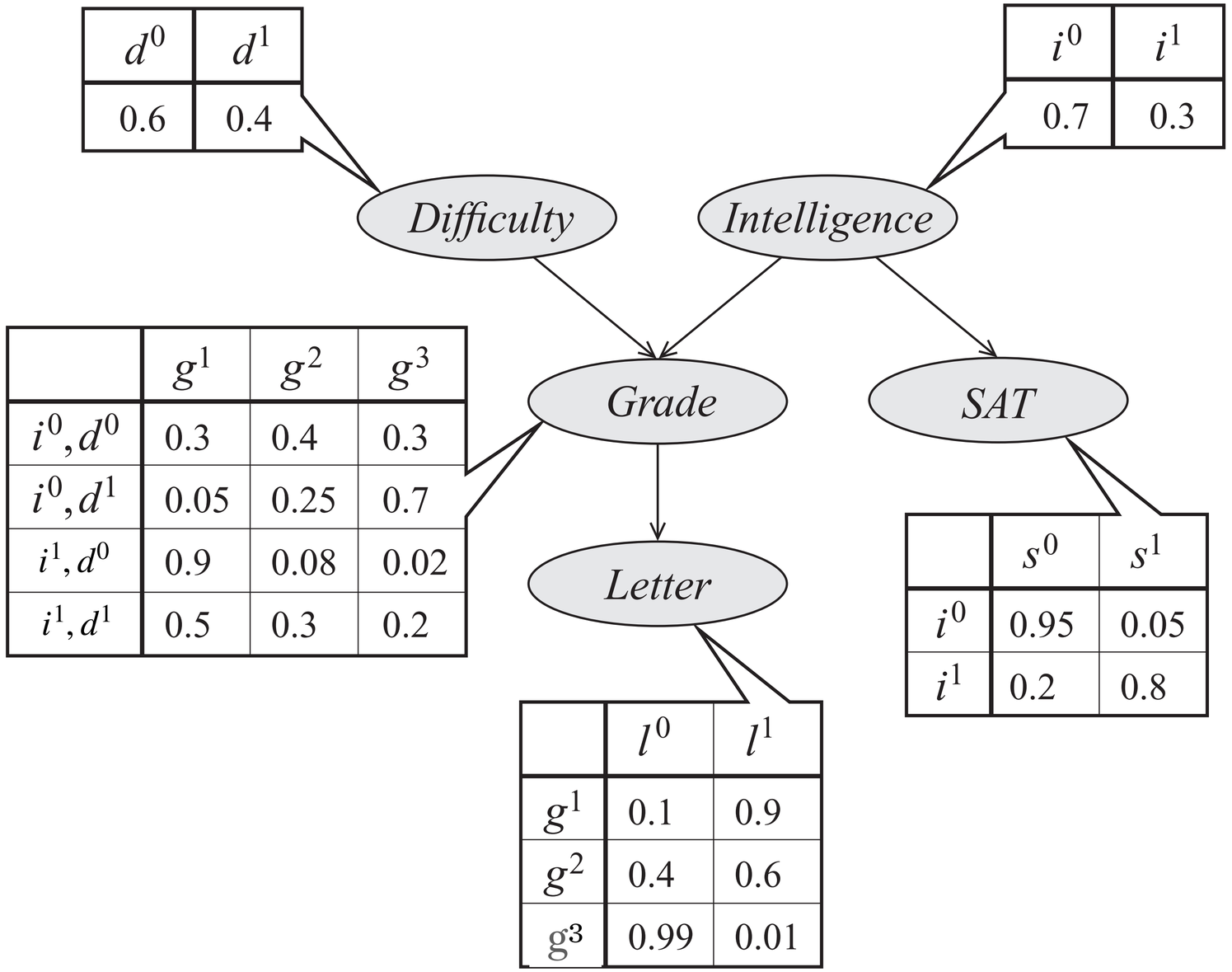
| | $s^0$ | $s^1$ |
|---|---|---|
| $i^0$ | 0.95 | 0.05 |
| $i^1$ | 0.2 | 0.8 |

| | $l^0$ | $l^1$ |
|---|---|---|
| $g^1$ | 0.1 | 0.9 |
| $g^2$ | 0.4 | 0.6 |
| $g^3$ | 0.99 | 0.01 |

$d^0$

# Forward sampling from a BN

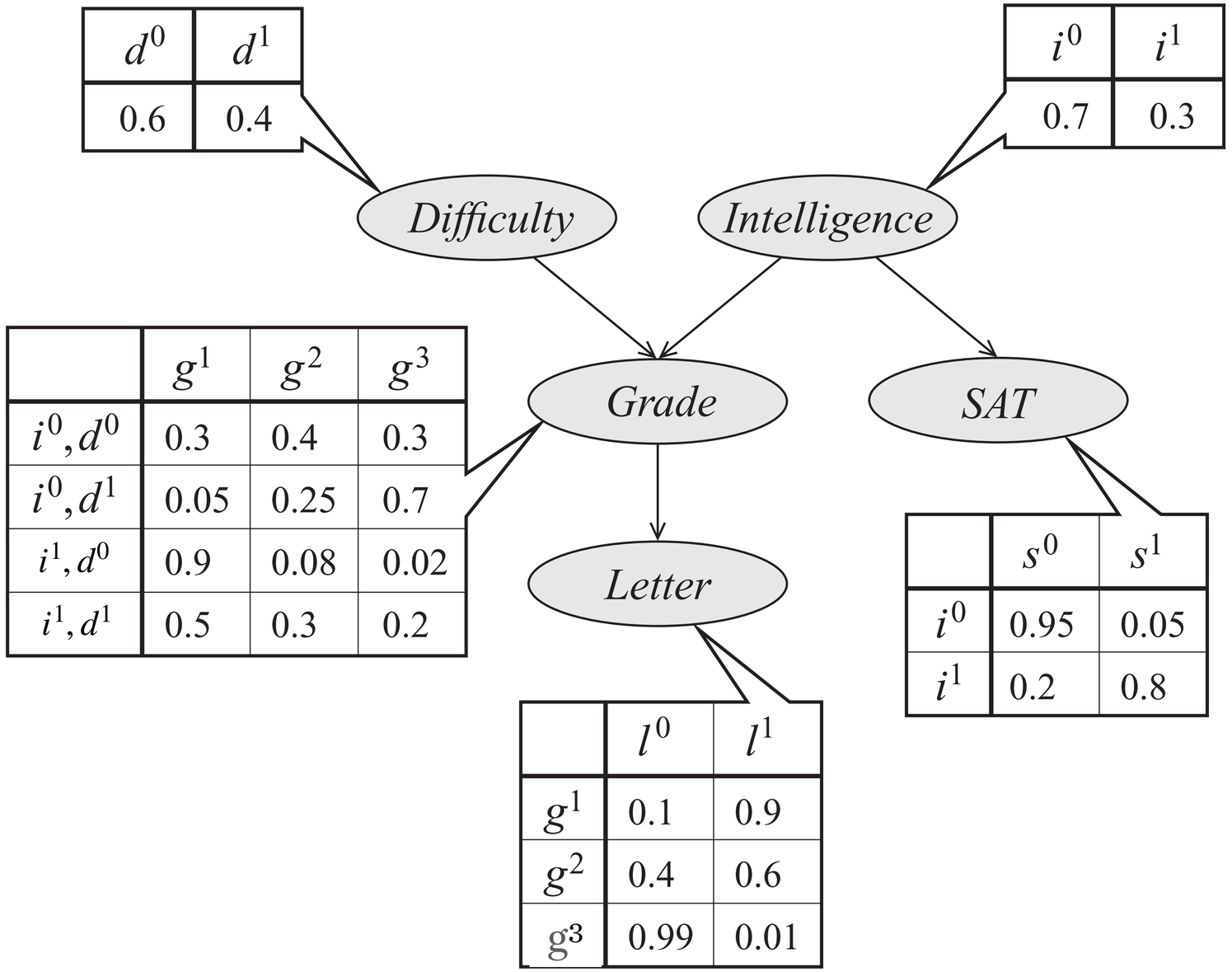
$P(D, I, G, S, L)$




$d^0 i^1$

# Forward sampling from a BN

$P(D, I, G, S, L)$




| $d^0$ | $d^1$ |
|---|---|
| 0.6 | 0.4 |

| $i^0$ | $i^1$ |
|---|---|
| 0.7 | 0.3 |

| | $g^1$ | $g^2$ | $g^3$ |
|---|---|---|---|
| $i^0, d^0$ | 0.3 | 0.4 | 0.3 |
| $i^0, d^1$ | 0.05 | 0.25 | 0.7 |
| $i^1, d^0$ | 0.9 | 0.08 | 0.02 |
| $i^1, d^1$ | 0.5 | 0.3 | 0.2 |

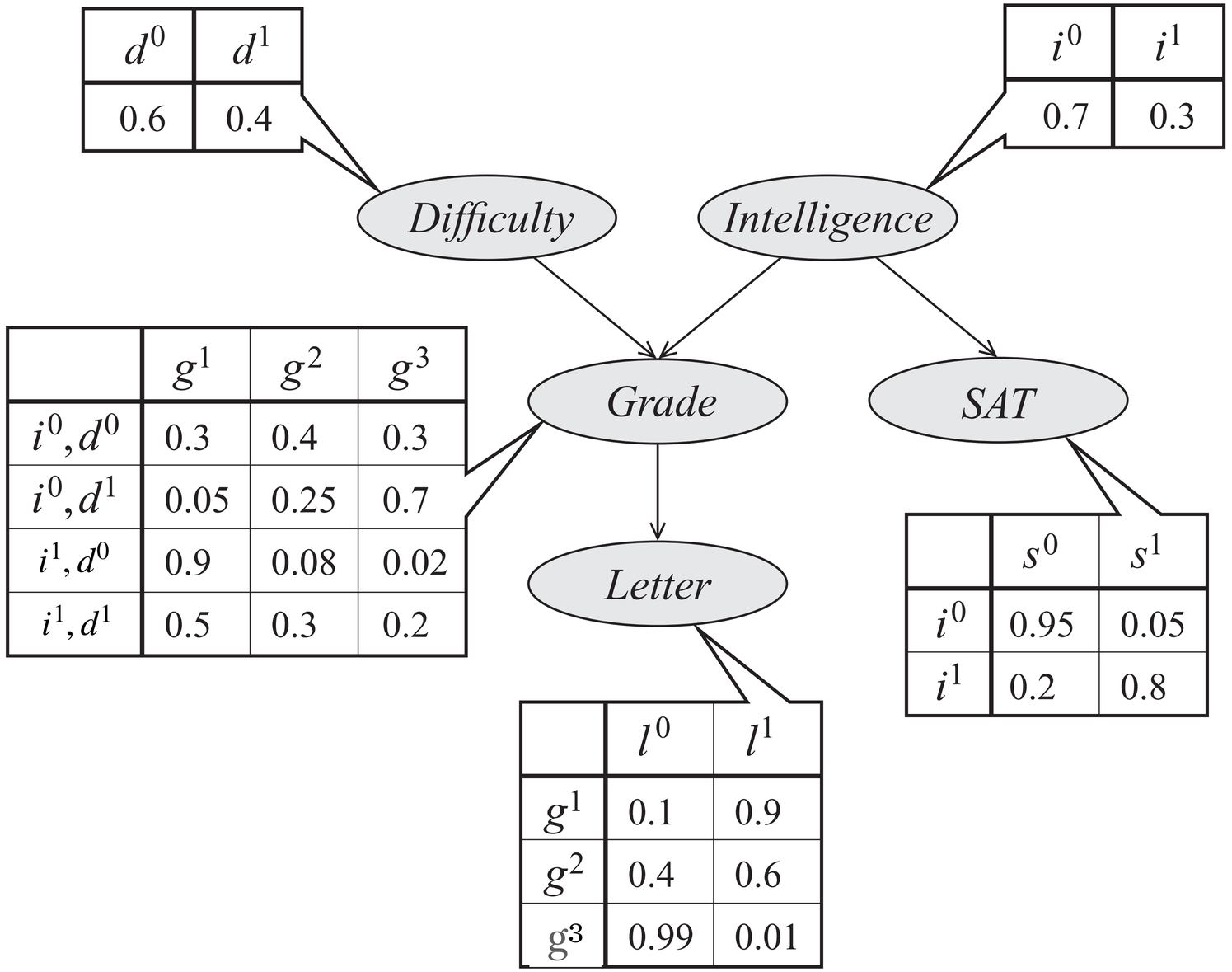
| | $s^0$ | $s^1$ |
|---|---|---|
| $i^0$ | 0.95 | 0.05 |
| $i^1$ | 0.2 | 0.8 |

| | $l^0$ | $l^1$ |
|---|---|---|
| $g^1$ | 0.1 | 0.9 |
| $g^2$ | 0.4 | 0.6 |
| $g^3$ | 0.99 | 0.01 |

$d^0 i^1 g^1$

# Forward sampling from a BN

$P(D, I, G, S, L)$




| $d^0$ | $d^1$ |
|---|---|
| 0.6 | 0.4 |

| $i^0$ | $i^1$ |
|---|---|
| 0.7 | 0.3 |

| | $g^1$ | $g^2$ | $g^3$ |
|---|---|---|---|
| $i^0, d^0$ | 0.3 | 0.4 | 0.3 |
| $i^0, d^1$ | 0.05 | 0.25 | 0.7 |
| $i^1, d^0$ | 0.9 | 0.08 | 0.02 |
| $i^1, d^1$ | 0.5 | 0.3 | 0.2 |

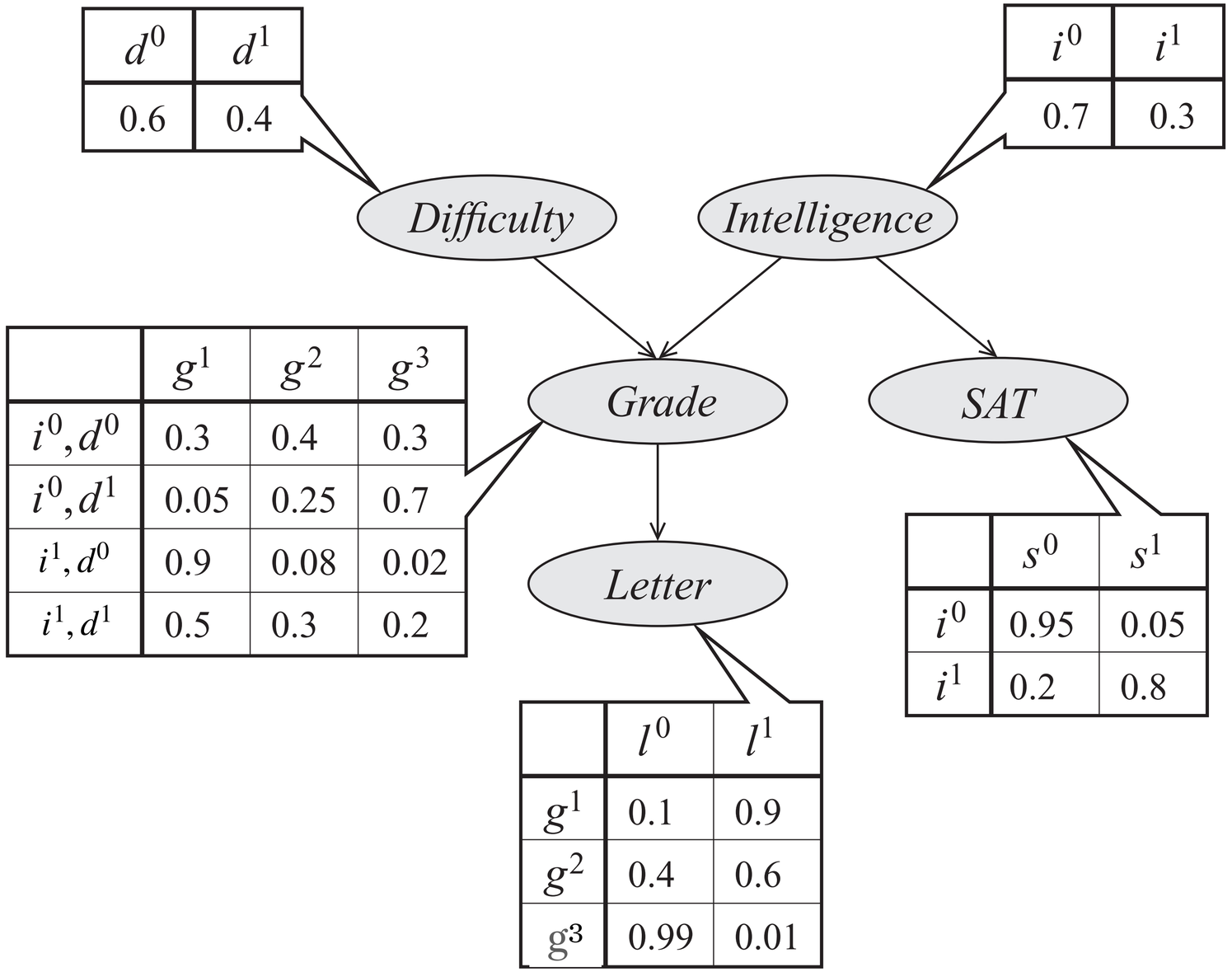
| | $s^0$ | $s^1$ |
|---|---|---|
| $i^0$ | 0.95 | 0.05 |
| $i^1$ | 0.2 | 0.8 |

| | $l^0$ | $l^1$ |
|---|---|---|
| $g^1$ | 0.1 | 0.9 |
| $g^2$ | 0.4 | 0.6 |
| $g^3$ | 0.99 | 0.01 |

$d^0 i^1 g^1 s^0$

# Forward sampling from a BN

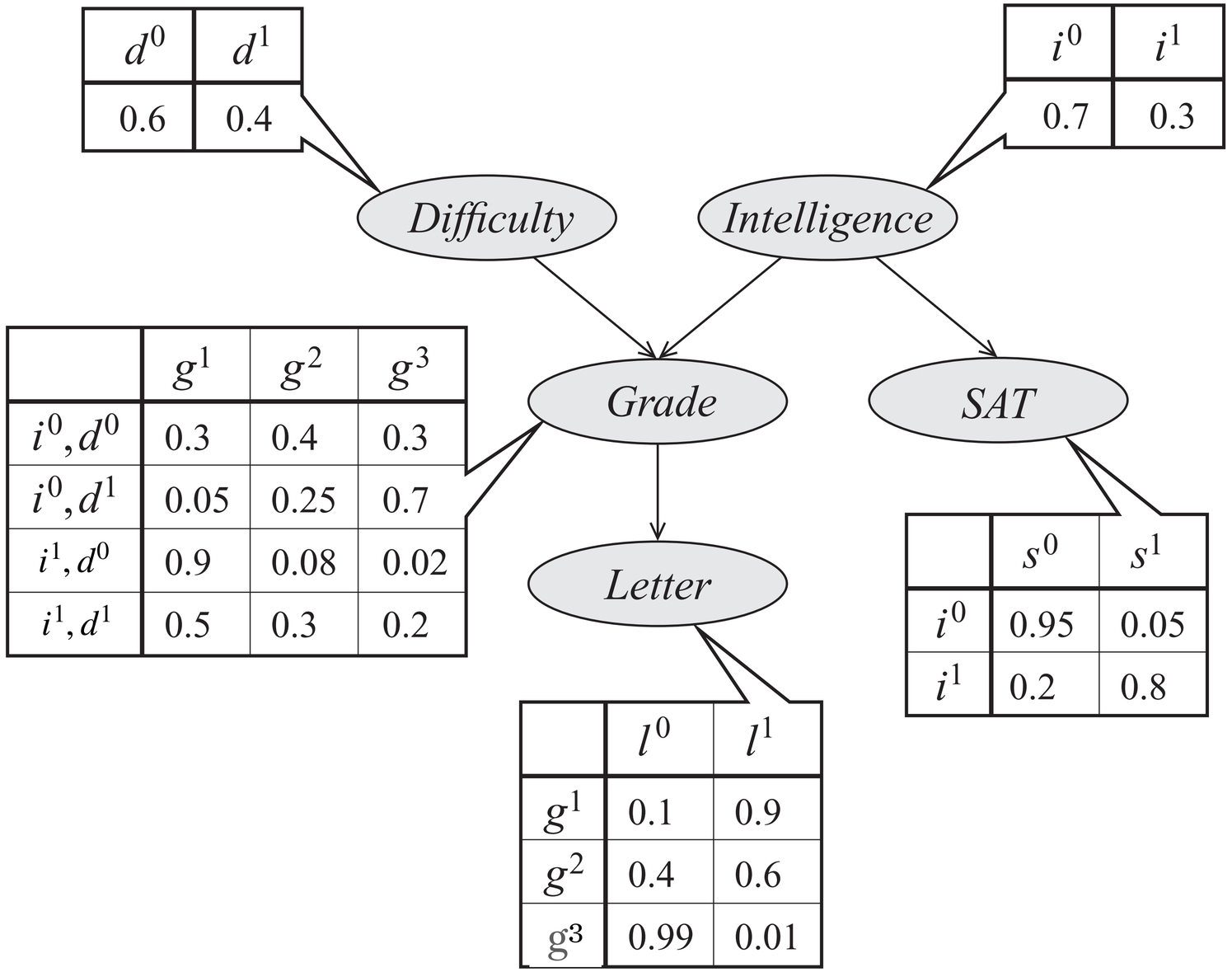
$P(D, I, G, S, L)$




$d^0 i^1 g^1 s^0 l^1$

# Forward sampling from a BN

$P(D, I, G, S, L)$




| $d^0$ | $d^1$ |
|---|---|
| 0.6 | 0.4 |

| $i^0$ | $i^1$ |
|---|---|
| 0.7 | 0.3 |

| | $g^1$ | $g^2$ | $g^3$ |
|---|---|---|---|
| $i^0, d^0$ | 0.3 | 0.4 | 0.3 |
| $i^0, d^1$ | 0.05 | 0.25 | 0.7 |
| $i^1, d^0$ | 0.9 | 0.08 | 0.02 |
| $i^1, d^1$ | 0.5 | 0.3 | 0.2 |

| | $s^0$ | $s^1$ |
|---|---|---|
| $i^0$ | 0.95 | 0.05 |
| $i^1$ | 0.2 | 0.8 |

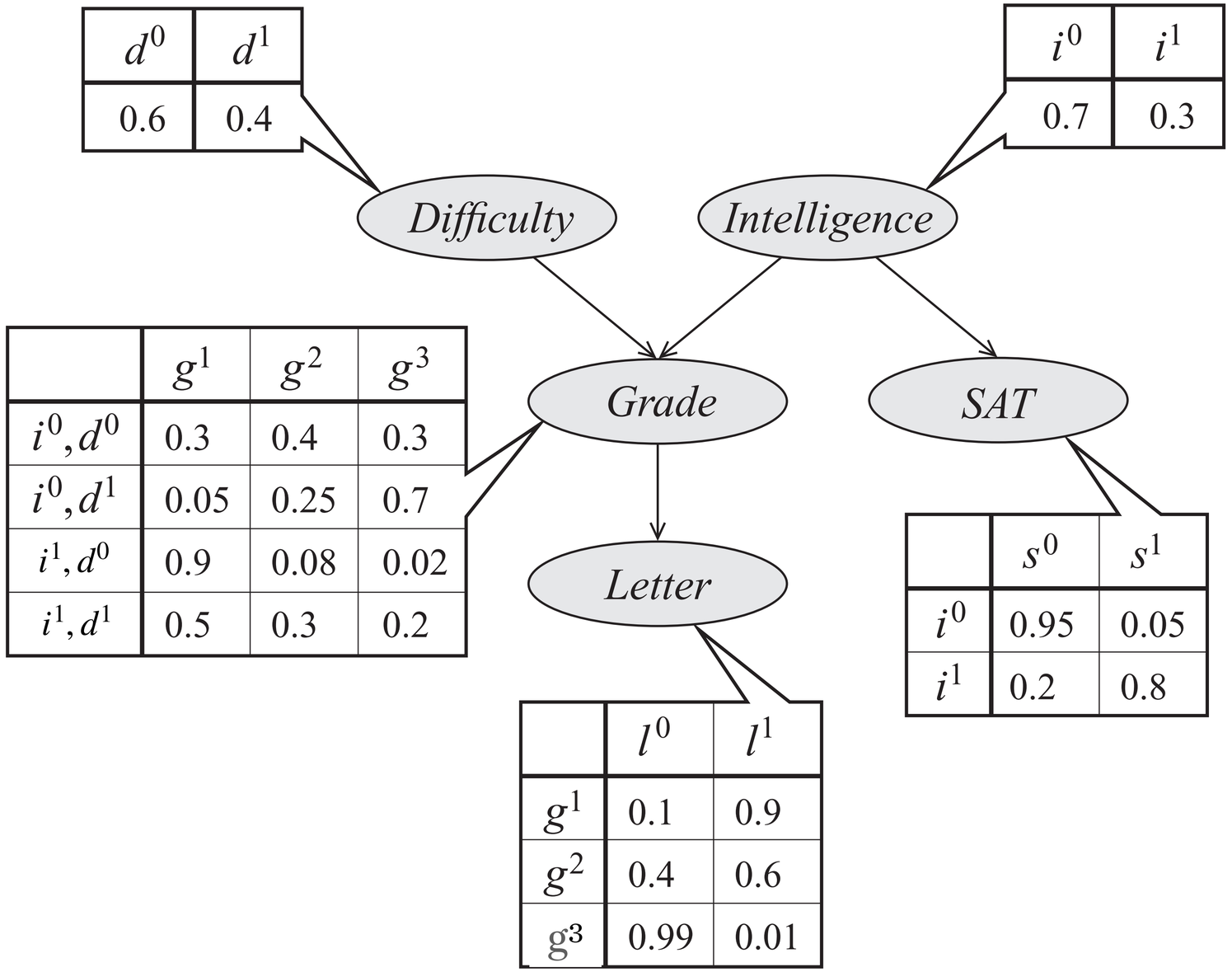
| | $l^0$ | $l^1$ |
|---|---|---|
| $g^1$ | 0.1 | 0.9 |
| $g^2$ | 0.4 | 0.6 |
| $g^3$ | 0.99 | 0.01 |

