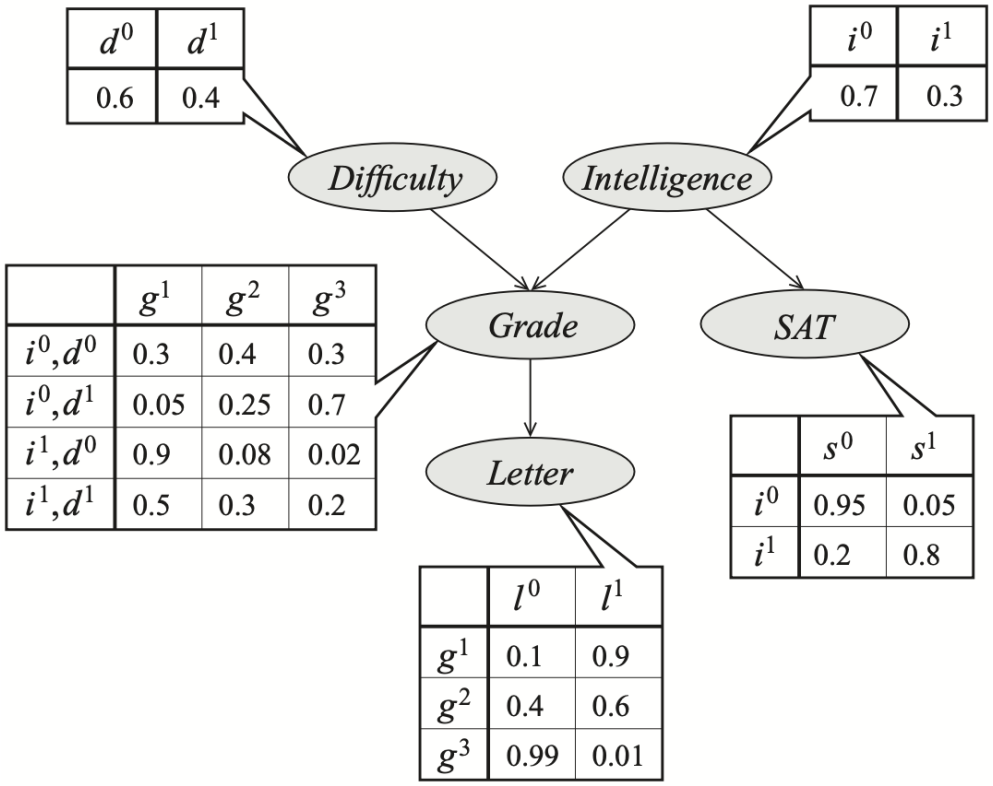
Sampling for graphical models

Inference in Bayesian Networks

Goal: compute queries like $P(I = i^1 \mid G = g^2, S = s^0)$ or in words "the probability of high intelligence given medium grade and low SAT?"



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

$$P(L, G, S, D, I) = P(L \mid G)P(G \mid D, I)P(S \mid I)P(D)P(I)$$

Inference in Bayesian Networks

Goal: compute queries like $P(I = i^1 \mid G = g^2, S = s^0)$ or in words "the probability of high intelligence given medium grade and low SAT?"

$$P(I = i^1 \mid G = g^2, S = s^0) = \frac{P(I = i^1, G = g^2, S = s^0)}{P(G = g^2, S = s^0)}$$

Approaches:

- **Exact inference by variable elimination:**

$$P(L, G, S, D, I) = P(L \mid G)P(G \mid D, I)P(S \mid I)P(D)P(I)$$

$$P(I = i^1, G = g^2, S = s^0) = \sum_{l,d} P(L = l, G = g^2, S = s^0, D = d, I = i^1)$$

Inference in Bayesian Networks

Goal: compute queries like $P(I = i^1 \mid G = g^2, S = s^0)$ or in words "the probability of high intelligence given medium grade and low SAT?"

$$P(I = i^1 \mid G = g^2, S = s^0) = \frac{P(I = i^1, G = g^2, S = s^0)}{P(G = g^2, S = s^0)}$$

Approaches:

- **Exact inference by variable elimination:**

$$P(L, G, S, D, I) = P(L \mid G)P(G \mid D, I)P(S \mid I)P(D)P(I)$$

$$P(I = i^1, G = g^2, S = s^0) = \sum_{l,d} P(L = l, G = g^2, S = s^0, D = d, I = i^1)$$

$$= \sum_{l,d} P(L = l \mid G = g^2)P(G = g^2 \mid D = d, I = i^1)P(S = s^0 \mid I = i^1)P(D = d)P(I = i^1)$$

Inference in Bayesian Networks

Goal: compute queries like $P(I = i^1 \mid G = g^2, S = s^0)$ or in words "the probability of high intelligence given medium grade and low SAT?"

$$P(I = i^1 \mid G = g^2, S = s^0) = \frac{P(I = i^1, G = g^2, S = s^0)}{P(G = g^2, S = s^0)}$$

Approaches:

- **Exact inference by variable elimination:**

$$P(L, G, S, D, I) = P(L \mid G)P(G \mid D, I)P(S \mid I)P(D)P(I)$$

$$P(I = i^1, G = g^2, S = s^0) = \sum_{l,d} P(L = l, G = g^2, S = s^0, D = d, I = i^1)$$

$$= \sum_{l,d} P(L = l \mid G = g^2)P(G = g^2 \mid D = d, I = i^1)P(S = s^0 \mid I = i^1)P(D = d)P(I = i^1)$$

$$= \left(\sum_{d} P(G = g^2 \mid D = d, I = i^1)P(D = d)\right)P(S = s^0 \mid I = i^1)P(I = i^1)$$

$$P(G = g^2, S = s^0) = \sum_{i} \left(\sum_{d} P(G = g^2 \mid D = d, I = i)P(D = d)\right)P(S = s^0 \mid I = i)P(I = i)$$

Inference in Bayesian Networks

Goal: compute queries like $P(I = i^1 \mid G = g^2, S = s^0)$ or in words "the probability of high intelligence given medium grade and low SAT?"

$$P(I = i^1 \mid G = g^2, S = s^0) = \frac{P(I = i^1, G = g^2, S = s^0)}{P(G = g^2, S = s^0)}$$

Approaches:

- **Exact inference by variable elimination**
  - Straightforward, but can be computationally prohibitive
- **Variational inference**
  - Approximate, biased
- **Sampling strategies**
  - Rejection sampling
  - Importance weighted sampling
  - MCMC

Rejection sampling

Goal: compute queries like $P(I = i^1 \mid G = g^2, S = s^0)$ or in words "the probability of high intelligence given medium grade and low SAT?"

$$P(I = i^1 \mid G = g^2, S = s^0) = \frac{P(I = i^1, G = g^2, S = s^0)}{P(G = g^2, S = s^0)}$$

Algorithm 12.1 Forward Sampling in a Bayesian network

Procedure Forward-Sample (
  $\mathcal{B}$ // Bayesian network over $\mathcal{X}$
)

1 Let $X_1, \ldots, X_n$ be a topological ordering of $\mathcal{X}$
2 **for** $i = 1, \ldots, n$
3 $\boldsymbol{u}_i \leftarrow \boldsymbol{x}\langle \mathrm{Pa}_{X_i} \rangle$ // Assignment to Pa_{X_i} in $x_1, \ldots, x_{i-1}$
4 Sample x_i from $P(X_i \mid \boldsymbol{u}_i)$
5 **return** $(x_1, \ldots, x_n)$

Rejection sampling