$d^0 i^1 g^1 s^0 l^1$

$d^1$

# Forward sampling from a BN

$P(D, I, G, S, L)$




| $d^0$ | $d^1$ |
|---|---|
| 0.6 | 0.4 |

| $i^0$ | $i^1$ |
|---|---|
| 0.7 | 0.3 |

| | $g^1$ | $g^2$ | $g^3$ |
|---|---|---|---|
| $i^0, d^0$ | 0.3 | 0.4 | 0.3 |
| $i^0, d^1$ | 0.05 | 0.25 | 0.7 |
| $i^1, d^0$ | 0.9 | 0.08 | 0.02 |
| $i^1, d^1$ | 0.5 | 0.3 | 0.2 |

| | $s^0$ | $s^1$ |
|---|---|---|
| $i^0$ | 0.95 | 0.05 |
| $i^1$ | 0.2 | 0.8 |

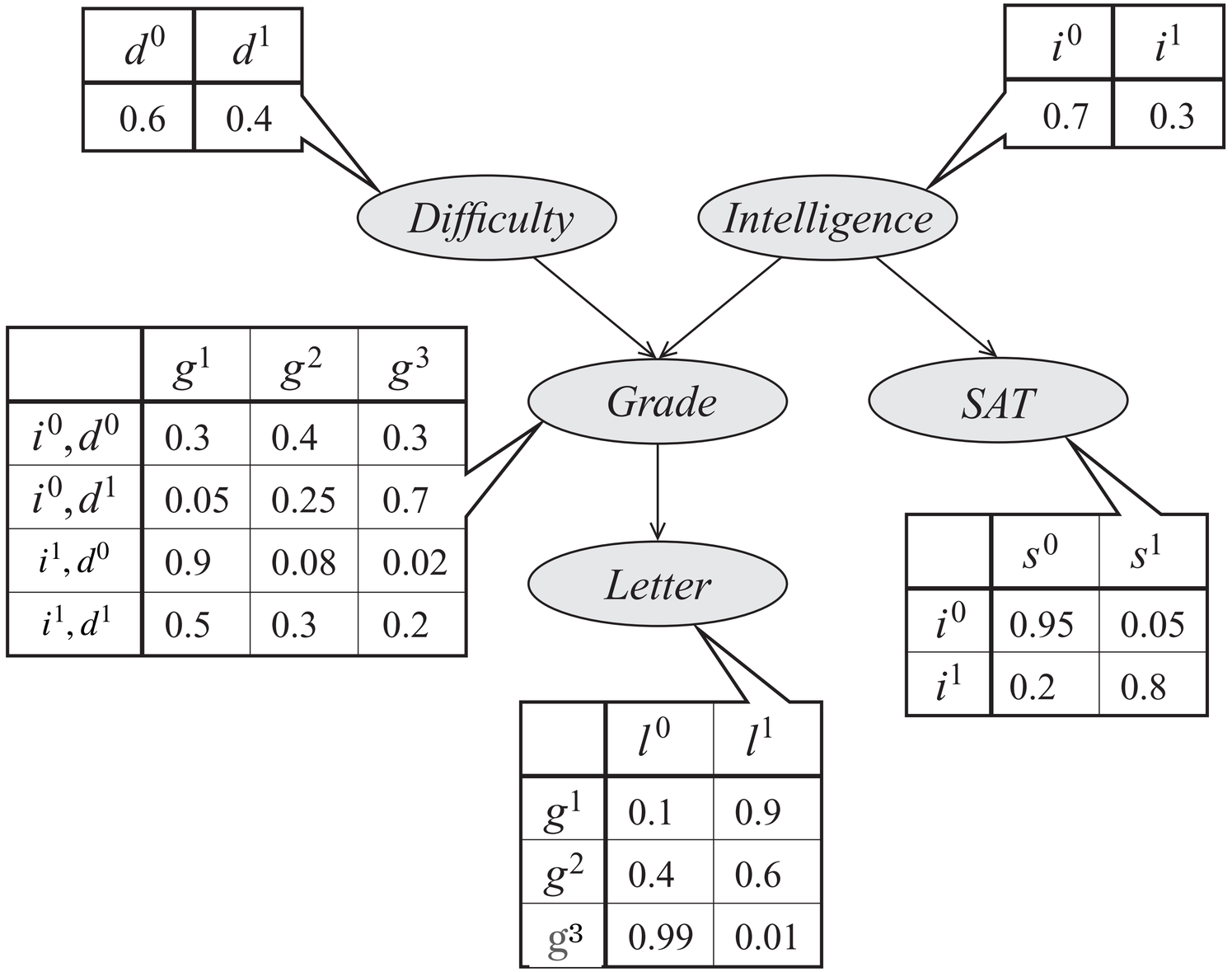
| | $l^0$ | $l^1$ |
|---|---|---|
| $g^1$ | 0.1 | 0.9 |
| $g^2$ | 0.4 | 0.6 |
| $g^3$ | 0.99 | 0.01 |

$d^0 i^1 g^1 s^0 l^1$

$d^1 i^1$

# Forward sampling from a BN

$P(D, I, G, S, L)$




| $d^0$ | $d^1$ |
|---|---|
| 0.6 | 0.4 |

| $i^0$ | $i^1$ |
|---|---|
| 0.7 | 0.3 |

| | $g^1$ | $g^2$ | $g^3$ |
|---|---|---|---|
| $i^0, d^0$ | 0.3 | 0.4 | 0.3 |
| $i^0, d^1$ | 0.05 | 0.25 | 0.7 |
| $i^1, d^0$ | 0.9 | 0.08 | 0.02 |
| $i^1, d^1$ | 0.5 | 0.3 | 0.2 |

| | $s^0$ | $s^1$ |
|---|---|---|
| $i^0$ | 0.95 | 0.05 |
| $i^1$ | 0.2 | 0.8 |

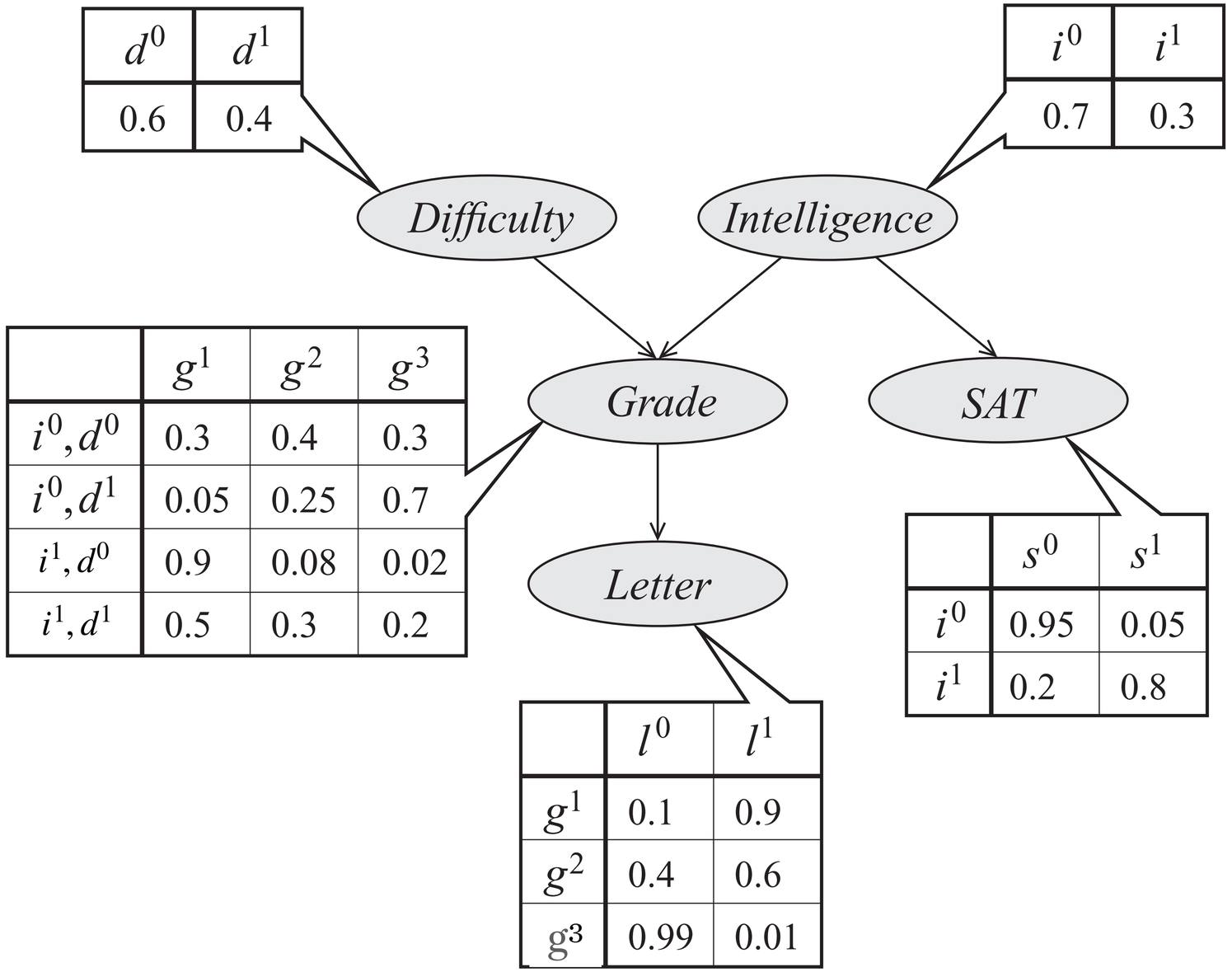
| | $l^0$ | $l^1$ |
|---|---|---|
| $g^1$ | 0.1 | 0.9 |
| $g^2$ | 0.4 | 0.6 |
| $g^3$ | 0.99 | 0.01 |

$d^0 i^1 g^1 s^0 l^1$

$d^1 i^1 g^2$

# Forward sampling from a BN

$P(D, I, G, S, L)$




| $d^0$ | $d^1$ |
|---|---|
| 0.6 | 0.4 |

| $i^0$ | $i^1$ |
|---|---|
| 0.7 | 0.3 |

| | $g^1$ | $g^2$ | $g^3$ |
|---|---|---|---|
| $i^0, d^0$ | 0.3 | 0.4 | 0.3 |
| $i^0, d^1$ | 0.05 | 0.25 | 0.7 |
| $i^1, d^0$ | 0.9 | 0.08 | 0.02 |
| $i^1, d^1$ | 0.5 | 0.3 | 0.2 |

| | $s^0$ | $s^1$ |
|---|---|---|
| $i^0$ | 0.95 | 0.05 |
| $i^1$ | 0.2 | 0.8 |

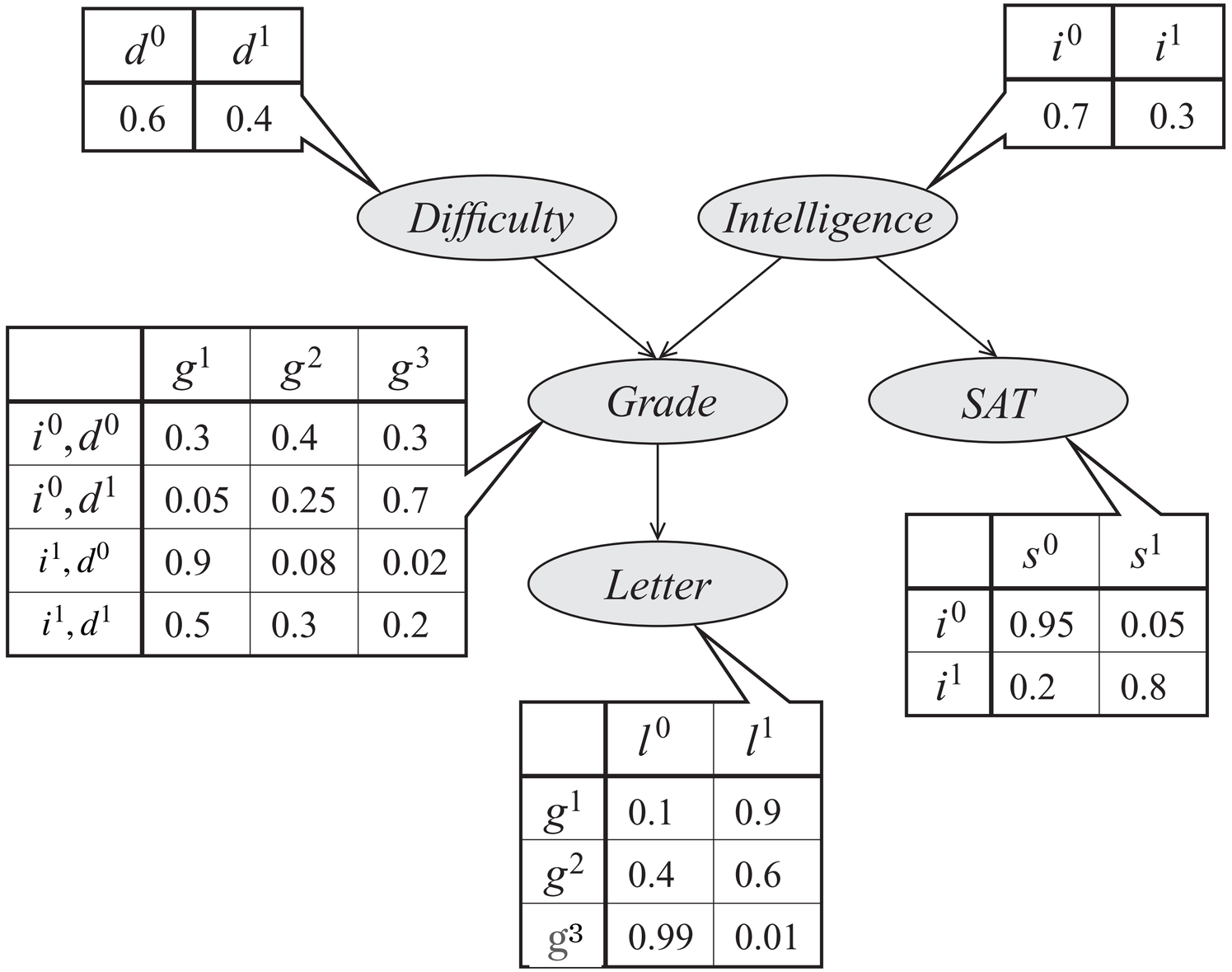
| | $l^0$ | $l^1$ |
|---|---|---|
| $g^1$ | 0.1 | 0.9 |
| $g^2$ | 0.4 | 0.6 |
| $g^3$ | 0.99 | 0.01 |

$d^0 i^1 g^1 s^0 l^1$

$d^1 i^1 g^2 s^1$

# Forward sampling from a BN

$P(D, I, G, S, L)$




| $d^0$ | $d^1$ |
|---|---|
| 0.6 | 0.4 |

| $i^0$ | $i^1$ |
|---|---|
| 0.7 | 0.3 |

| | $g^1$ | $g^2$ | $g^3$ |
|---|---|---|---|
| $i^0, d^0$ | 0.3 | 0.4 | 0.3 |
| $i^0, d^1$ | 0.05 | 0.25 | 0.7 |
| $i^1, d^0$ | 0.9 | 0.08 | 0.02 |
| $i^1, d^1$ | 0.5 | 0.3 | 0.2 |

| | $s^0$ | $s^1$ |
|---|---|---|
| $i^0$ | 0.95 | 0.05 |
| $i^1$ | 0.2 | 0.8 |

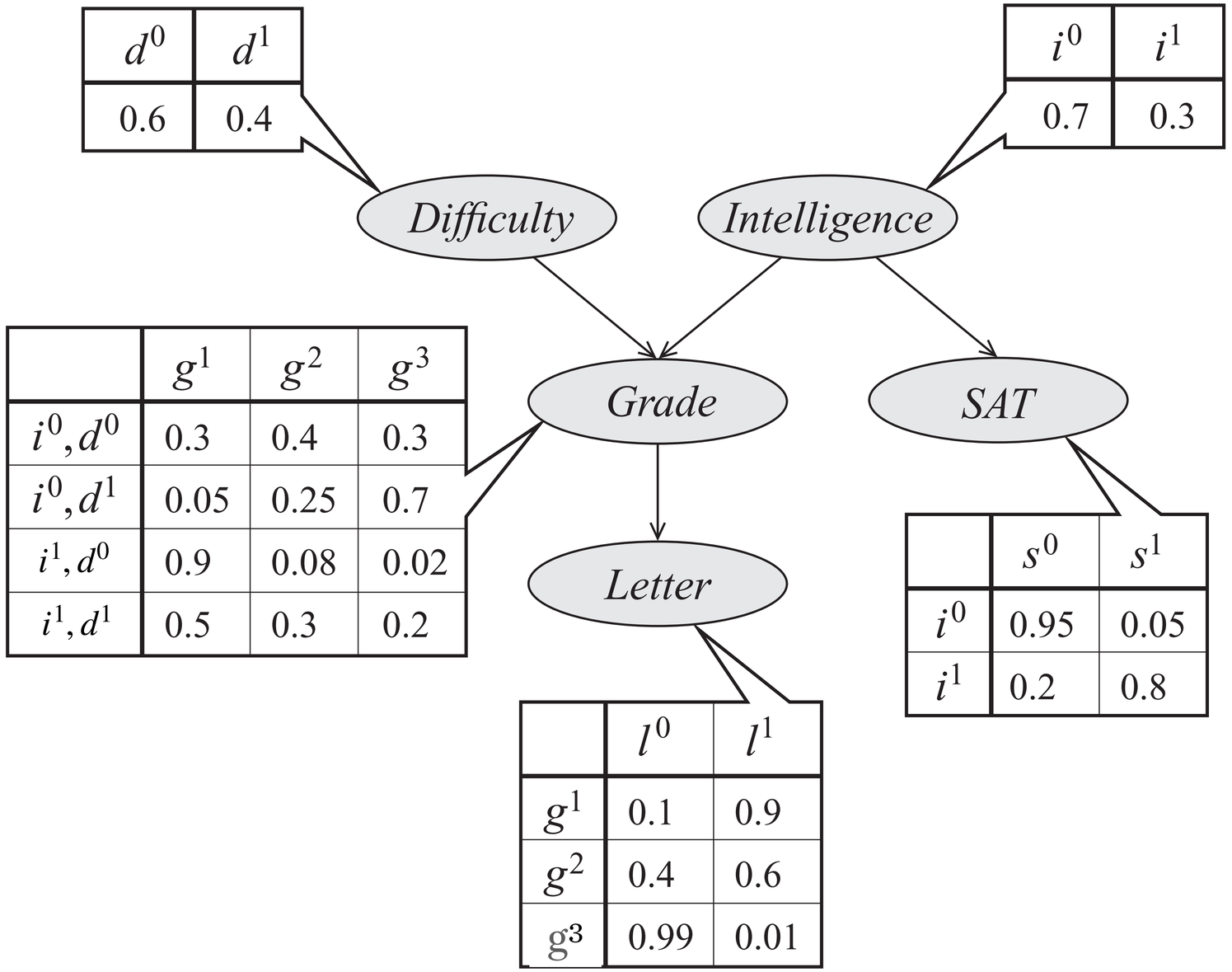
| | $l^0$ | $l^1$ |
|---|---|---|
| $g^1$ | 0.1 | 0.9 |
| $g^2$ | 0.4 | 0.6 |
| $g^3$ | 0.99 | 0.01 |

$d^0 i^1 g^1 s^0 l^1$

$d^1 i^1 g^2 s^1 l^0$

# Forward sampling from a BN

$P(D, I, G, S, L)$




| $d^0$ | $d^1$ |
|---|---|
| 0.6 | 0.4 |

| $i^0$ | $i^1$ |
|---|---|
| 0.7 | 0.3 |

| | $g^1$ | $g^2$ | $g^3$ |
|---|---|---|---|
| $i^0, d^0$ | 0.3 | 0.4 | 0.3 |
| $i^0, d^1$ | 0.05 | 0.25 | 0.7 |
| $i^1, d^0$ | 0.9 | 0.08 | 0.02 |
| $i^1, d^1$ | 0.5 | 0.3 | 0.2 |

| | $s^0$ | $s^1$ |
|---|---|---|
| $i^0$ | 0.95 | 0.05 |
| $i^1$ | 0.2 | 0.8 |

| | $l^0$ | $l^1$ |
|---|---|---|
| $g^1$ | 0.1 | 0.9 |
| $g^2$ | 0.4 | 0.6 |
| $g^3$ | 0.99 | 0.01 |

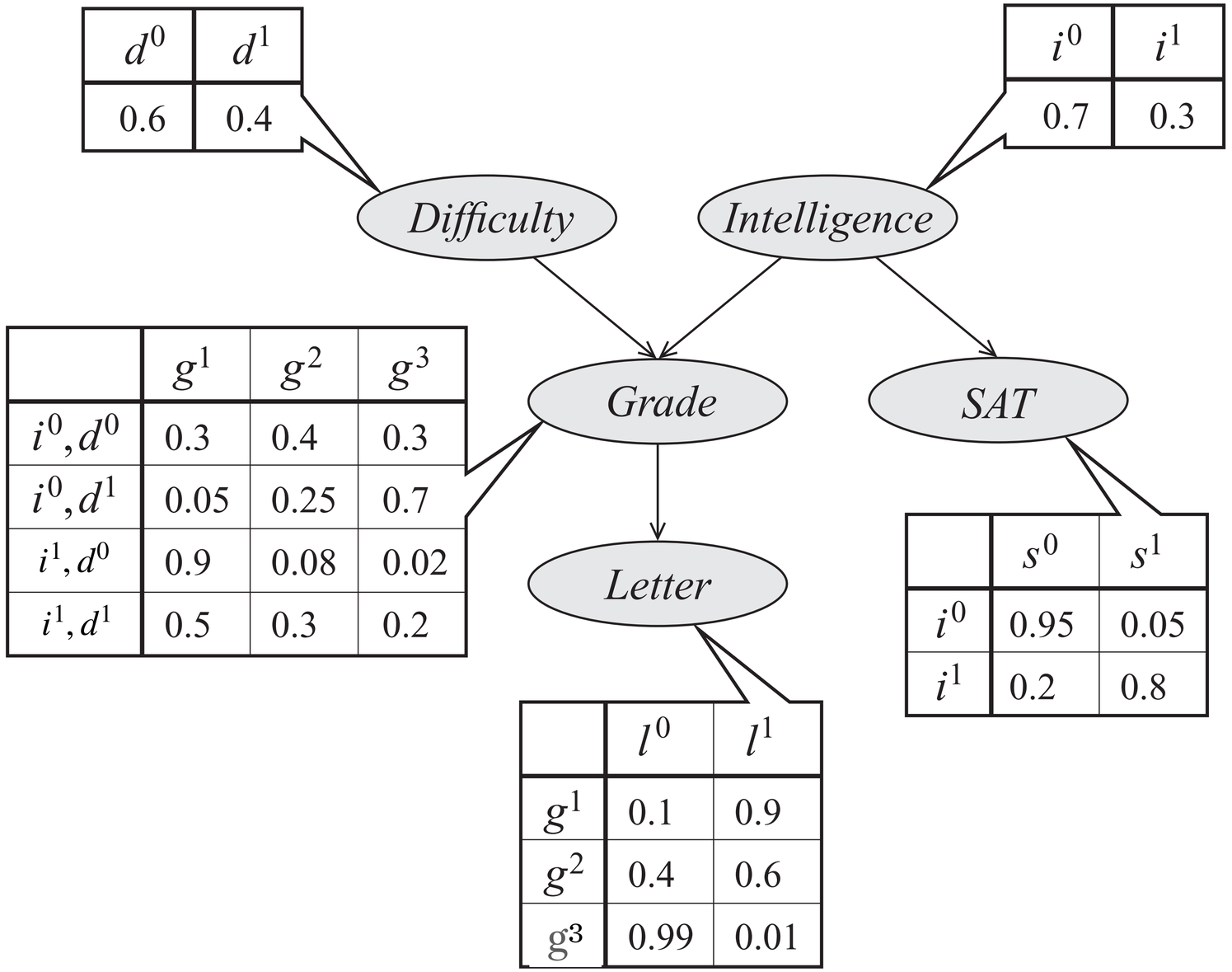
$d^0 i^1 g^1 s^0 l^1$

$d^1 i^1 g^2 s^1 l^0$

Sampling in topological order

# Forward sampling from a BN

$P(D, I, G, S, L)$




| $d^0$ | $d^1$ |
|---|---|
| 0.6 | 0.4 |

| $i^0$ | $i^1$ |
|---|---|
| 0.7 | 0.3 |

| | $g^1$ | $g^2$ | $g^3$ |
|---|---|---|---|
| $i^0, d^0$ | 0.3 | 0.4 | 0.3 |
| $i^0, d^1$ | 0.05 | 0.25 | 0.7 |
| $i^1, d^0$ | 0.9 | 0.08 | 0.02 |
| $i^1, d^1$ | 0.5 | 0.3 | 0.2 |

| | $s^0$ | $s^1$ |
|---|---|---|
| $i^0$ | 0.95 | 0.05 |
| $i^1$ | 0.2 | 0.8 |

| | $l^0$ | $l^1$ |
|---|---|---|
| $g^1$ | 0.1 | 0.9 |
| $g^2$ | 0.4 | 0.6 |
| $g^3$ | 0.99 | 0.01 |

$d^0 i^1 g^1 s^0 l^1$

$d^1 i^1 g^2 s^1 l^0$

Sampling in topological order

Parents appear before children

# Forward sampling for Querying

- Goal: estimate $P(Y = y)$
  - Generate samples from BN
  - Compute fraction where $Y = y$

# Forward sampling for Querying

- Goal: estimate $P(Y = y)$
  - Generate samples from BN
  - Compute fraction where $Y = y$

For additive bound $\epsilon$ on error with probability $> 1 - \delta$: $M \geq \frac{\ln(2/\delta)}{2\epsilon^2}$

# Forward sampling for Querying

- Goal: estimate $P(Y = y)$
  - Generate samples from BN
  - Compute fraction where $Y = y$

For additive bound $\epsilon$ on error with probability $> 1 - \delta$: $M \geq \frac{\ln(2/\delta)}{2\epsilon^2}$

For multiplicative bound $\epsilon$ on error with probability $> 1 - \delta$: $M \geq 3\frac{\ln(2/\delta)}{P(y)\epsilon^2}$

# Queries with evidence

- Goal: estimate $P(Y = y \mid \boldsymbol{E} = \boldsymbol{e})$
- Rejection sampling algorithm
  - Generate samples from BN
  - Throw away all those where $\boldsymbol{E} \neq \boldsymbol{e}$
  - Compute fraction where $Y = y$

# Queries with evidence

- Goal: estimate $P(Y = y \mid \boldsymbol{E} = \boldsymbol{e})$
- Rejection sampling algorithm
  - Generate samples from BN
  - Throw away all those where $\boldsymbol{E} \neq \boldsymbol{e}$
  - Compute fraction where $Y = y$

**Remaining samples are sampled from** $P(\cdot \mid E = e)$

# Queries with evidence

- Goal: estimate $P(Y = y \mid \boldsymbol{E} = \boldsymbol{e})$
- Rejection sampling algorithm
  - Generate samples from BN
  - Throw away all those where $\boldsymbol{E} \neq \boldsymbol{e}$
  - Compute fraction where $Y = y$
- Expected fraction of samples kept $\sim P(e)$

**Remaining samples are sampled from** $P(\cdot \mid E = e)$

# Queries with evidence

- Goal: estimate $P(Y = y \,|\, \boldsymbol{E} = \boldsymbol{e})$
- Rejection sampling algorithm
  - Generate samples from BN
  - Throw away all those where $\boldsymbol{E} \neq \boldsymbol{e}$
  - Compute fraction where $Y = y$
- Expected fraction of samples kept $\sim P(e)$

# samples needed grows exponentially
with # of observed variables

**Remaining samples are sampled from** $P(\,\cdot\,|\,E = e)$

# Summary simple sampling

- Generating samples from a BN is easy

# Summary simple sampling

- Generating samples from a BN is easy
- $(\epsilon, \delta)$-bounds exist but usefulness is limited:
  - Additive bounds: useless for low-probability events
  - Multiplicative bounds: # samples grows as $1/P(y)$

# Summary simple sampling

- Generating samples from a BN is easy
- $(\epsilon, \delta)$-bounds exist but usefulness is limited:
  - Additive bounds: useless for low-probability events
  - Multiplicative bounds: # samples grows as $1/P(y)$
- With evidence, # required samples grows exponentially with # of observed variables

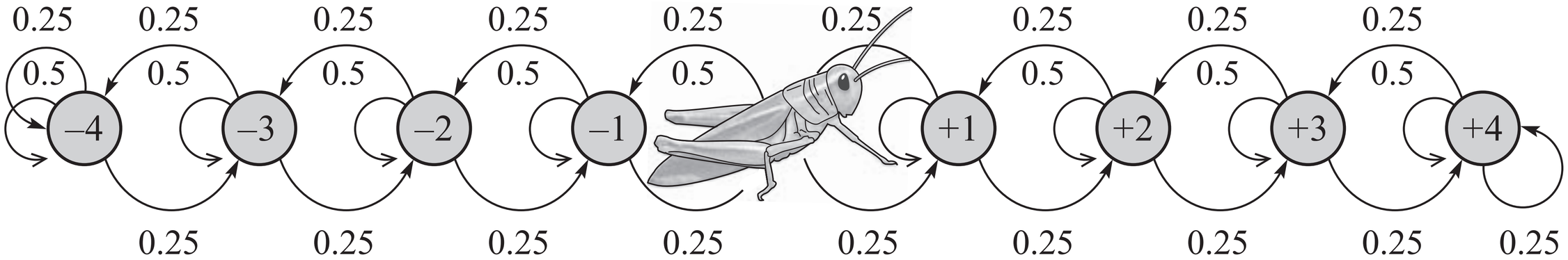
# Summary simple sampling

- Generating samples from a BN is easy
- $(\epsilon, \delta)$-bounds exist but usefulness is limited:
  - Additive bounds: useless for low-probability events
  - Multiplicative bounds: # samples grows as $1/P(y)$
- With evidence, # required samples grows exponentially with # of observed variables
- Forward sampling generally infeasible for Markov Networks


# Markov Chain Monte Carlo

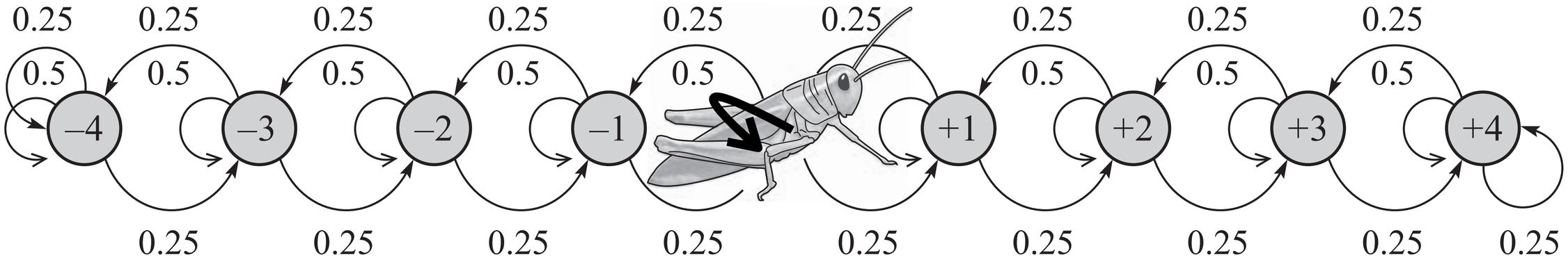
Photo by Matthew Hartley from Helmshore, Lancashire, United Kingdom - Casino de Monte-Carlo, CC BY-SA 2.0, wikimedia

# Markov Chain



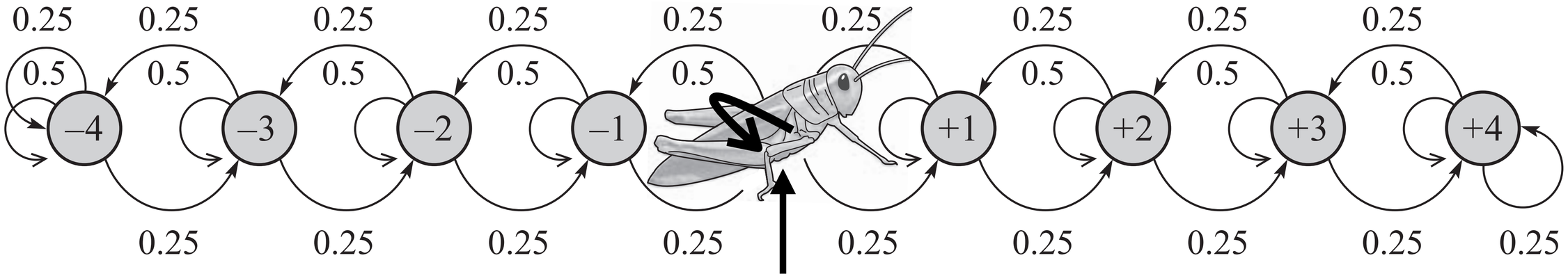

- A Markov chain defines a probabilistic transition model $T(x \rightarrow x')$ over states $x$:
  - For all $x$: $\sum_{x'} T(x \rightarrow x') = 1$

# Markov Chain




- A Markov chain defines a probabilistic transition model $T(x \rightarrow x')$ over states $x$:
  - For all $x$: $\sum_{x'} T(x \rightarrow x') = 1$

# Markov Chain




- A Markov chain defines a probabilistic transition model $T(x \rightarrow x')$ over states $x$:
  - For all $x$: $\sum_{x'} T(x \rightarrow x') = 1$

# Temporal dynamics

$$p^{(t+1)}(X^{(t+1)} = x') = \sum_x P^{(t)}(X^{(t)} = x)\, T(x \to x')$$

# Temporal dynamics

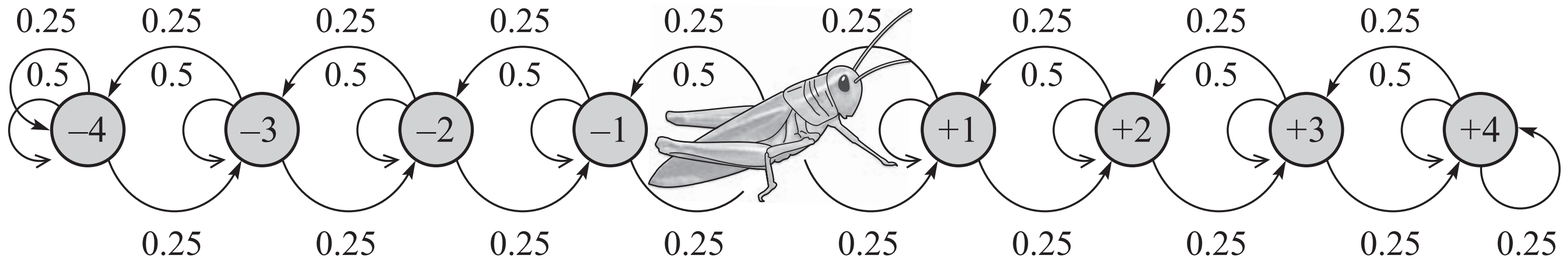
$$p^{(t+1)}(X^{(t+1)} = x') = \sum_x P^{(t)}(X^{(t)} = x)\, T(x \to x')$$

# Temporal dynamics

$$p^{(t+1)}(X^{(t+1)} = x') = \sum_{x} \underbrace{P^{(t)}(X^{(t)} = x)\, T(x \to x')}_{\text{Pairs } x, x'}$$

# Temporal dynamics

$$p^{(t+1)}(X^{(t+1)} = x') = \sum_{x} \underbrace{P^{(t)}(X^{(t)} = x)\, T(x \to x')}_{\text{Pairs } x, x'}$$




| | -2 | -1 | 0 | +1 | +2 |
|---|---|---|---|---|---|
| $\mathbf{p^{(0)}}$ | 0 | 0 | 1 | 0 | 0 |
| $\mathbf{p^{(1)}}$ | 0 | 0.25 | 0.5 | 0.25 | 0 |
| $\mathbf{p^{(2)}}$ | $0.25^2 = 0.0625$ | $2 \times (0.5 \times 0.25) = 0.25$ | $0.5^2 + 2 \times 0.25^2 = 0.375$ | $2 \times (0.5 \times 0.25) = 0.25$ | $0.25^2 = 0.0625$ |

# Stationary distribution

$$P^{(t)}(x') \approx P^{(t+1)}(x') = \sum_{x} P^{(t)}(x)\, T(x \to x')$$

# Stationary distribution

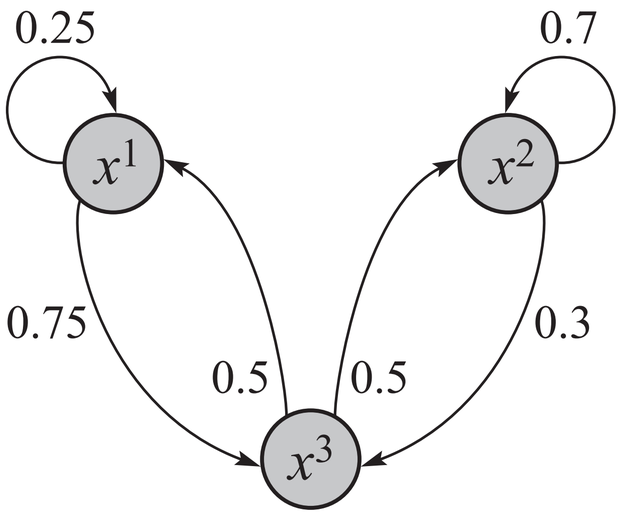
$$P^{(t)}(x') \approx P^{(t+1)}(x') = \sum_x P^{(t)}(x)\, T(x \to x')$$

$$\pi(x') = \sum_x \pi(x)\, T(x \to x')$$

# Stationary distribution

$$P^{(t)}(x') \approx P^{(t+1)}(x') = \sum_x P^{(t)}(x)\, T(x \to x')$$

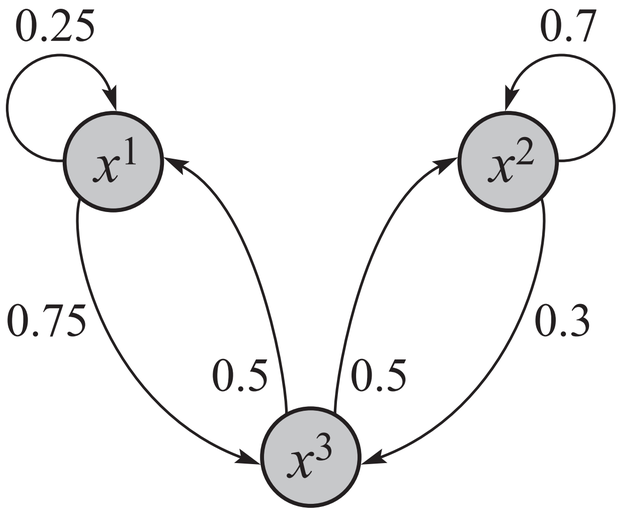
$$\pi(x') = \sum_x \pi(x)\, T(x \to x')$$



# Stationary distribution

$$P^{(t)}(x') \approx P^{(t+1)}(x') = \sum_x P^{(t)}(x)\, T(x \to x')$$

$$\pi(x') = \sum_x \pi(x)\, T(x \to x')$$

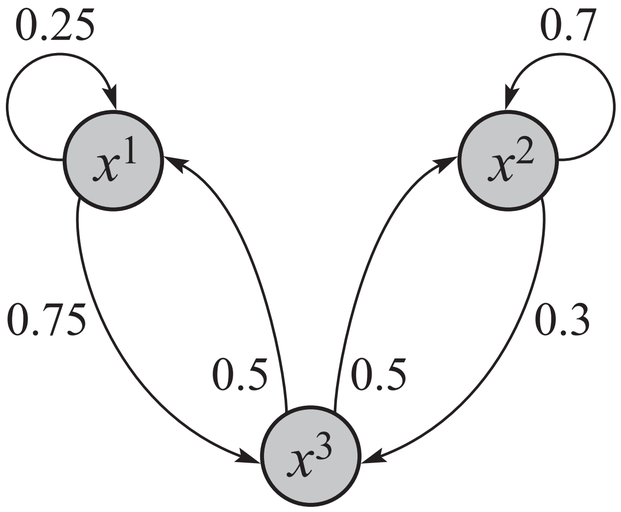



$$\pi(x^1) = 0.25\pi(x^1) + 0.5\pi(x^3)$$
$$\pi(x^2) = 0.7\pi(x^2) + 0.5\pi(x^3)$$
$$\pi(x^3) = 0.75\pi(x^1) + 0.3\pi(x^2)$$

# Stationary distribution

$$P^{(t)}(x') \approx P^{(t+1)}(x') = \sum_x P^{(t)}(x)\, T(x \to x')$$

$$\pi(x') = \sum_x \pi(x)\, T(x \to x')$$



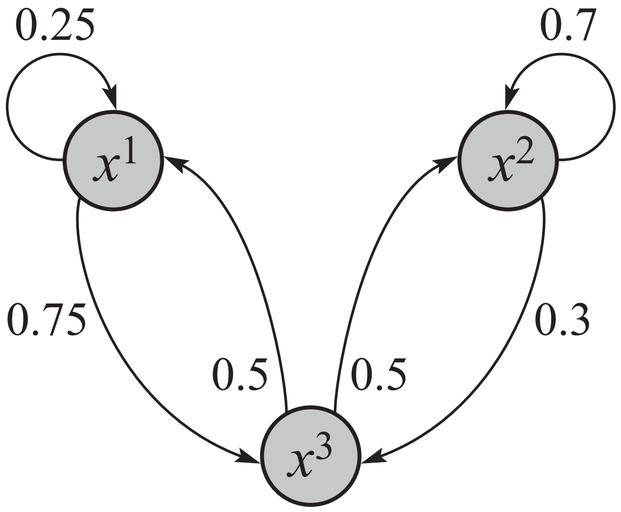

$$\pi(x^1) = 0.25\pi(x^1) + 0.5\pi(x^3)$$
$$\pi(x^2) = 0.7\pi(x^2) + 0.5\pi(x^3)$$
$$\pi(x^3) = 0.75\pi(x^1) + 0.3\pi(x^2)$$

$$\pi(x^1) + \pi(x^2) + \pi(x^3) = 1$$

# Stationary distribution

$$P^{(t)}(x') \approx P^{(t+1)}(x') = \sum_x P^{(t)}(x)\, T(x \to x')$$

$$\pi(x') = \sum_x \pi(x)\, T(x \to x')$$




$$\pi(x^1) = 0.25\pi(x^1) + 0.5\pi(x^3)$$
$$\pi(x^2) = 0.7\pi(x^2) + 0.5\pi(x^3)$$
$$\pi(x^3) = 0.75\pi(x^1) + 0.3\pi(x^2)$$

$$\pi(x^1) + \pi(x^2) + \pi(x^3) = 1$$

$$\pi(x^1) = 0.2$$
$$\pi(x^2) = 0.5$$
$$\pi(x^3) = 0.3$$

# Regular Markov chains

- A Markov chain is *regular* if there exists a $k$ such that, for every $x, x'$, the probability of getting from $x$ to $x'$ in exactly $k$ steps is $> 0$

- Theorem: A regular Markov chain converges to a unique stationary distribution regardless of start state

# Regular Markov chains

- A Markov chain is regular if there exists a $k$ such that, for every $x, x'$, the probability of getting from $x$ to $x'$ in exactly $k$ steps is $> 0$

- Sufficient conditions for regularity:
  - Every two states $x, x'$ are connected (with a path of probability $> 0$)
  - For every state, there is a self-transition

# Regular Markov chains

- A Markov chain is regular if there exists a $k$ such that, for every $x, x'$, the probability of getting from $x$ to $x'$ in exactly $k$ steps is $> 0$

  $k$ distance between furthest $x, x'$

- Sufficient conditions for regularity:
  - Every two states $x, x'$ are connected (with a path of probability $> 0$)
  - For every state, there is a self-transition

# Summary Markov chains

- A Markov chain defines a dynamical system from which we can sample trajectories
- Under certain conditions (e.g., regularity), this process is guaranteed to converge to a stationary distribution at the limit
- This allows us to sample from a distribution indirectly, and thereby provides a mechanism for sampling from an intractable distribution

# Using a Markov chain

# Using a Markov chain

- Goal: compute $P(x \in S)$
  - but $P$ is too hard to sample from directly
- Construct a (regular) Markov chain $T$ whose unique stationary distribution is $P$
- Sample $x^{(0)}$ from some initial distribution $P^{(0)}$
- For $t = 0,1,2,\ldots$
  - Generate $x^{(t+1)}$ from $T(x^{(t)} \to x')$

# Using a Markov chain

- We only want to use samples that are sampled from a distribution close to $P$
- At early iterations, $P^{(t)}$ is usually far from $P$
- Start collecting samples only after the chain has long run enough to “mix”

# Using a Markov chain

- We only want to use samples that are sampled from a distribution close to $P$
- At early iterations, $P^{(t)}$ is usually far from $P$
- Start collecting samples only after the chain has long run enough to "mix"

**$P^{(t)}$ close enough to $\pi$**

# Mixing

- How do you know if a chain has mixed or not?
  - In general, you can never “prove” a chain has mixed
  - But in many cases, you can show that it has NOT

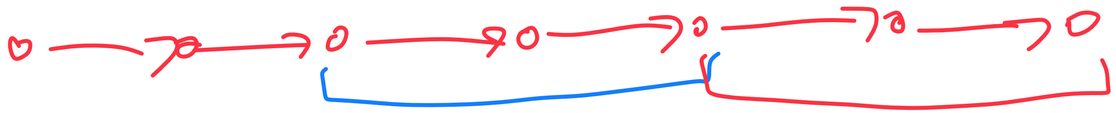
# Mixing

- How do you know if a chain has mixed or not?
  - In general, you can never “prove” a chain has mixed
  - But in many cases, you can show that it has NOT

- How do you know a chain has not mixed?
  - Compare chain statistics in different windows with a single run of the chain
  - and across different runs initialised differently

# Mixing

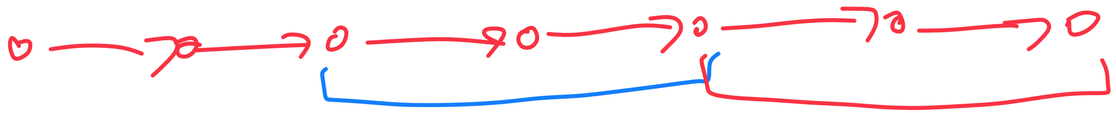
- How do you know if a chain has mixed or not?
  - In general, you can never “prove” a chain has mixed
  - But in many cases, you can show that it has NOT



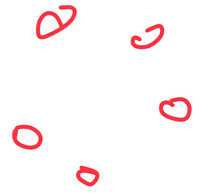
- How do you know a chain has not mixed?
  - Compare chain statistics in different windows with a single run of the chain
  - and across different runs initialised differently

# Mixing

- How do you know if a chain has mixed or not?
  - In general, you can never “prove” a chain has mixed
  - But in many cases, you can show that it has NOT

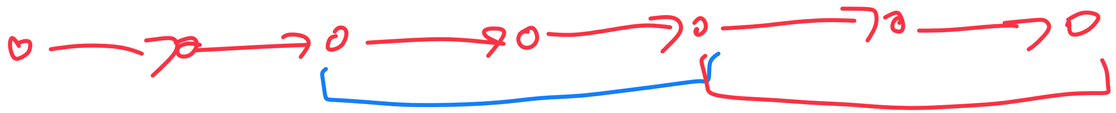


- How do you know a chain has not mixed?
  - Compare chain statistics in different windows with a single run of the chain
  - and across different runs initialised differently

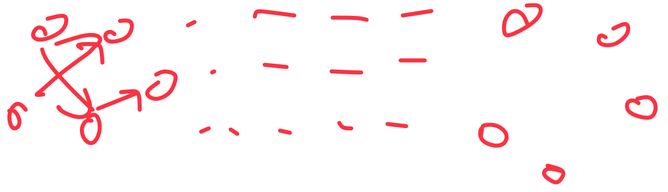


# Mixing

- How do you know if a chain has mixed or not?
  - In general, you can never “prove” a chain has mixed
  - But in many cases, you can show that it has NOT

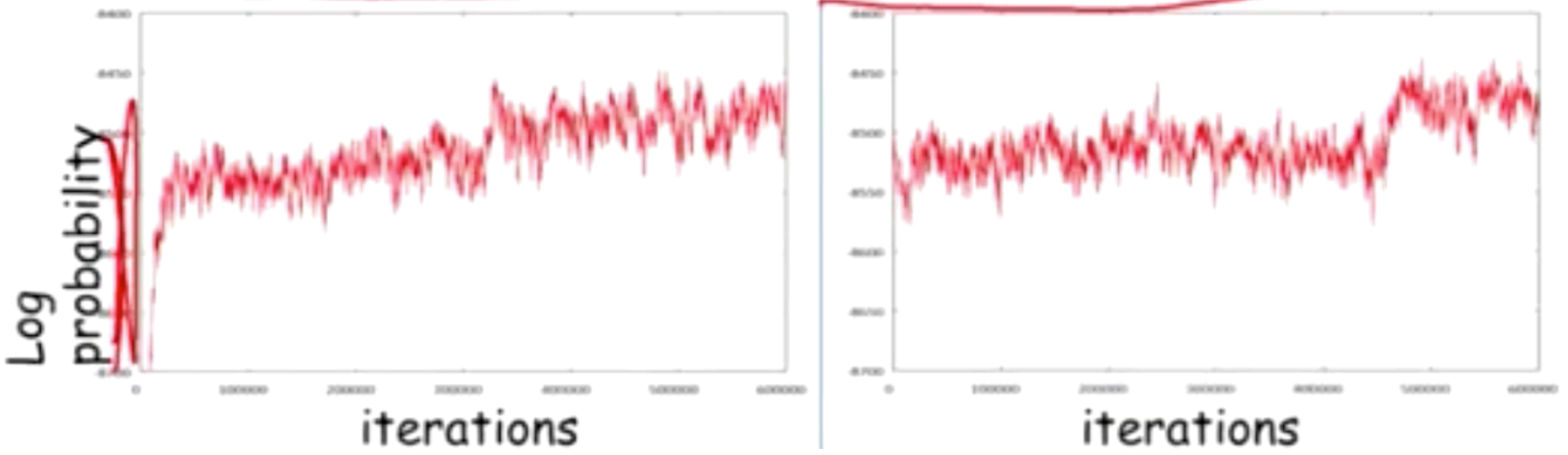


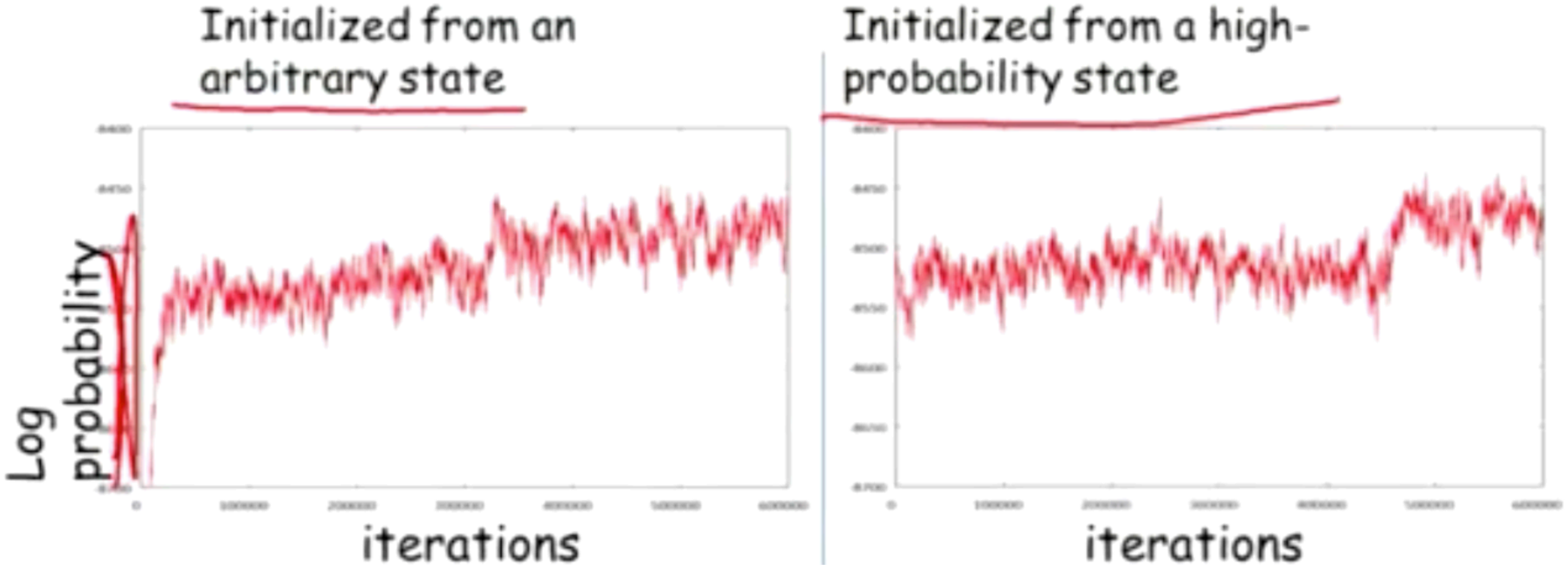
- How do you know a chain has not mixed?
  - Compare chain statistics in different windows with a single run of the chain
  - and across different runs initialised differently



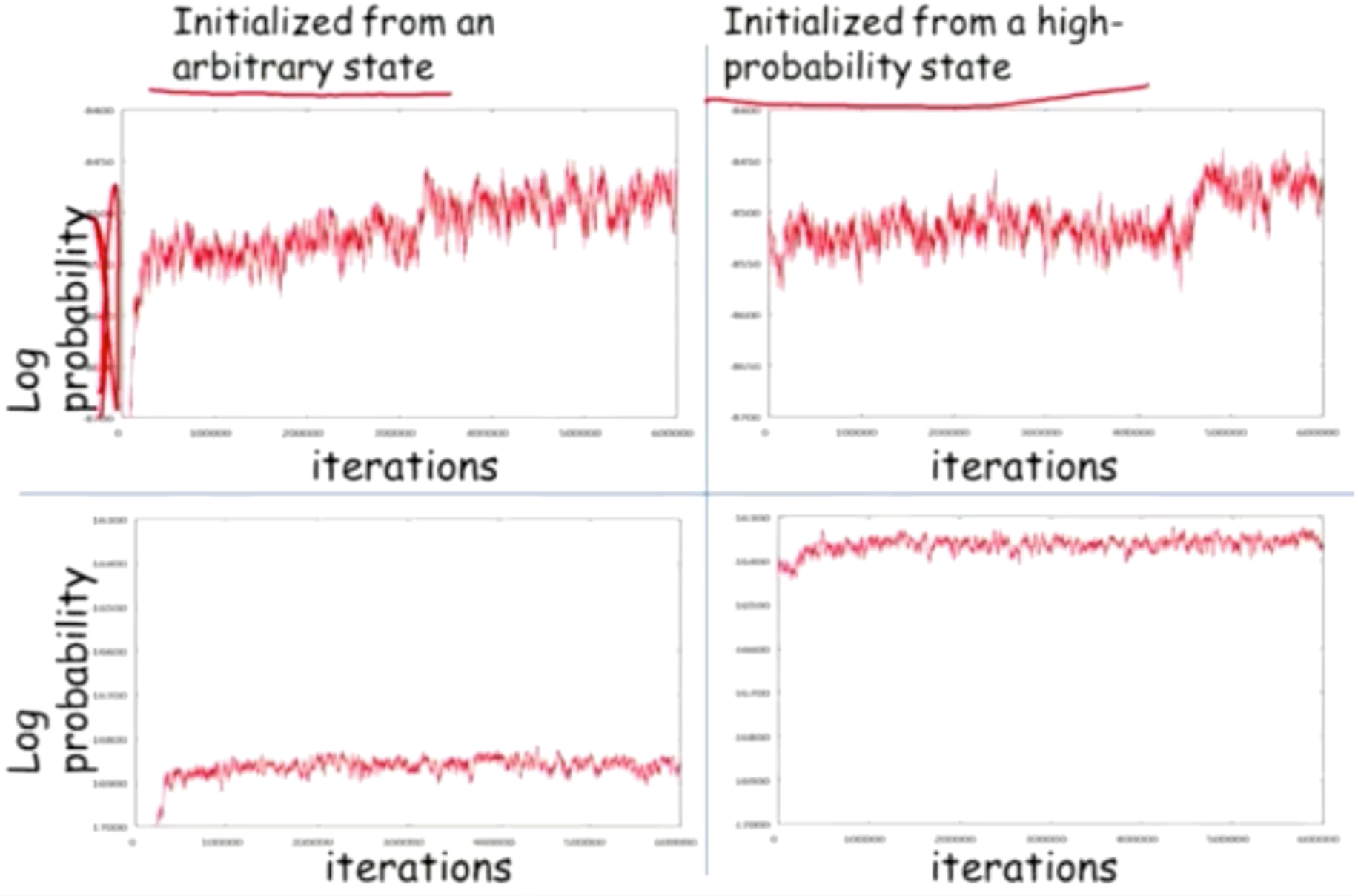
Initialized from an arbitrary state

Initialized from a high-probability state

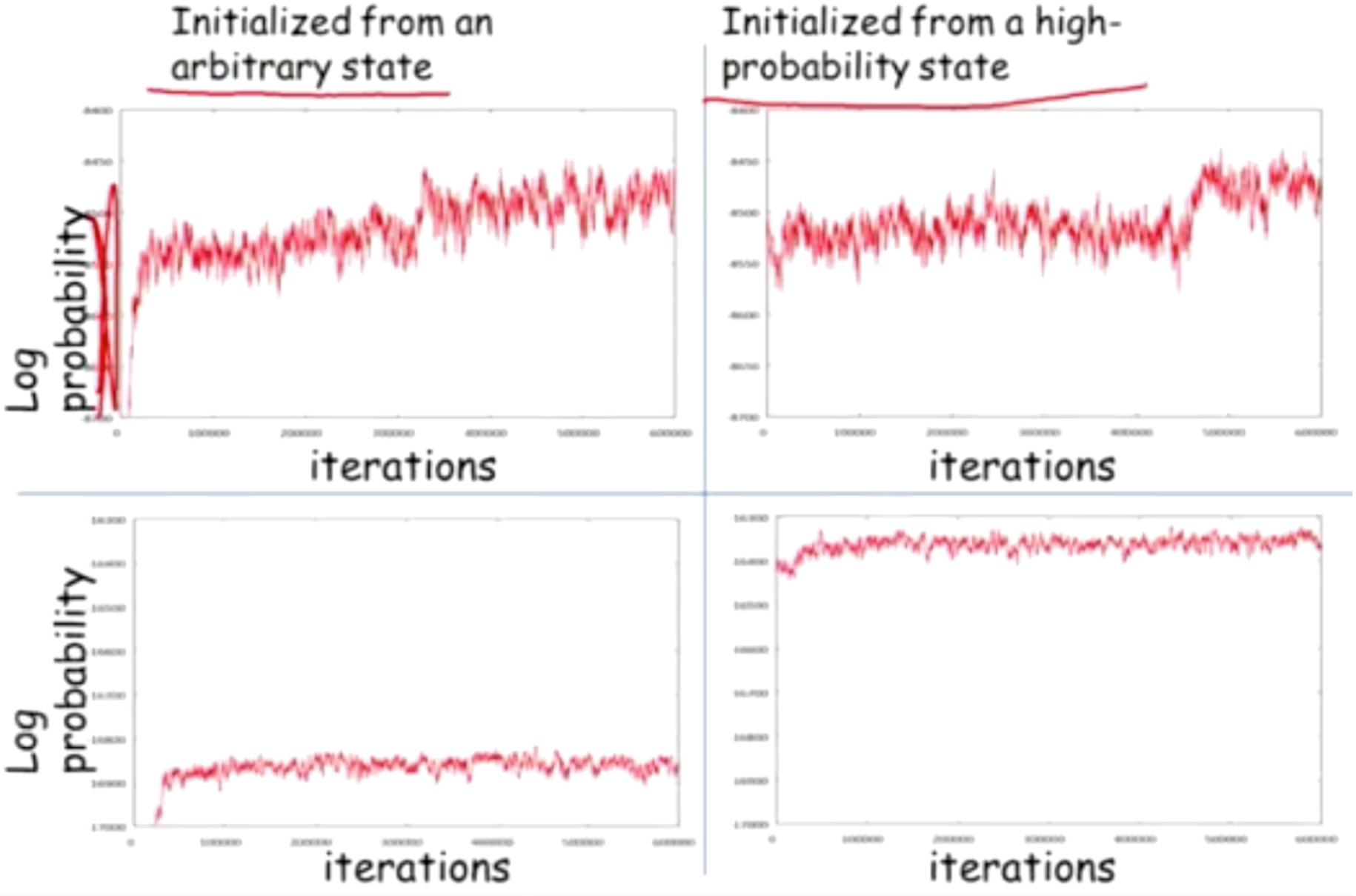





**Possibly mixed**




**Possibly mixed**




**Possibly mixed**

**Not mixed**

# Using the samples

# Using the samples

- Once the chain mixes, all samples $x^{(t)}$ are from the stationary distribution $\pi$
  - So we can (and should) use all $x^{(t)}$ for $t > T_{\text{mix}}$

# Using the samples

- Once the chain mixes, all samples $x^{(t)}$ are from the stationary distribution $\pi$
  - So we can (and should) use all $x^{(t)}$ for $t > T_{\text{mix}}$
- However, nearby samples are correlated!
  - So we shouldn't overestimate the quality of our estimate by simply counting samples (not IID)

# Using the samples

- Once the chain mixes, all samples $x^{(t)}$ are from the stationary distribution $\pi$
  - So we can (and should) use all $x^{(t)}$ for $t > T_{\text{mix}}$
- However, nearby samples are correlated!
  - So we shouldn't overestimate the quality of our estimate by simply counting samples (not IID)
- The faster the chain mixes, the less correlated (more useful) the samples

# MCMC Algorithm Summary I

- For $c = 1,...,C$

# MCMC Algorithm Summary I

- For $c = 1,...,C$
  - Sample $x^{(c,0)}$ from $P^{(0)}$

# MCMC Algorithm Summary I

- For $c = 1,\ldots,C$
  - Sample $x^{(c,0)}$ from $P^{(0)}$
- Repeat until mixing

# MCMC Algorithm Summary I

- For $c = 1,\ldots,C$
  - Sample $x^{(c,0)}$ from $P^{(0)}$
- Repeat until mixing
  - For $c = 1,\ldots,C$

# MCMC Algorithm Summary I

- For $c = 1,\ldots,C$
  - Sample $x^{(c,0)}$ from $P^{(0)}$
- Repeat until mixing
  - For $c = 1,\ldots,C$
    - Generate $x^{(c,t+1)}$ from $T(x^{(c,t)} \rightarrow x')$

# MCMC Algorithm Summary I

- For $c = 1,\ldots,C$
  - Sample $x^{(c,0)}$ from $P^{(0)}$
- Repeat until mixing
  - For $c = 1,\ldots,C$
    - Generate $x^{(c,t+1)}$ from $T(x^{(c,t)} \rightarrow x')$
  - Compare window statistics in different chains to determine mixing

# MCMC Algorithm Summary I

- For $c = 1,\ldots,C$
  - Sample $x^{(c,0)}$ from $P^{(0)}$
- Repeat until mixing
  - For $c = 1,\ldots,C$
    - Generate $x^{(c,t+1)}$ from $T(x^{(c,t)} \rightarrow x')$
  - Compare window statistics in different chains to determine mixing
  - $t := t + 1$

# MCMC Algorithm Summary II

# MCMC Algorithm Summary II

- Repeat until sufficient samples

# MCMC Algorithm Summary II

- Repeat until sufficient samples
  - $D := \emptyset$

# MCMC Algorithm Summary II

- Repeat until sufficient samples
  - $D := \emptyset$
  - For $c = 1,\ldots,C$

# MCMC Algorithm Summary II

- Repeat until sufficient samples
  - $D := \emptyset$
  - For $c = 1,\ldots,C$
    - Generate $x^{(c,t+1)}$ from $T(x^{(c,t)} \rightarrow x')$

# MCMC Algorithm Summary II

- Repeat until sufficient samples
  - $D := \emptyset$
  - For $c = 1,\ldots,C$
    - Generate $x^{(c,t+1)}$ from $T(x^{(c,t)} \rightarrow x')$
    - $D := D \cup \{x^{(c,t+1)}\}$

# MCMC Algorithm Summary II

- Repeat until sufficient samples
  - $D := \emptyset$
  - For $c = 1,\ldots,C$
    - Generate $x^{(c,t+1)}$ from $T(x^{(c,t)} \rightarrow x')$
    - $D := D \cup \{x^{(c,t+1)}\}$
  - $t := t + 1$

# MCMC Algorithm Summary II

- Repeat until sufficient samples
  - $D := \emptyset$
  - For $c = 1,\ldots,C$
    - Generate $x^{(c,t+1)}$ from $T(x^{(c,t)} \rightarrow x')$
    - $D := D \cup \{x^{(c,t+1)}\}$
  - $t := t + 1$

$$\boldsymbol{E}_P[f] \approx \frac{1}{M} \sum_{m=1}^{M} f(x[m])$$

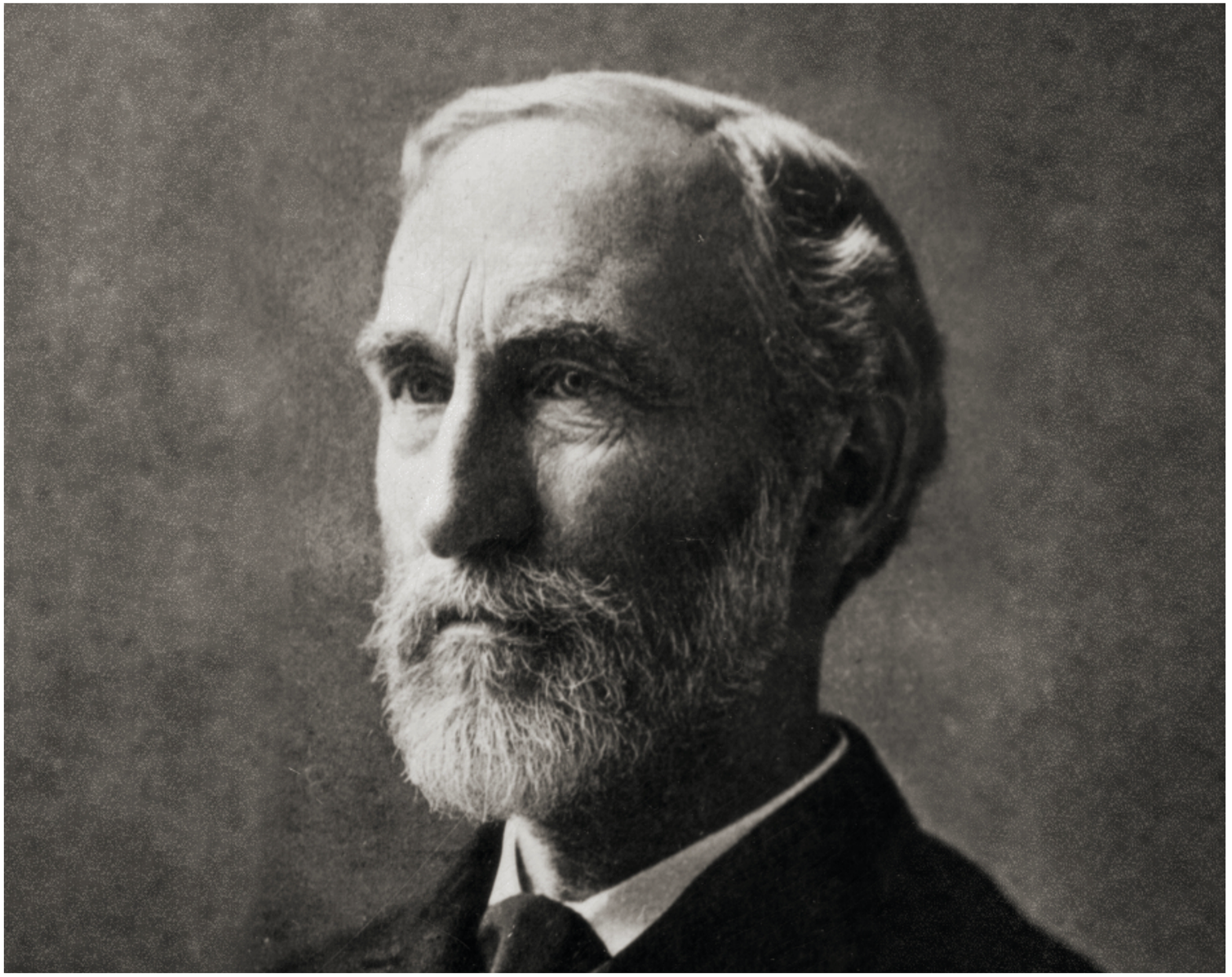
# Summary

- Pros:
  - Very general purpose
  - Often easy to implement
  - Good theoretical guarantees as $t \to \infty$

# Summary

- Pros:
  - Very general purpose
  - Often easy to implement
  - Good theoretical guarantees as $t \to \infty$
- Cons:
  - Lots of tunable parameters / design choices
  - Can be quite slow to converge
  - Difficult to tell whether it's working

# Gibbs sampling



# MCMC for PGM: The Gibbs Chain

J. W. Gibbs
(1839 — 1903)

# Gibbs chain

- Target distribution $P_\Phi(X_1, \ldots, X_n)$
- Markov chain state space: complete set of assignments $\boldsymbol{x}$ to $\boldsymbol{X} = \{X_1, \ldots, X_n\}$
- Transition model given starting state $\boldsymbol{x}$:
  - For $i = 1,..,n$
    - Sample $x_i \sim P_\Phi(X_i \mid \boldsymbol{x}_{-i})$
  - Set $\boldsymbol{x}' = \boldsymbol{x}$

# Gibbs chain

- Target distribution $P_\Phi(X_1, \ldots, X_n)$
- Markov chain state space: complete set of assignments $\boldsymbol{x}$ to $\boldsymbol{X} = \{X_1, \ldots, X_n\}$
- Transition model given starting state $\boldsymbol{x}$:
  - For $i = 1,..,n$
    - Sample $x_i \sim P_\Phi(X_i \mid \boldsymbol{x}_{-i})$
  - Set $\boldsymbol{x}' = \boldsymbol{x}$

**Assignment to all $X_1, \ldots, X_n$ except $X_i$**

# Gibbs chain

- Target distribution $P_\Phi(X_1, \ldots, X_n)$
- Markov chain state space: complete set of assignments $\boldsymbol{x}$ to $\boldsymbol{X} = \{X_1, \ldots, X_n\}$
- Transition model given starting state $\boldsymbol{x}$:
  - For $i = 1,..,n$
    - Sample $x_i \sim P_\Phi(X_i \mid \boldsymbol{x}_{-i})$
  - Set $\boldsymbol{x}' = \boldsymbol{x}$

**Assignment to all $X_1, \ldots, X_n$ except $X_i$**

$x_1$ $x_2$ $x_3$

# Gibbs chain

- Target distribution $P_\Phi(X_1, \ldots, X_n)$
- Markov chain state space: complete set of assignments $\boldsymbol{x}$ to $\boldsymbol{X} = \{X_1, \ldots, X_n\}$
- Transition model given starting state $\boldsymbol{x}$:
  - For $i = 1,..,n$
    - Sample $x_i \sim P_\Phi(X_i \mid \boldsymbol{x}_{-i})$
  - Set $\boldsymbol{x}' = \boldsymbol{x}$

**Assignment to all $X_1, \ldots, X_n$ except $X_i$**

| $x_1$ | $x_2$ | $x_3$ |
|---|---|---|
| 0 | 0 | 0 |

# Gibbs chain

- Target distribution $P_\Phi(X_1, \ldots, X_n)$
- Markov chain state space: complete set of assignments $\boldsymbol{x}$ to $\boldsymbol{X} = \{X_1, \ldots, X_n\}$
- Transition model given starting state $\boldsymbol{x}$:
  - For $i = 1,..,n$
    - Sample $x_i \sim P_\Phi(X_i \mid \boldsymbol{x}_{-i})$
  - Set $\boldsymbol{x}' = \boldsymbol{x}$

**Assignment to all $X_1, \ldots, X_n$ except $X_i$**

| $x_1$ | $x_2$ | $x_3$ | |
|---|---|---|---|
| 0 | 0 | 0 | $p(X_1 \mid X_2 = 0, X_3 = 0)$ |

# Gibbs chain

- Target distribution $P_\Phi(X_1, \ldots, X_n)$
- Markov chain state space: complete set of assignments $\boldsymbol{x}$ to $\boldsymbol{X} = \{X_1, \ldots, X_n\}$
- Transition model given starting state $\boldsymbol{x}$:
  - For $i = 1,..,n$
    - Sample $x_i \sim P_\Phi(X_i \mid \boldsymbol{x}_{-i})$
  - Set $\boldsymbol{x}' = \boldsymbol{x}$

**Assignment to all $X_1, \ldots, X_n$ except $X_i$**

| $x_1$ | $x_2$ | $x_3$ | |
|---|---|---|---|
| 0 | 0 | 0 | $p(X_1 \mid X_2 = 0, X_3 = 0)$ |
| 1 | 0 | 0 | |

# Gibbs chain

- Target distribution $P_\Phi(X_1, \ldots, X_n)$
- Markov chain state space: complete set of assignments $\boldsymbol{x}$ to $\boldsymbol{X} = \{X_1, \ldots, X_n\}$
- Transition model given starting state $\boldsymbol{x}$:
  - For $i = 1,..,n$
    - Sample $x_i \sim P_\Phi(X_i \mid \boldsymbol{x}_{-i})$
  - Set $\boldsymbol{x}' = \boldsymbol{x}$

**Assignment to all $X_1, \ldots, X_n$ except $X_i$**

| $x_1$ | $x_2$ | $x_3$ | |
|---|---|---|---|
| 0 | 0 | 0 | $p(X_1 \mid X_2 = 0, X_3 = 0)$ |
| 1 | 0 | 0 | $p(X_2 \mid X_1 = 1, X_3 = 0)$ |

# Gibbs chain

- Target distribution $P_\Phi(X_1, \ldots, X_n)$
- Markov chain state space: complete set of assignments $\boldsymbol{x}$ to $\boldsymbol{X} = \{X_1, \ldots, X_n\}$
- Transition model given starting state $\boldsymbol{x}$:
  - For $i = 1,..,n$
    - Sample $x_i \sim P_\Phi(X_i \mid \boldsymbol{x}_{-i})$
  - Set $\boldsymbol{x}' = \boldsymbol{x}$

**Assignment to all $X_1, \ldots, X_n$ except $X_i$**

| $x_1$ | $x_2$ | $x_3$ | |
|---|---|---|---|
| 0 | 0 | 0 | $p(X_1 \mid X_2 = 0, X_3 = 0)$ |
| 1 | 0 | 0 | $p(X_2 \mid X_1 = 1, X_3 = 0)$ |
| 1 | 0 | 0 | |

# Gibbs chain

- Target distribution $P_\Phi(X_1, \ldots, X_n)$
- Markov chain state space: complete set of assignments $\boldsymbol{x}$ to $\boldsymbol{X} = \{X_1, \ldots, X_n\}$
- Transition model given starting state $\boldsymbol{x}$:
  - For $i = 1,..,n$
    - Sample $x_i \sim P_\Phi(X_i \mid \boldsymbol{x}_{-i})$
  - Set $\boldsymbol{x}' = \boldsymbol{x}$

**Assignment to all $X_1, \ldots, X_n$ except $X_i$**

| $x_1$ | $x_2$ | $x_3$ | |
|---|---|---|---|
| 0 | 0 | 0 | $p(X_1 \mid X_2 = 0, X_3 = 0)$ |
| 1 | 0 | 0 | $p(X_2 \mid X_1 = 1, X_3 = 0)$ |
| 1 | 0 | 0 | $p(X_3 \mid X_1 = 1, X_2 = 0)$ |

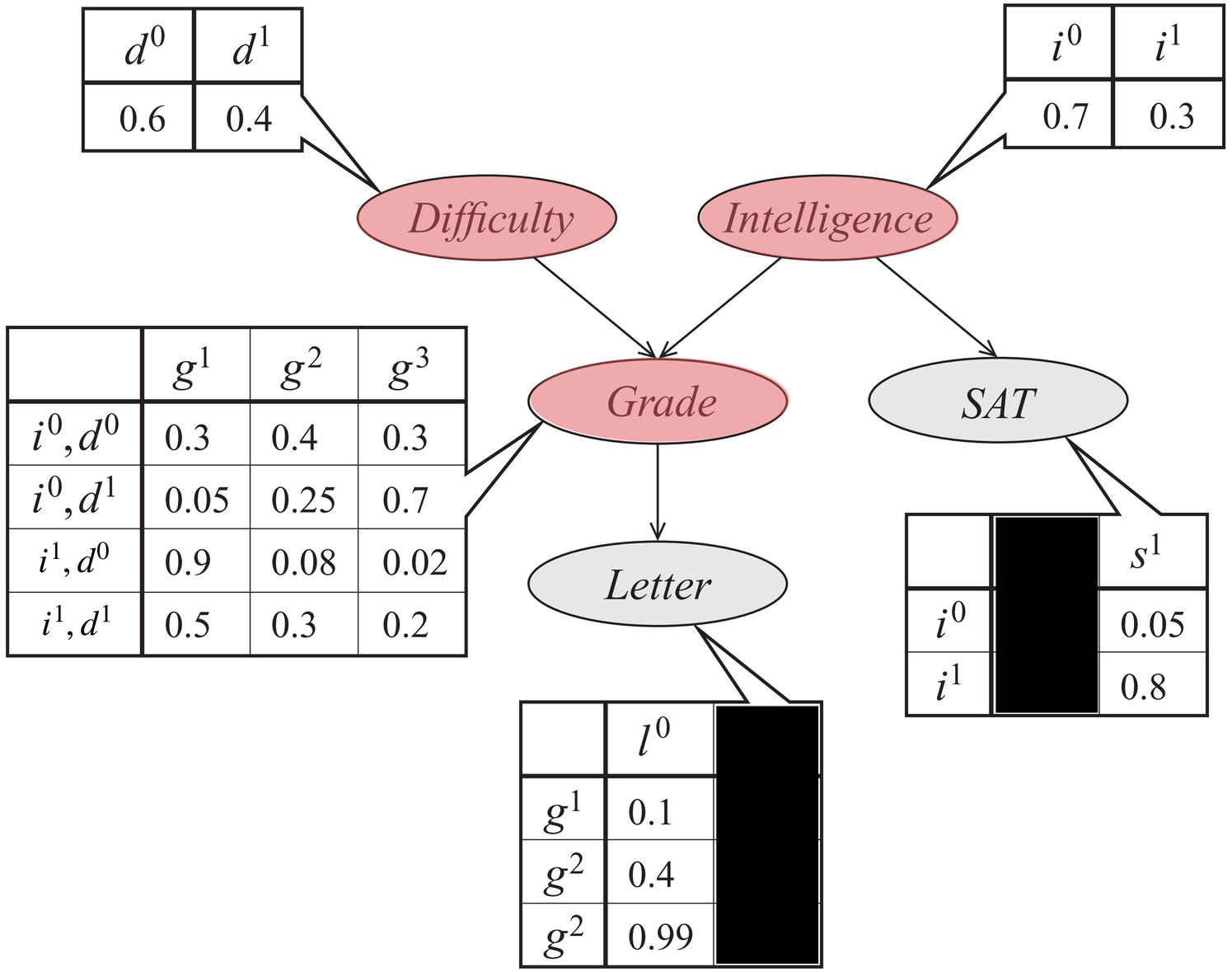
# Gibbs chain

- Target distribution $P_\Phi(X_1, \ldots, X_n)$
- Markov chain state space: complete set of assignments $\boldsymbol{x}$ to $\boldsymbol{X} = \{X_1, \ldots, X_n\}$
- Transition model given starting state $\boldsymbol{x}$:
  - For $i = 1,..,n$
    - Sample $x_i \sim P_\Phi(X_i \mid \boldsymbol{x}_{-i})$
  - Set $\boldsymbol{x}' = \boldsymbol{x}$

**Assignment to all $X_1, \ldots, X_n$ except $X_i$**

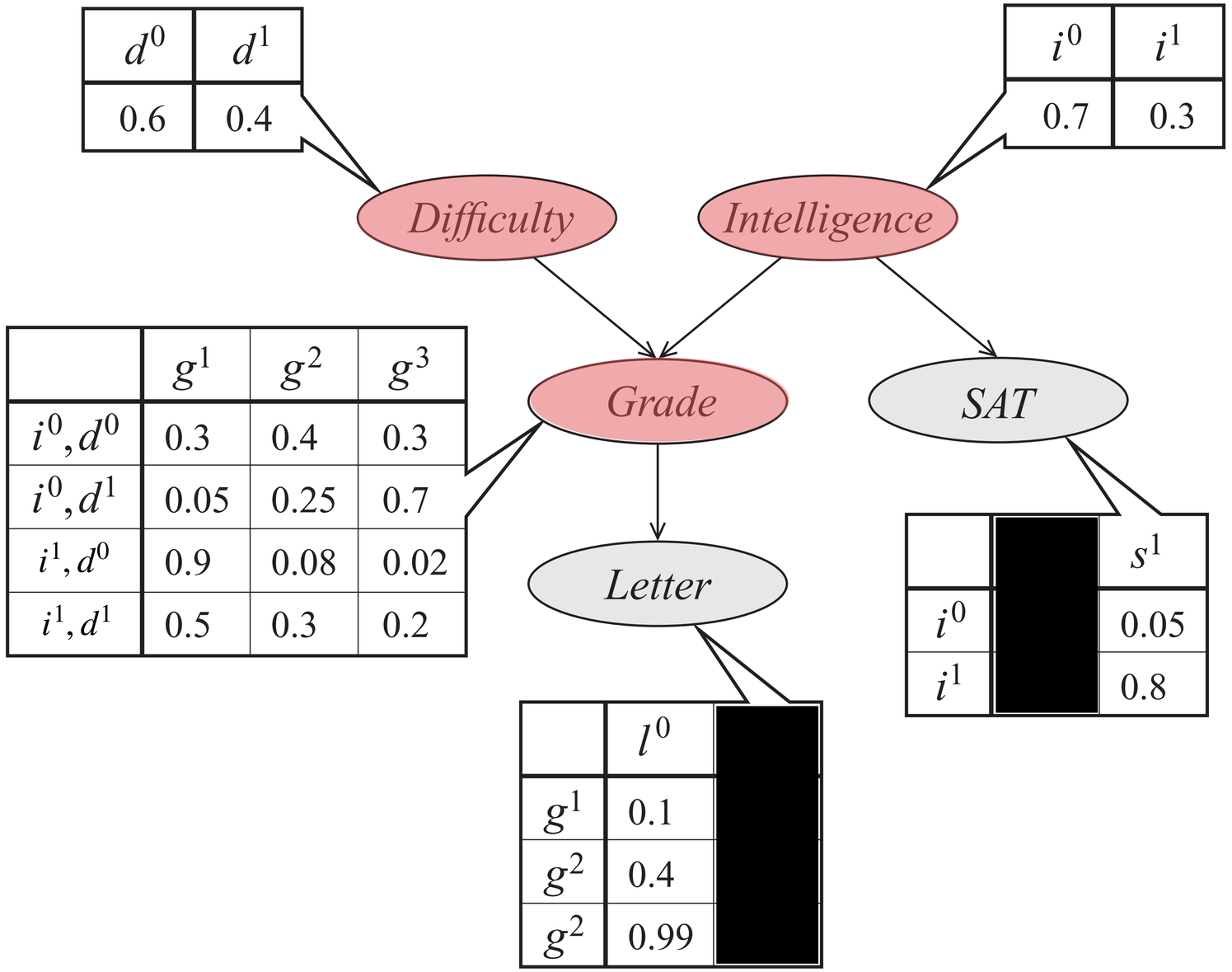
| $x_1$ | $x_2$ | $x_3$ | |
|---|---|---|---|
| 0 | 0 | 0 | $p(X_1 \mid X_2 = 0, X_3 = 0)$ |
| 1 | 0 | 0 | $p(X_2 \mid X_1 = 1, X_3 = 0)$ |
| 1 | 0 | 0 | $p(X_3 \mid X_1 = 1, X_2 = 0)$ |
| 1 | 0 | 1 | |

# Example


$d^0$ $d^1$
0.6 0.4
$i^0$ $i^1$
0.7 0.3
Difficulty
Intelligence
Grade
SAT
Letter
$g^1$ $g^2$ $g^3$
$i^0, d^0$ 0.3 0.4 0.3
$i^0, d^1$ 0.05 0.25 0.7
$i^1, d^0$ 0.9 0.08 0.02
$i^1, d^1$ 0.5 0.3 0.2
$s^1$
$i^0$ 0.05
$i^1$ 0.8
$l^0$
$g^1$ 0.1
$g^2$ 0.4
$g^2$ 0.99

# Example

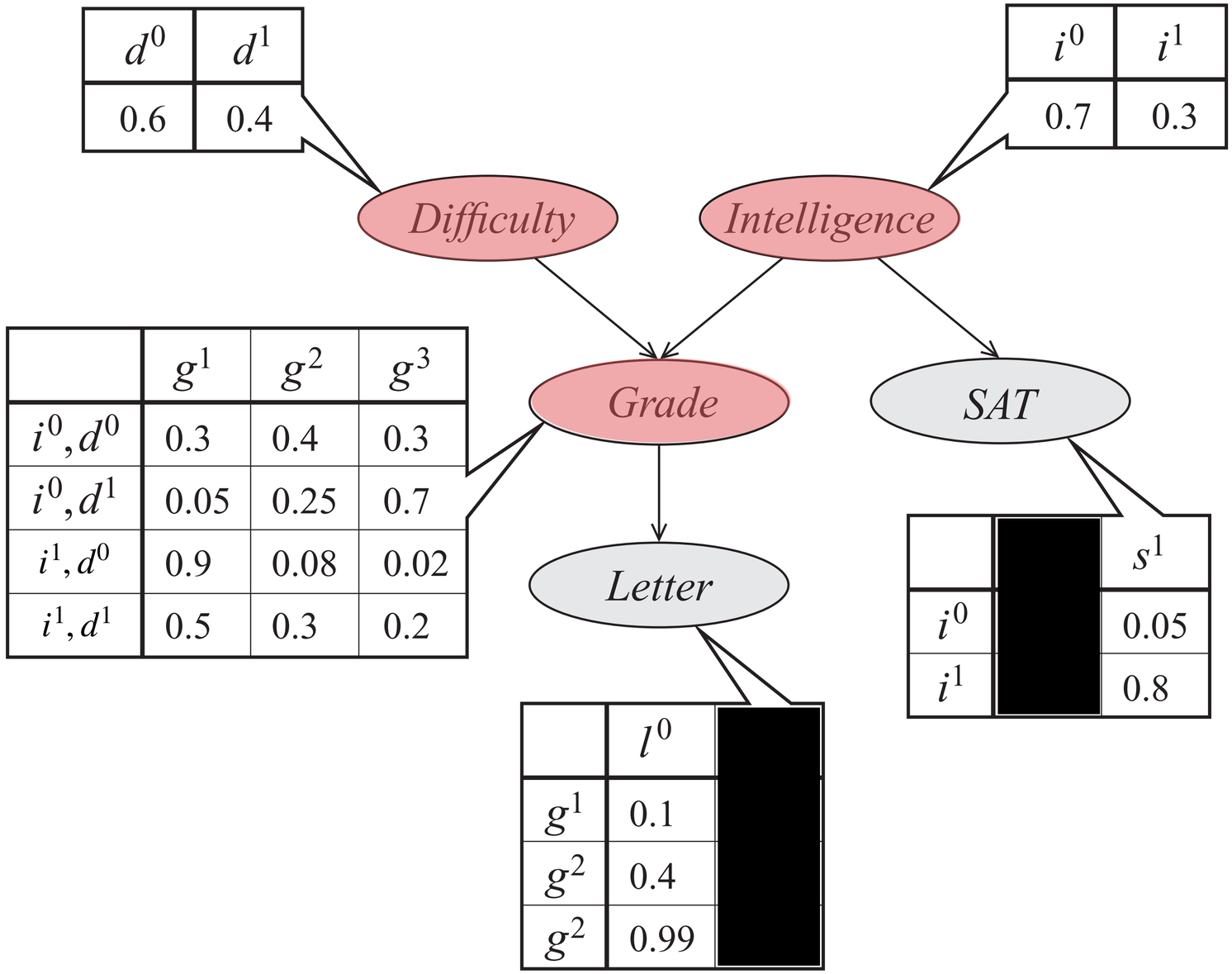
$\tilde{P}_{\Phi}(D, I, G \mid s^1, l^0)$



# Example

$\tilde{P}_\Phi(D, I, G \mid s^1, l^0)$

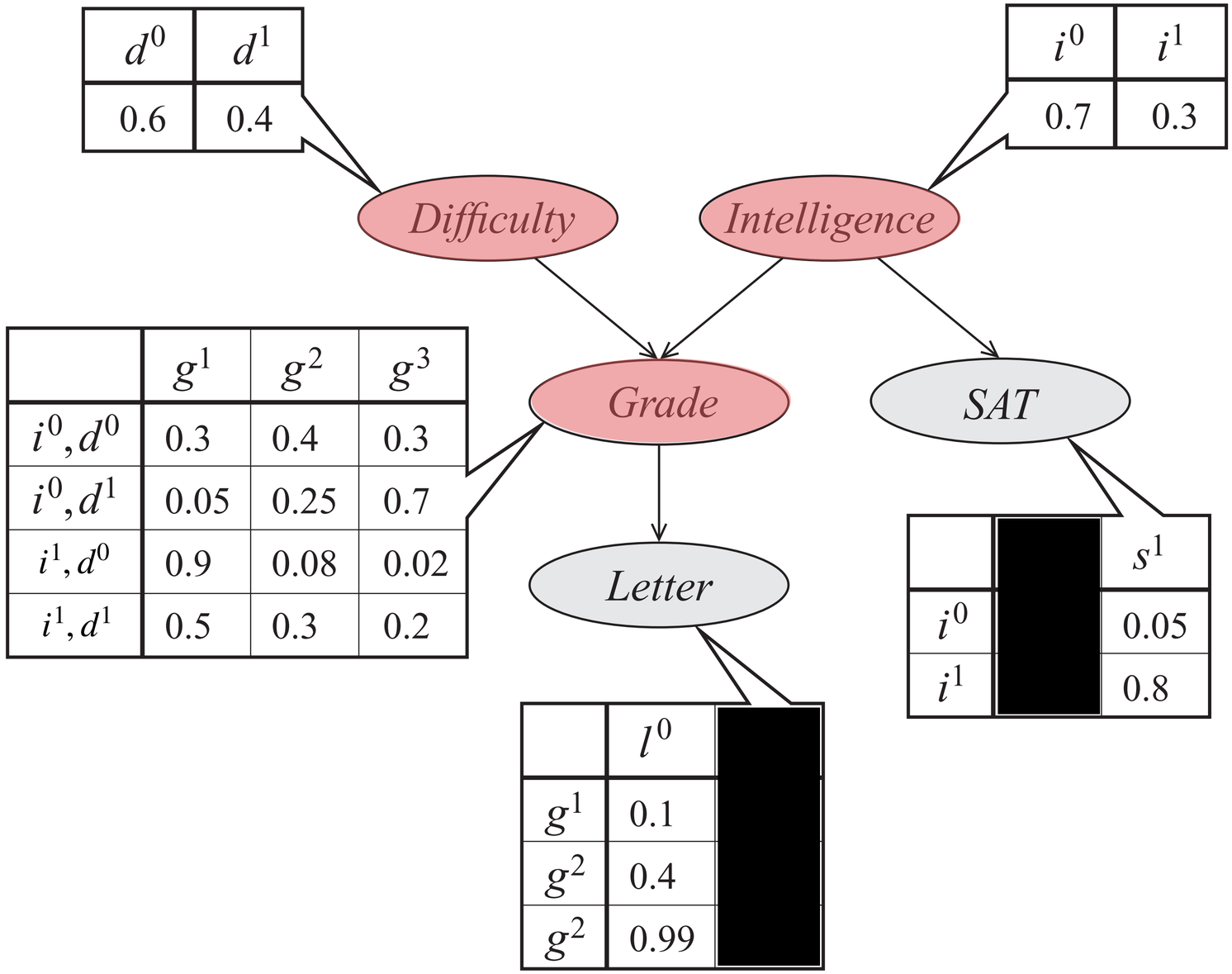
$d^0\ i^0\ g^1$



# Example

$\tilde{P}_{\Phi}(D, I, G \mid s^1, l^0)$

$d^0\ i^0\ g^1$




| $d^0$ | $d^1$ |
|---|---|
| 0.6 | 0.4 |

| $i^0$ | $i^1$ |
|---|---|
| 0.7 | 0.3 |

| | $g^1$ | $g^2$ | $g^3$ |
|---|---|---|---|
| $i^0, d^0$ | 0.3 | 0.4 | 0.3 |
| $i^0, d^1$ | 0.05 | 0.25 | 0.7 |
| $i^1, d^0$ | 0.9 | 0.08 | 0.02 |
| $i^1, d^1$ | 0.5 | 0.3 | 0.2 |

| | | $s^1$ |
|---|---|---|
| $i^0$ | | 0.05 |
| $i^1$ | | 0.8 |

| | $l^0$ |
|---|---|
| $g^1$ | 0.1 |
| $g^2$ | 0.4 |
| $g^2$ | 0.99 |

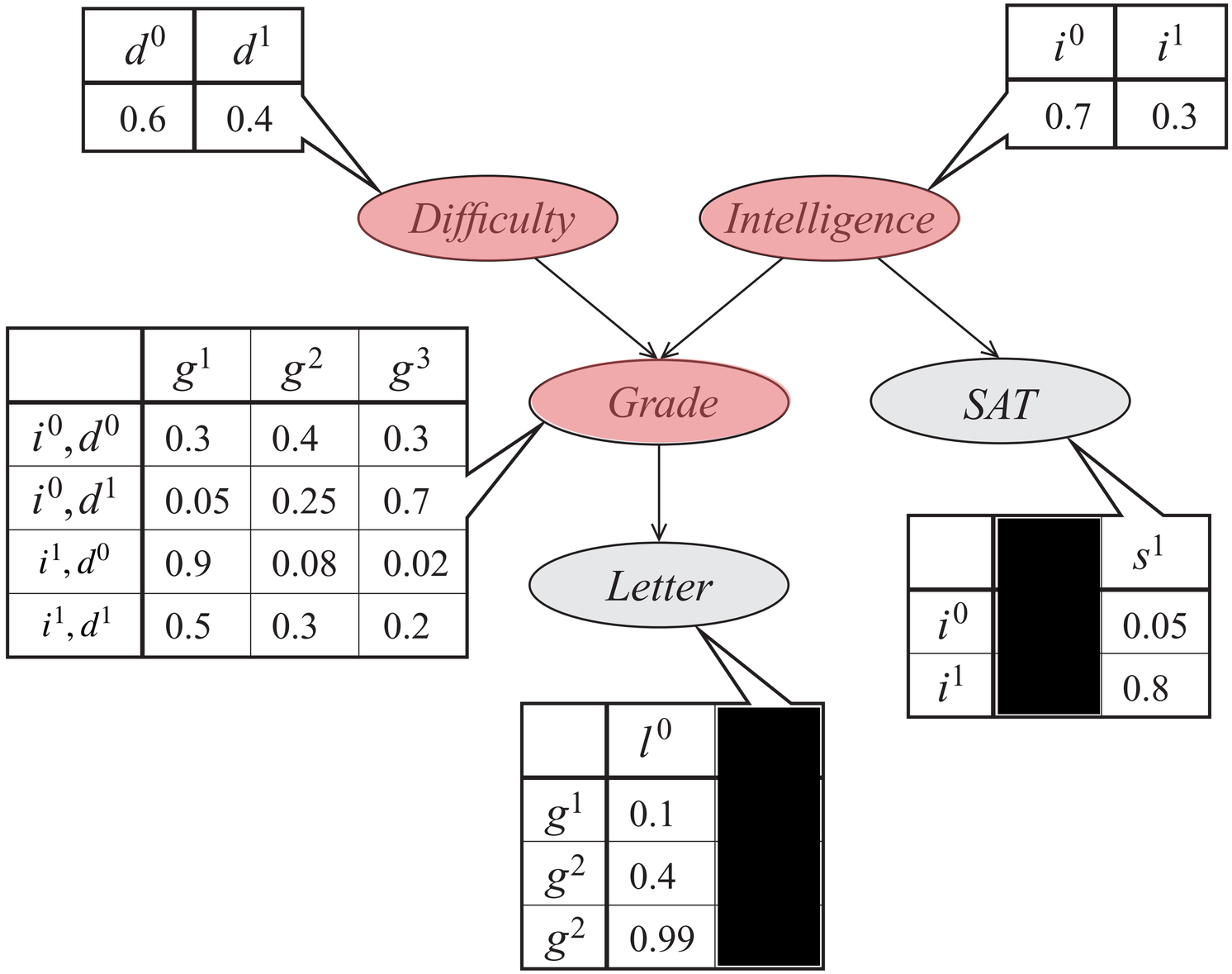
$P(D \mid i^0, g^1, l^0, s^1)$

# Example

$\tilde{P}_\Phi(D, I, G \mid s^1, l^0)$

$d^0\ i^0\ g^1$

$d^1\ i^0\ g^1$



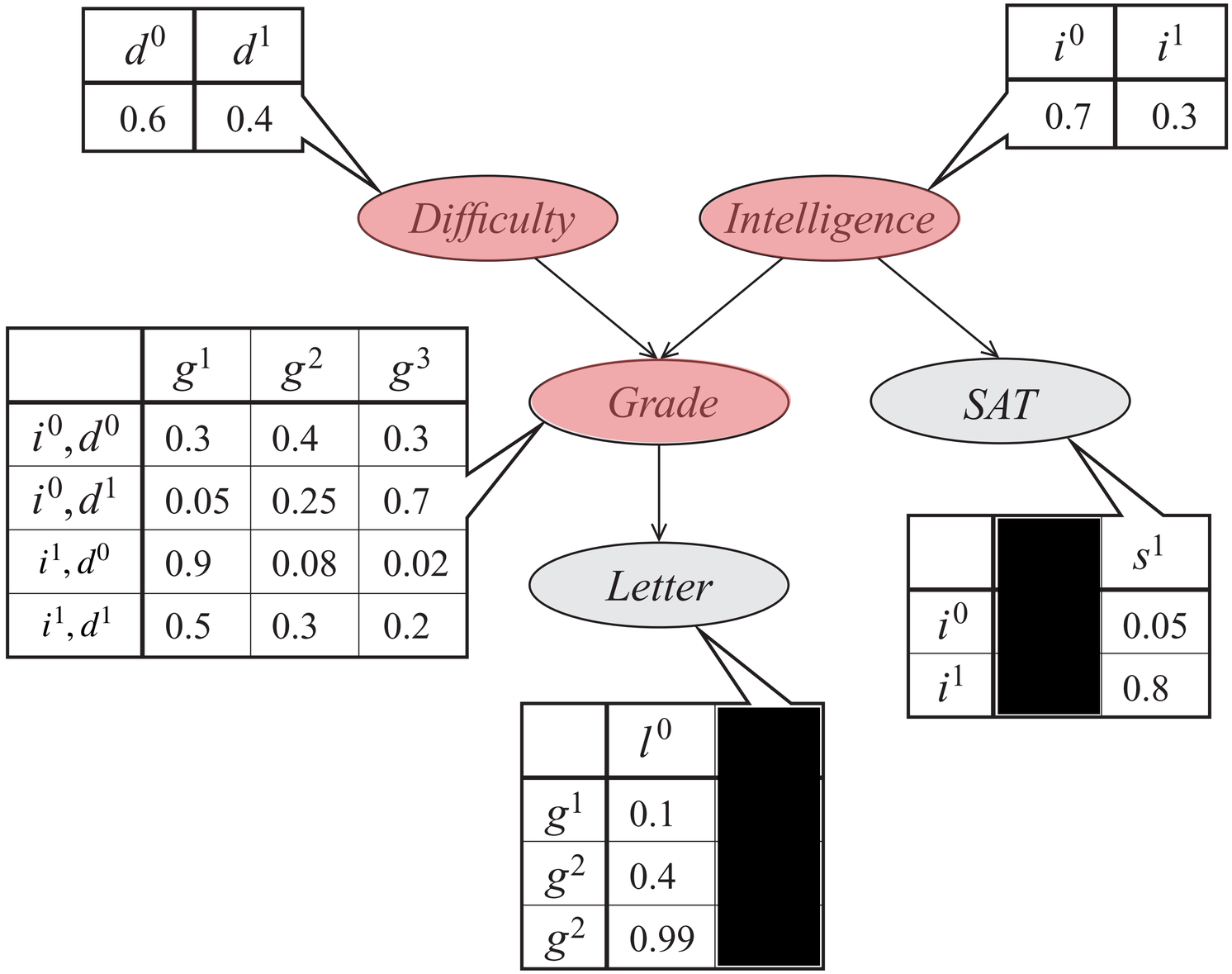

$P(D \mid i^0, g^1, l^0, s^1)$

# Example

$\tilde{P}_\Phi(D, I, G \mid s^1, l^0)$

$d^0\ i^0\ g^1$

$d^1\ i^0\ g^1$




$P(D \mid i^0, g^1, l^0, s^1)$

$P(I \mid d^1, g^1, l^0, s^1)$

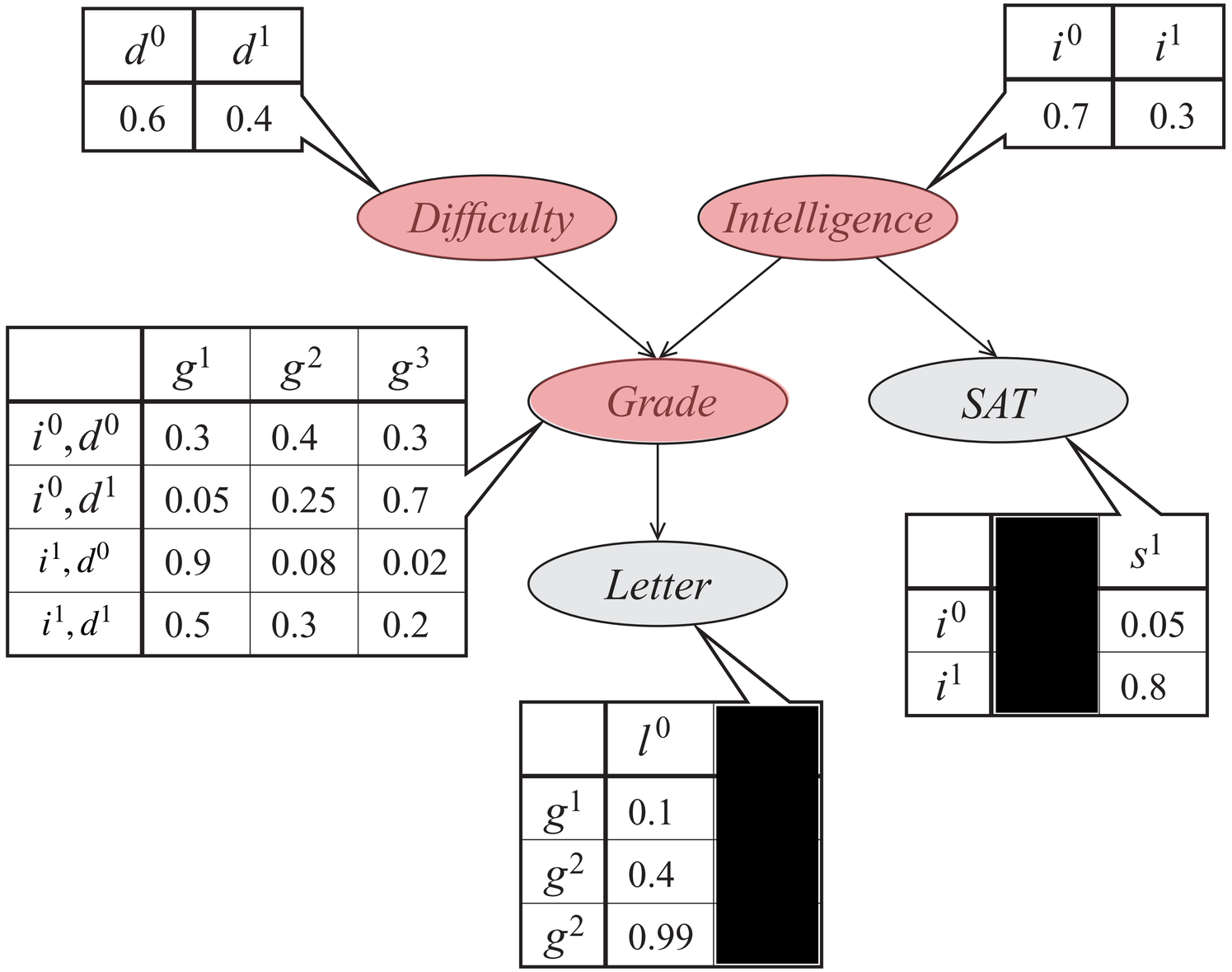
# Example

$\tilde{P}_\Phi(D, I, G \mid s^1, l^0)$

$d^0\ i^0\ g^1$

$d^1\ i^0\ g^1$

$d^1\ i^1\ g^1$




| $d^0$ | $d^1$ |
|---|---|
| 0.6 | 0.4 |

| $i^0$ | $i^1$ |
|---|---|
| 0.7 | 0.3 |

| | $g^1$ | $g^2$ | $g^3$ |
|---|---|---|---|
| $i^0, d^0$ | 0.3 | 0.4 | 0.3 |
| $i^0, d^1$ | 0.05 | 0.25 | 0.7 |
| $i^1, d^0$ | 0.9 | 0.08 | 0.02 |
| $i^1, d^1$ | 0.5 | 0.3 | 0.2 |

| | | $s^1$ |
|---|---|---|
| $i^0$ | | 0.05 |
| $i^1$ | | 0.8 |

| | $l^0$ |
|---|---|
| $g^1$ | 0.1 |
| $g^2$ | 0.4 |
| $g^2$ | 0.99 |

$P(D \mid i^0, g^1, l^0, s^1)$

$P(I \mid d^1, g^1, l^0, s^1)$

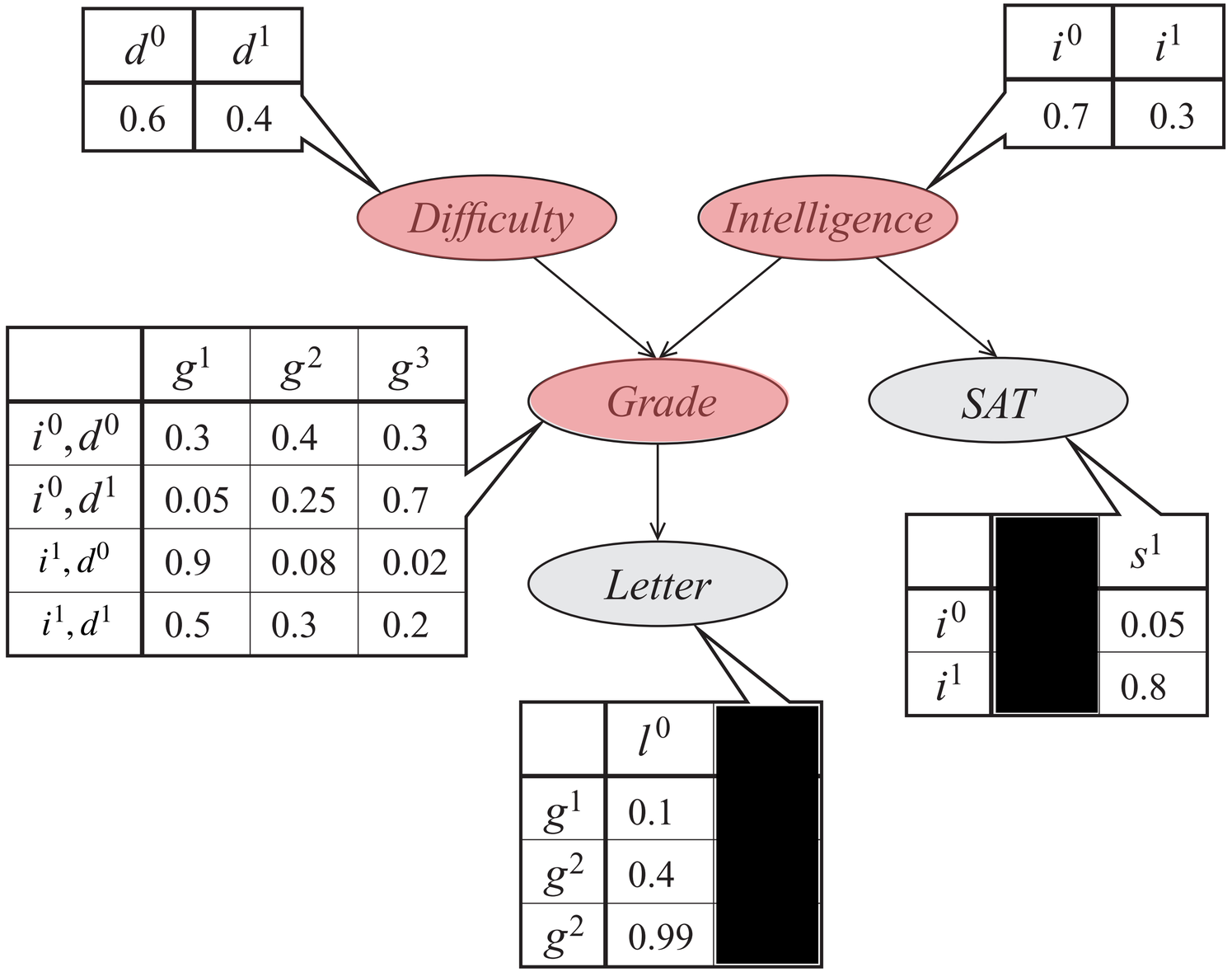
# Example

$\tilde{P}_\Phi(D, I, G \mid s^1, l^0)$

$d^0\ i^0\ g^1$

$d^1\ i^0\ g^1$

$d^1\ i^1\ g^1$




| $d^0$ | $d^1$ |
|---|---|
| 0.6 | 0.4 |

| $i^0$ | $i^1$ |
|---|---|
| 0.7 | 0.3 |

| | $g^1$ | $g^2$ | $g^3$ |
|---|---|---|---|
| $i^0, d^0$ | 0.3 | 0.4 | 0.3 |
| $i^0, d^1$ | 0.05 | 0.25 | 0.7 |
| $i^1, d^0$ | 0.9 | 0.08 | 0.02 |
| $i^1, d^1$ | 0.5 | 0.3 | 0.2 |

| | | $s^1$ |
|---|---|---|
| $i^0$ | | 0.05 |
| $i^1$ | | 0.8 |

| | $l^0$ |
|---|---|
| $g^1$ | 0.1 |
| $g^2$ | 0.4 |
| $g^2$ | 0.99 |

$P(D \mid i^0, g^1, l^0, s^1)$

$P(I \mid d^1, g^1, l^0, s^1)$

$P(G \mid d^1, i^1, l^0, s^1)$

# Example

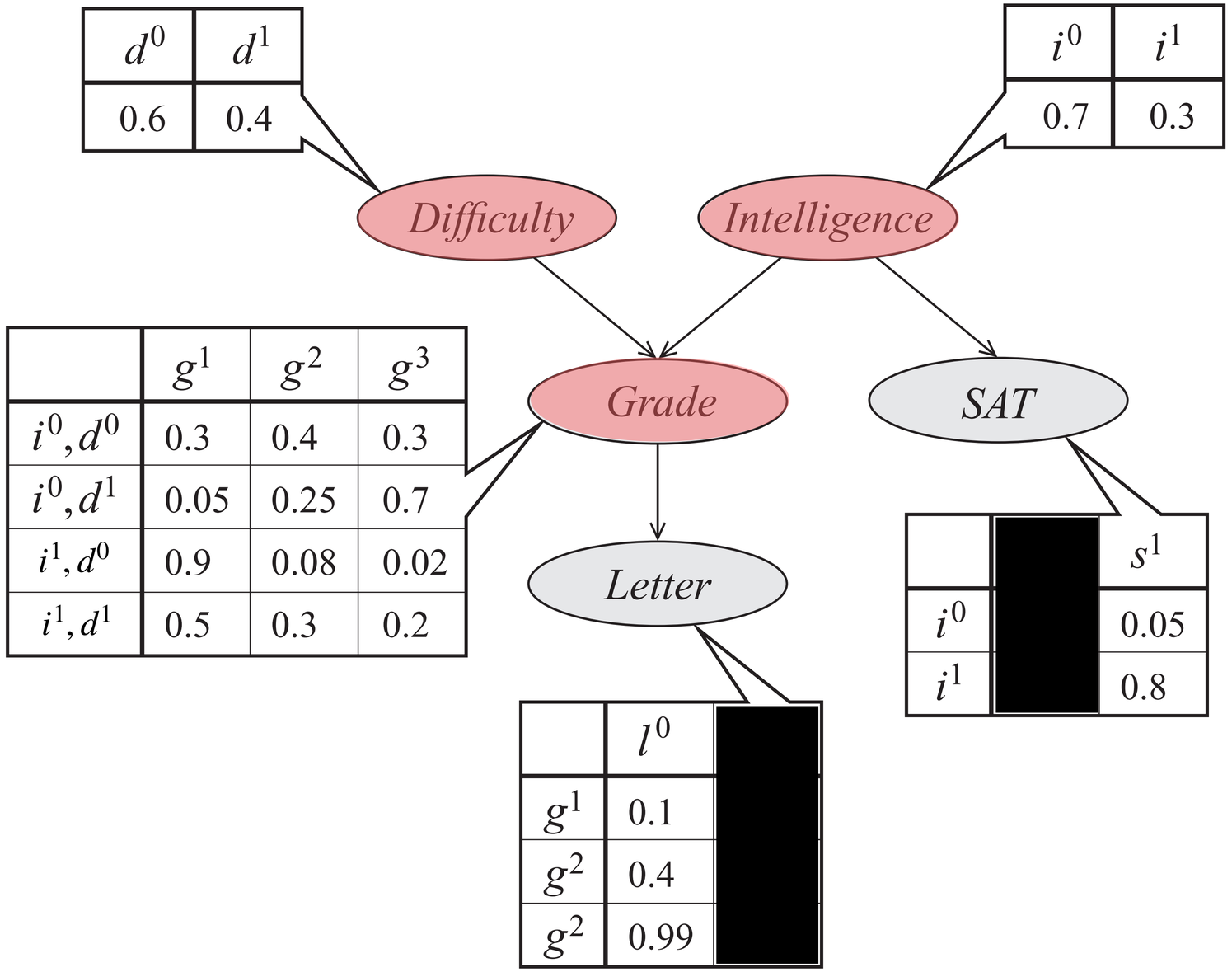
$\tilde{P}_\Phi(D, I, G \mid s^1, l^0)$

$d^0\ i^0\ g^1$

$d^1\ i^0\ g^1$

$d^1\ i^1\ g^1$

$d^1\ i^1\ g^3$




| $d^0$ | $d^1$ |
|---|---|
| 0.6 | 0.4 |

| $i^0$ | $i^1$ |
|---|---|
| 0.7 | 0.3 |

| | $g^1$ | $g^2$ | $g^3$ |
|---|---|---|---|
| $i^0, d^0$ | 0.3 | 0.4 | 0.3 |
| $i^0, d^1$ | 0.05 | 0.25 | 0.7 |
| $i^1, d^0$ | 0.9 | 0.08 | 0.02 |
| $i^1, d^1$ | 0.5 | 0.3 | 0.2 |

| | [redacted] | $s^1$ |
|---|---|---|
| $i^0$ | | 0.05 |
| $i^1$ | | 0.8 |

| | $l^0$ | [redacted] |
|---|---|---|
| $g^1$ | 0.1 | |
| $g^2$ | 0.4 | |
| $g^2$ | 0.99 | |

$P(D \mid i^0, g^1, l^0, s^1)$

$P(I \mid d^1, g^1, l^0, s^1)$

$P(G \mid d^1, i^1, l^0, s^1)$

# Example

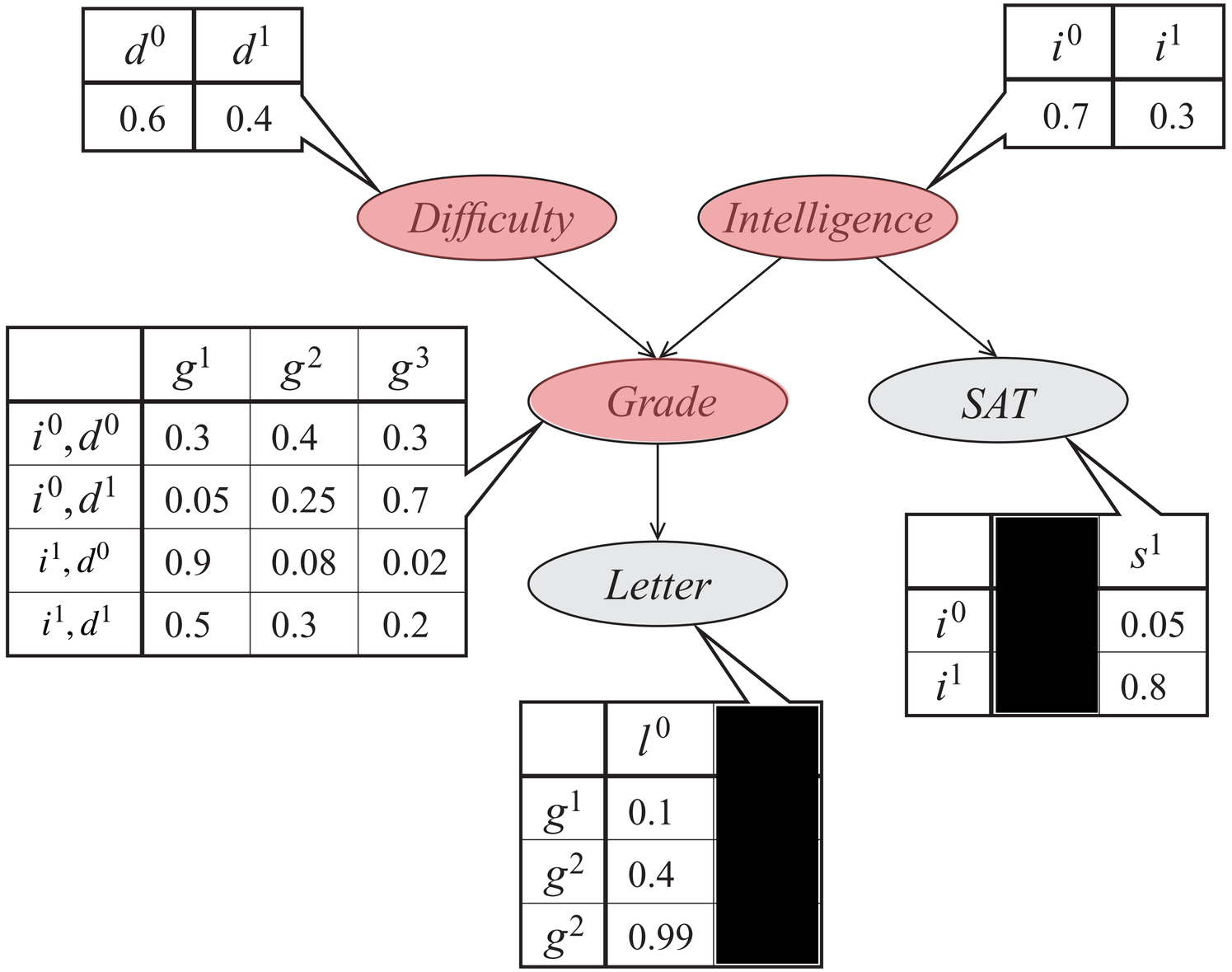
$\tilde{P}_\Phi(D, I, G \mid s^1, l^0)$

$d^0\ i^0\ g^1$

$d^1\ i^0\ g^1$

$d^1\ i^1\ g^1$

$d^1\ i^1\ g^3$




| $d^0$ | $d^1$ |
|---|---|
| 0.6 | 0.4 |

| $i^0$ | $i^1$ |
|---|---|
| 0.7 | 0.3 |

| | $g^1$ | $g^2$ | $g^3$ |
|---|---|---|---|
| $i^0, d^0$ | 0.3 | 0.4 | 0.3 |
| $i^0, d^1$ | 0.05 | 0.25 | 0.7 |
| $i^1, d^0$ | 0.9 | 0.08 | 0.02 |
| $i^1, d^1$ | 0.5 | 0.3 | 0.2 |

| | [redacted] | $s^1$ |
|---|---|---|
| $i^0$ | | 0.05 |
| $i^1$ | | 0.8 |

| | $l^0$ | [redacted] |
|---|---|---|
| $g^1$ | 0.1 | |
| $g^2$ | 0.4 | |
| $g^2$ | 0.99 | |

$P(D \mid i^0, g^1, l^0, s^1)$

$P(I \mid d^1, g^1, l^0, s^1)$

$P(G \mid d^1, i^1, l^0, s^1)$

# Computational cost

- For $i = 1,\ldots,n$
  - Sample $x_i \sim P_\Phi(X_i \mid \boldsymbol{x}_{-i})$

$$P_\Phi(X_i \mid \boldsymbol{x}_{-i}) = \frac{P_\Phi(X_i, \boldsymbol{x}_{-i})}{P_\Phi(\boldsymbol{x}_{-i})} = \frac{\tilde{P}_\Phi(X_i, \boldsymbol{x}_{-i})}{\tilde{P}_\Phi(\boldsymbol{x}_{-i})}$$

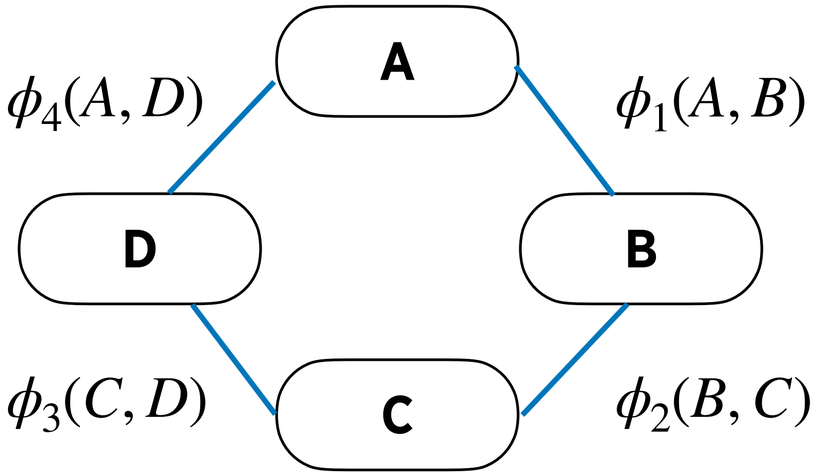
# Computational cost

- For $i = 1,\ldots,n$
  - Sample $x_i \sim P_\Phi(X_i \mid \boldsymbol{x}_{-i})$

$$P_\Phi(X_i \mid \boldsymbol{x}_{-i}) = \frac{P_\Phi(X_i, \boldsymbol{x}_{-i})}{P_\Phi(\boldsymbol{x}_{-i})} = \frac{\tilde{P}_\Phi(X_i, \boldsymbol{x}_{-i})}{\tilde{P}_\Phi(\boldsymbol{x}_{-i})}$$

**Complete assignment**
**Product of factors**

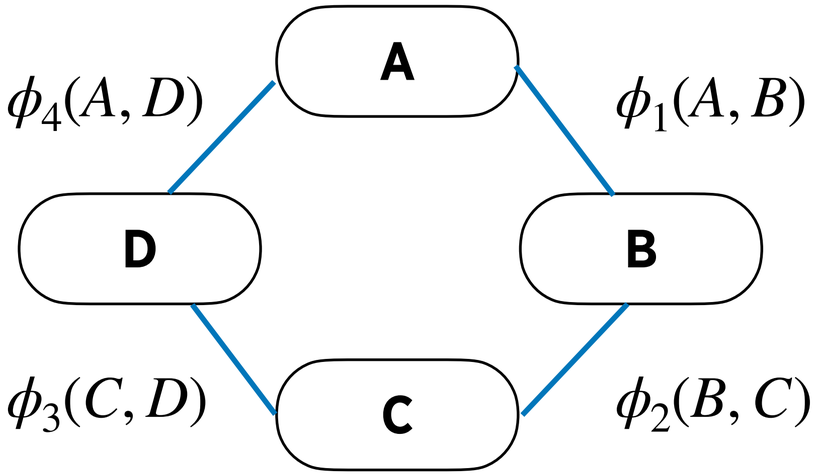
# Another example




$$P_\Phi(A \mid b, c, d) = \frac{\tilde{P}_\Phi(a, b, c, d)}{\sum_{A'} \tilde{P}_\Phi(A', b, c, d)}$$

$$\frac{\phi_1(A, b)\, \phi_2(b, c)\, \phi_3(c, d)\, \phi_4(A, d)}{\sum_{A'} \phi_1(A', b)\, \phi_2(b, c)\, \phi_3(c, d)\, \phi_4(A', d)}$$

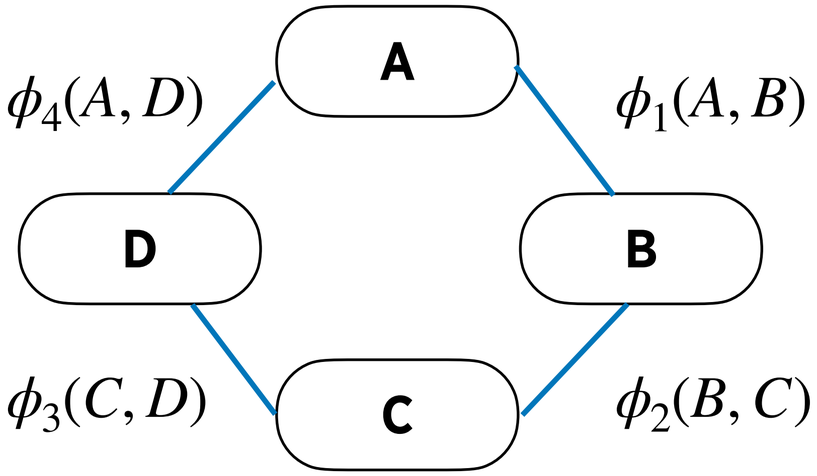
# Another example




$$P_\Phi(A \mid b, c, d) = \frac{\tilde{P}_\Phi(a, b, c, d)}{\sum_{A'} \tilde{P}_\Phi(A', b, c, d)}$$

$$\frac{\phi_1(A, b)\, \phi_2(b, c)\, \phi_3(c, d)\, \phi_4(A, d)}{\sum_{A'} \phi_1(A', b)\, \phi_2(b, c)\, \phi_3(c, d)\, \phi_4(A', d)}$$

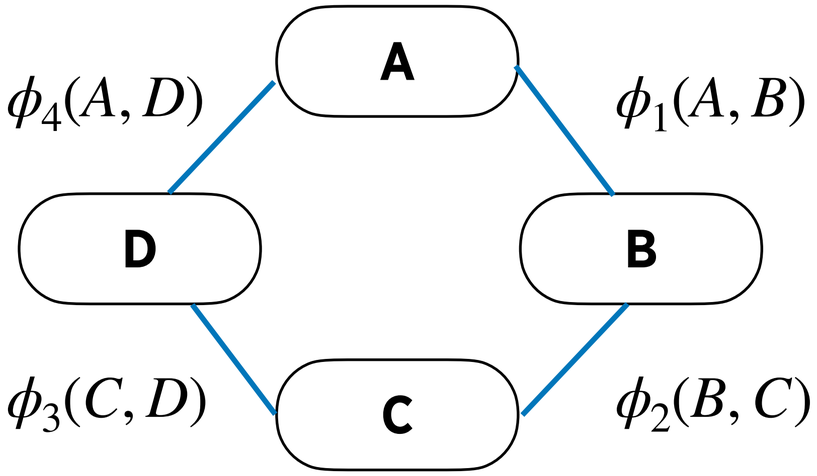
# Another example




$$P_\Phi(A \mid b, c, d) = \frac{\tilde{P}_\Phi(a, b, c, d)}{\sum_{A'} \tilde{P}_\Phi(A', b, c, d)}$$

$$\frac{\phi_1(A, b)\, \phi_2(b, c)\, \phi_3(c, d)\, \phi_4(A, d)}{\sum_{A'} \phi_1(A', b)\, \phi_2(b, c)\, \phi_3(c, d)\, \phi_4(A', d)}$$

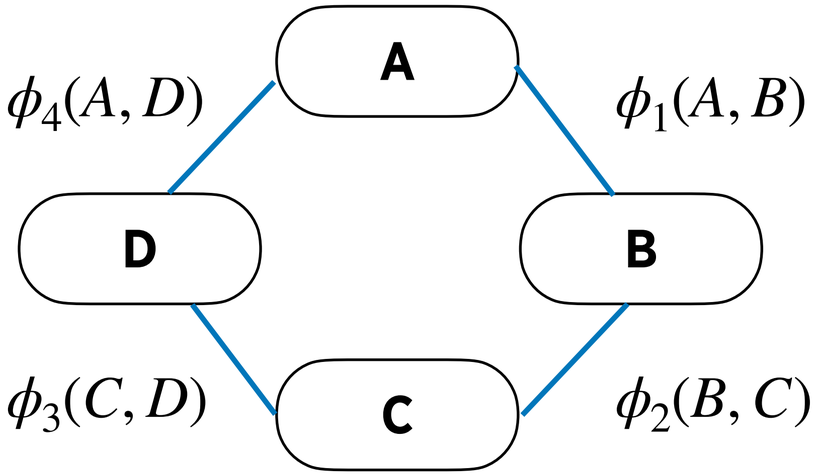
# Another example




$$P_\Phi(A \mid b, c, d) = \frac{\tilde{P}_\Phi(a, b, c, d)}{\sum_{A'} \tilde{P}_\Phi(A', b, c, d)}$$

$$\frac{\phi_1(A, b)\, \phi_2(b, c)\, \phi_3(c, d)\, \phi_4(A, d)}{\sum_{A'} \phi_1(A', b)\, \phi_2(b, c)\, \phi_3(c, d)\, \phi_4(A', d)}$$

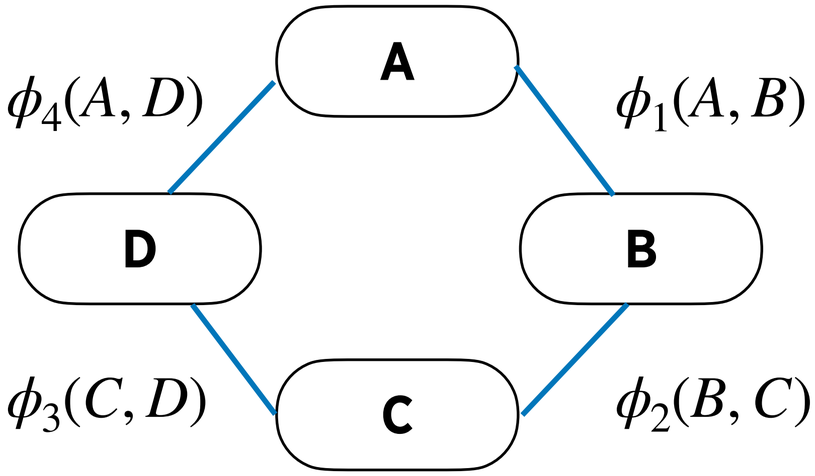
# Another example




$$P_\Phi(A \mid b, c, d) = \frac{\tilde{P}_\Phi(a, b, c, d)}{\sum_{A'} \tilde{P}_\Phi(A', b, c, d)}$$

$$\frac{\phi_1(A, b)\, \phi_2(b, c)\, \phi_3(c, d)\, \phi_4(A, d)}{\sum_{A'} \phi_1(A', b)\, \phi_2(b, c)\, \phi_3(c, d)\, \phi_4(A', d)}$$

# Another example

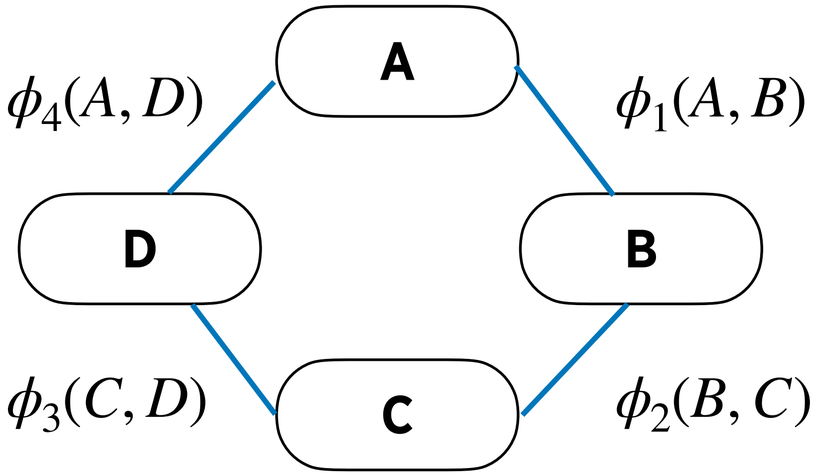



$$P_\Phi(A \mid b, c, d) = \frac{\tilde{P}_\Phi(a, b, c, d)}{\sum_{A'} \tilde{P}_\Phi(A', b, c, d)}$$

$$\frac{\phi_1(A, b)\, \phi_2(b, c)\, \phi_3(c, d)\, \phi_4(A, d)}{\sum_{A'} \phi_1(A', b)\, \phi_2(b, c)\, \phi_3(c, d)\, \phi_4(A', d)}$$

**Factors that involve $A$**

# Another example



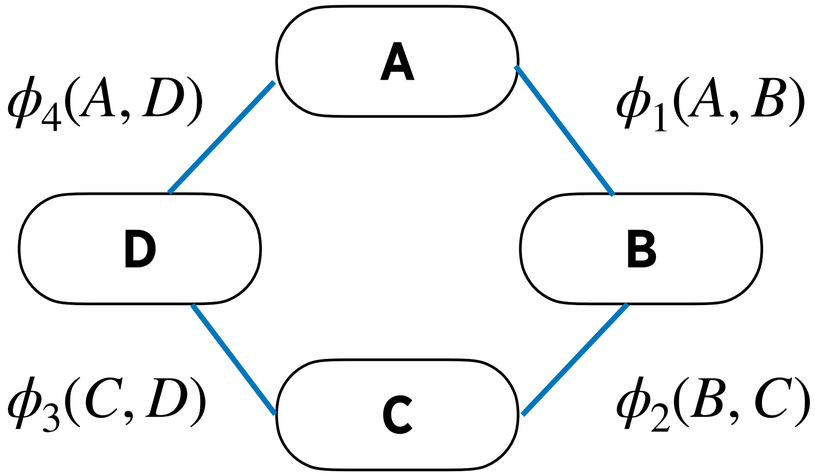

$$P_\Phi(A \mid b, c, d) = \frac{\tilde{P}_\Phi(a, b, c, d)}{\sum_{A'} \tilde{P}_\Phi(A', b, c, d)}$$

$$\frac{\phi_1(A, b)\, \phi_2(b, c)\, \phi_3(c, d)\, \phi_4(A, d)}{\sum_{A'} \phi_1(A', b)\, \phi_2(b, c)\, \phi_3(c, d)\, \phi_4(A', d)}$$

**Factors that involve $A$**

**normalising constant**

# Another example




$$P_\Phi(A \mid b, c, d) = \frac{\tilde{P}_\Phi(a, b, c, d)}{\sum_{A'} \tilde{P}_\Phi(A', b, c, d)}$$

$$\frac{\phi_1(A, b)\, \phi_2(b, c)\, \phi_3(c, d)\, \phi_4(A, d)}{\sum_{A'} \phi_1(A', b)\, \phi_2(b, c)\, \phi_3(c, d)\, \phi_4(A', d)}$$

**Factors that involve $A$** **normalising constant**

$$\propto \phi_1(A, b)\, \phi_4(A, d)$$

# Computational cost revisited

- For $i = 1,\ldots,n$
  - Sample $x_i \sim P_\Phi(X_i \mid \boldsymbol{x}_{-i})$

$$P_\Phi(X_i \mid \boldsymbol{x}_{-i}) = \frac{P_\Phi(X_i, \boldsymbol{x}_{-i})}{P_\Phi(\boldsymbol{x}_{-i})} = \frac{\tilde{P}_\Phi(X_i, \boldsymbol{x}_{-i})}{\tilde{P}_\Phi(\boldsymbol{x}_{-i})}$$

$$\propto \prod_{j : X_i \in \text{scope}[C_j]} \phi_j(X_i, \boldsymbol{x}_{j,-i})$$

# Computational cost revisited

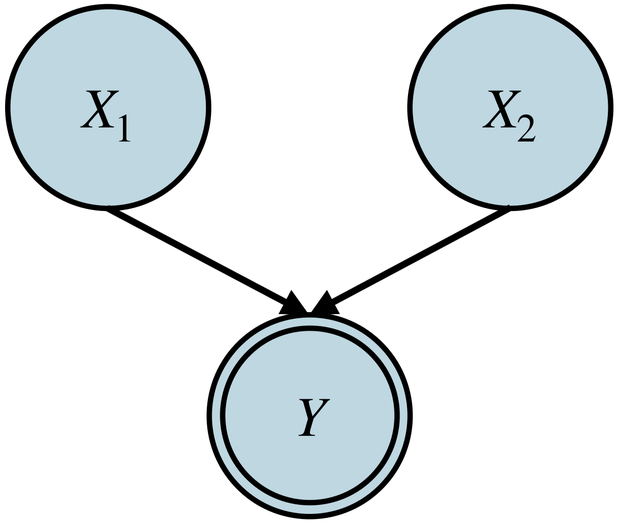
- For $i = 1,\ldots,n$
  - Sample $x_i \sim P_\Phi(X_i \mid \boldsymbol{x}_{-i})$

$$P_\Phi(X_i \mid \boldsymbol{x}_{-i}) = \frac{P_\Phi(X_i, \boldsymbol{x}_{-i})}{P_\Phi(\boldsymbol{x}_{-i})} = \frac{\tilde{P}_\Phi(X_i, \boldsymbol{x}_{-i})}{\tilde{P}_\Phi(\boldsymbol{x}_{-i})}$$

only $X_i$ and its neighbours

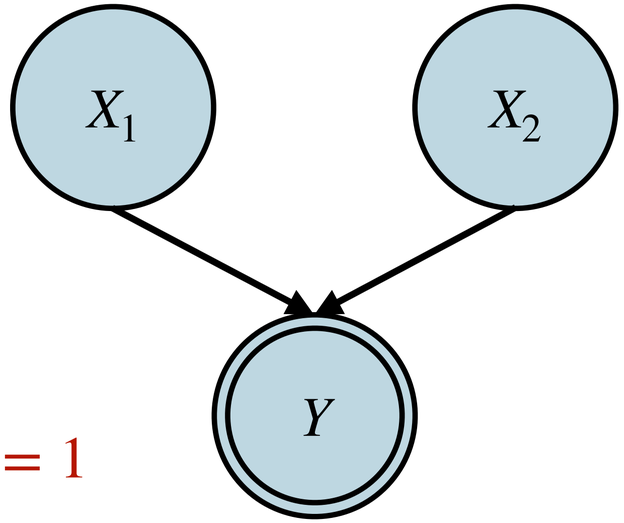
$$\propto \prod_{j:X_i \in \text{scope}[C_j]} \phi_j(X_i, \boldsymbol{x}_{j,-i})$$

# Gibbs chain and regularity




| $X_1$ | $X_2$ | Y | Prob |
|---|---|---|---|
| 0 | 0 | 0 | 0.25 |
| 0 | 1 | 1 | 0.25 |
| 1 | 0 | 1 | 0.25 |
| 1 | 1 | 0 | 0.25 |

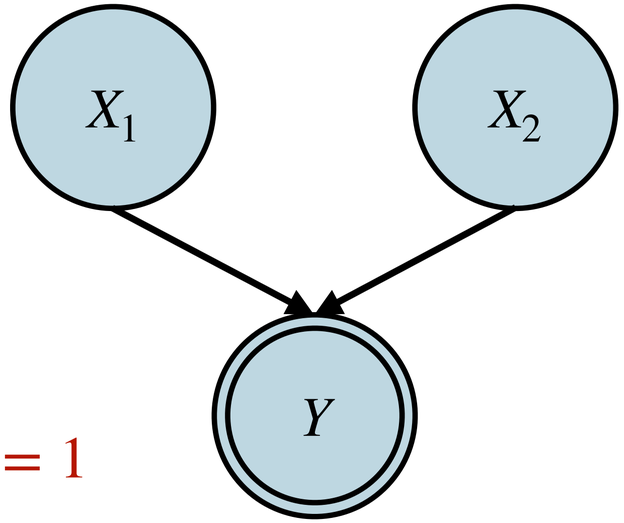
# Gibbs chain and regularity




$Y = 1$

| $X_1$ | $X_2$ | Y | Prob |
|---|---|---|---|
| 0 | 0 | 0 | 0.25 |
| 0 | 1 | 1 | 0.25 |
| 1 | 0 | 1 | 0.25 |
| 1 | 1 | 0 | 0.25 |

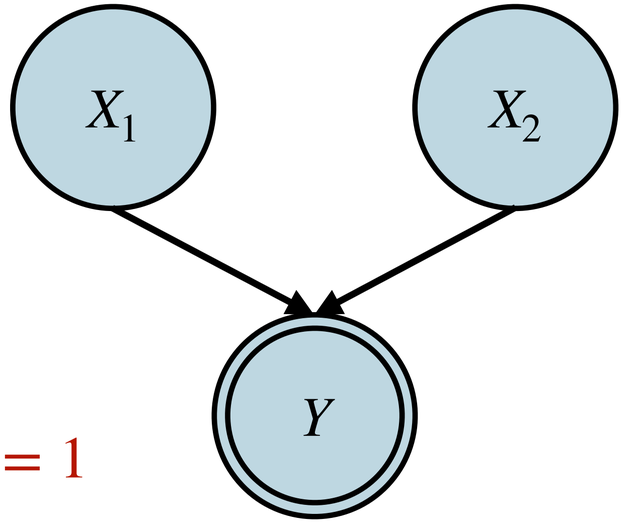
# Gibbs chain and regularity




$Y = 1$

| $X_1$ | $X_2$ | Y | Prob |
|---|---|---|---|
| ~~0~~ | ~~0~~ | ~~0~~ | ~~0.25~~ |
| 0 | 1 | 1 | 0.25 |
| 1 | 0 | 1 | 0.25 |
| ~~1~~ | ~~1~~ | ~~0~~ | ~~0.25~~ |

# Gibbs chain and regularity

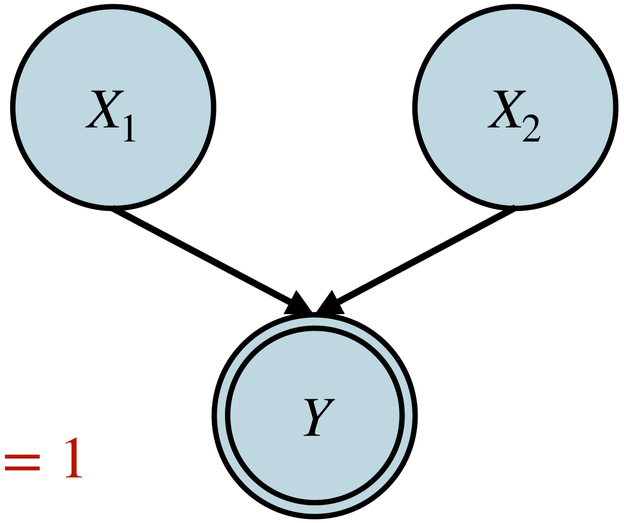



$Y = 1$

| $X_1$ | $X_2$ | Y | Prob |
|---|---|---|---|
| ~~0~~ | ~~0~~ | ~~0~~ | ~~0.25~~ |
| 0 | 1 | 1 | 0.25 |
| 1 | 0 | 1 | 0.25 |
| ~~1~~ | ~~1~~ | ~~0~~ | ~~0.25~~ |

# Gibbs chain and regularity

$X_1$ $X_2$ $Y = 1$



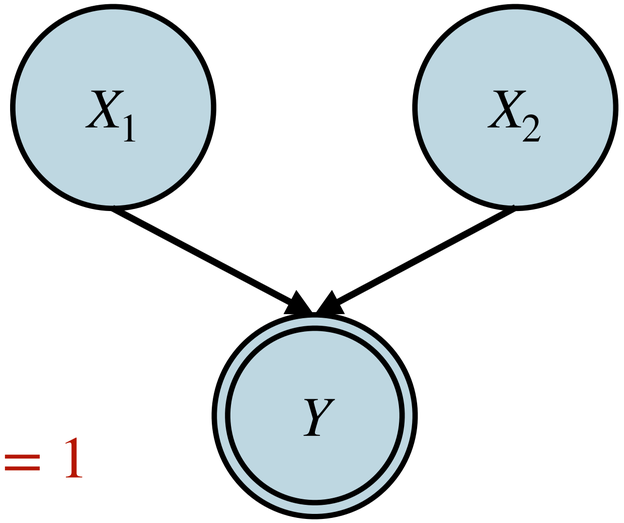

$Y = 1$

| $X_1$ | $X_2$ | Y | Prob |
|---|---|---|---|
| ~~0~~ | ~~0~~ | ~~0~~ | ~~0.25~~ |
| 0 | 1 | 1 | 0.25 |
| 1 | 0 | 1 | 0.25 |
| ~~1~~ | ~~1~~ | ~~0~~ | ~~0.25~~ |

# Gibbs chain and regularity

$X_1$ $X_2$ $Y = 1$

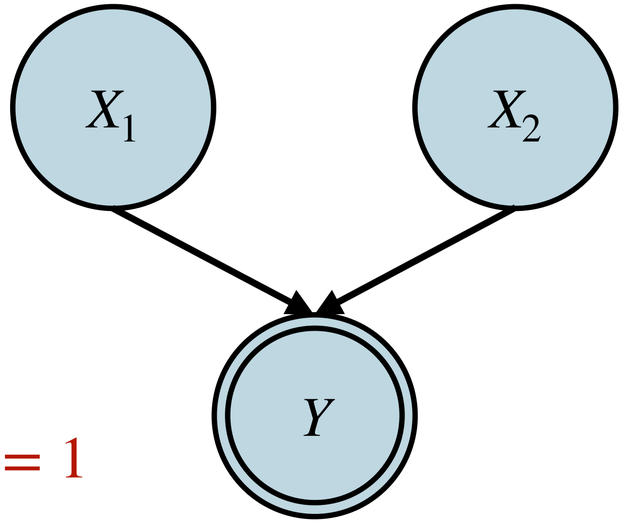
0 1




$Y = 1$

| $X_1$ | $X_2$ | Y | Prob |
|---|---|---|---|
| ~~0~~ | ~~0~~ | ~~0~~ | ~~0.25~~ |
| 0 | 1 | 1 | 0.25 |
| 1 | 0 | 1 | 0.25 |
| ~~1~~ | ~~1~~ | ~~0~~ | ~~0.25~~ |

# Gibbs chain and regularity




$Y = 1$

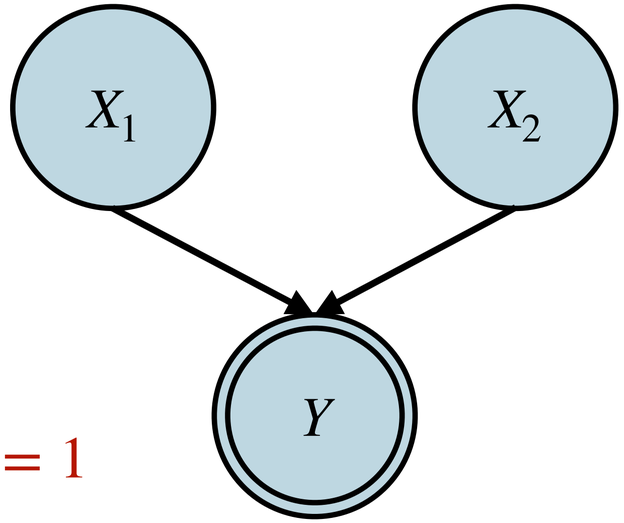
| $X_1$ | $X_2$ | Y | Prob |
|---|---|---|---|
| ~~0~~ | ~~0~~ | ~~0~~ | ~~0.25~~ |
| 0 | 1 | 1 | 0.25 |
| 1 | 0 | 1 | 0.25 |
| ~~1~~ | ~~1~~ | ~~0~~ | ~~0.25~~ |

$X_1$ $X_2$ $\quad Y = 1$

0 1

0 1

# Gibbs chain and regularity




$Y = 1$

| $X_1$ | $X_2$ | Y | Prob |
|---|---|---|---|
| ~~0~~ | ~~0~~ | ~~0~~ | ~~0.25~~ |
| 0 | 1 | 1 | 0.25 |
| 1 | 0 | 1 | 0.25 |
| ~~1~~ | ~~1~~ | ~~0~~ | ~~0.25~~ |

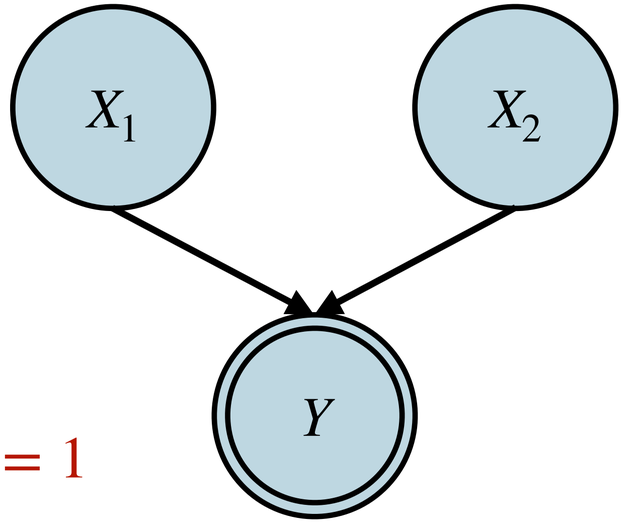
$X_1$ $X_2$ $\quad Y = 1$

0 1

0 1

0 1

# Gibbs chain and regularity




$Y = 1$

| $X_1$ | $X_2$ | Y | Prob |
|---|---|---|---|
| ~~0~~ | ~~0~~ | ~~0~~ | ~~0.25~~ |
| 0 | 1 | 1 | 0.25 |
| 1 | 0 | 1 | 0.25 |
| ~~1~~ | ~~1~~ | ~~0~~ | ~~0.25~~ |

$X_1$ $X_2$ $\quad Y = 1$

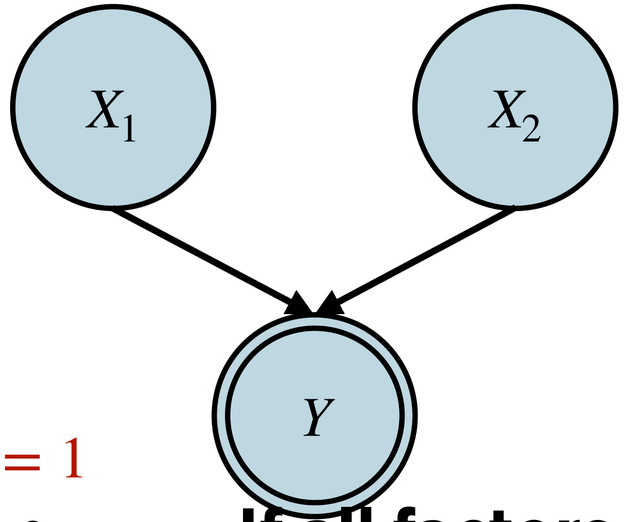
0 1

0 1

0 1

0 1

# Gibbs chain and regularity




$Y = 1$

| $X_1$ | $X_2$ | Y | Prob |
|---|---|---|---|
| ~~0~~ | ~~0~~ | ~~0~~ | ~~0.25~~ |
| 0 | 1 | 1 | 0.25 |
| 1 | 0 | 1 | 0.25 |
| ~~1~~ | ~~1~~ | ~~0~~ | ~~0.25~~ |

$X_1$ $X_2$ $Y = 1$

0 1

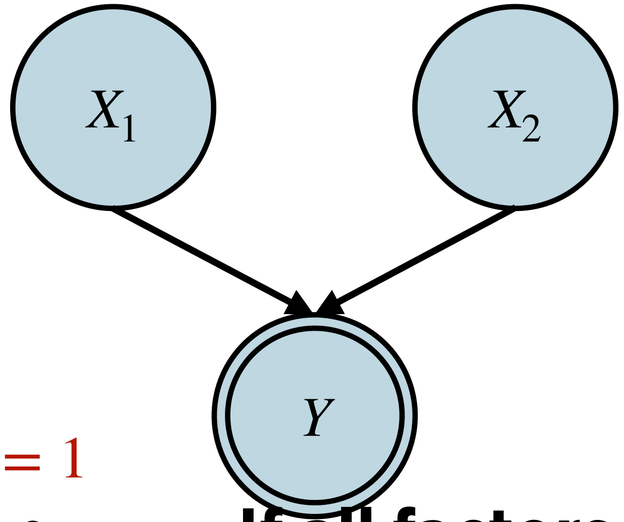
0 1

0 1

0 1

- If all factors are positive, Gibbs chain is regular (if there are no zero entries in any of the factors)

# Gibbs chain and regularity




$Y = 1$

| $X_1$ | $X_2$ | Y | Prob |
|---|---|---|---|
| ~~0~~ | ~~0~~ | ~~0~~ | ~~0.25~~ |
| 0 | 1 | 1 | 0.25 |
| 1 | 0 | 1 | 0.25 |
| ~~1~~ | ~~1~~ | ~~0~~ | ~~0.25~~ |

$X_1$ $X_2$ $Y = 1$

0 1

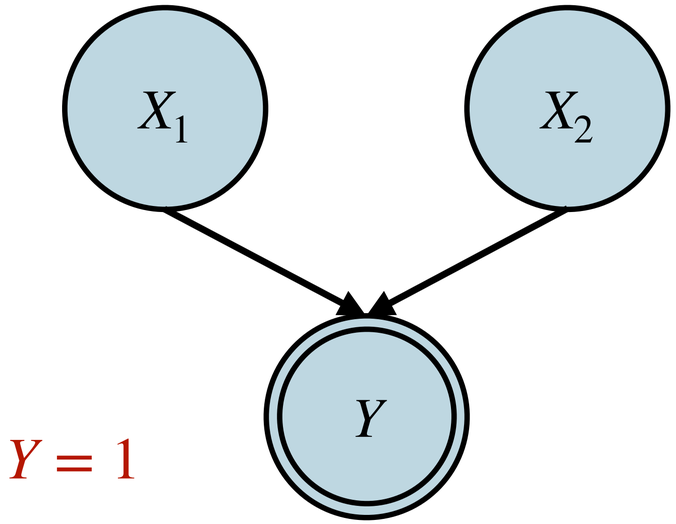
0 1

0 1

0 1

- If all factors are positive, Gibbs chain is regular (if there are no zero entries in any of the factors)
- However, mixing can still be very slow

# Gibbs chain and regularity




| $X_1$ | $X_2$ | Y | Prob |
|---|---|---|---|
| ~~0~~ | ~~0~~ | ~~0~~ | ~~0.25~~ |
| 0 | 1 | 1 | 0.25 |
| 1 | 0 | 1 | 0.25 |
| ~~1~~ | ~~1~~ | ~~0~~ | ~~0.25~~ |

- If all factors are positive, Gibbs chain is regular (if there are no zero entries in any of the factors)
- However, mixing can still be very slow

Image: https://lime.data-imaginist.com

# Summary

# Summary

- Converts the hard problem of inference to a sequence of “easy” sampling steps

# Summary

- Converts the hard problem of inference to a sequence of “easy” sampling steps
- Pros:
  - Probably the simplest Markov chain for PGMss
  - Computationally efficient to sample

# Summary

- Converts the hard problem of inference to a sequence of “easy” sampling steps
- Pros:
  - Probably the simplest Markov chain for PGMss
  - Computationally efficient to sample
- Cons:
  - Often slow to mix, esp. when probabilities are peaked
  - Only applies if we can sample from product of factors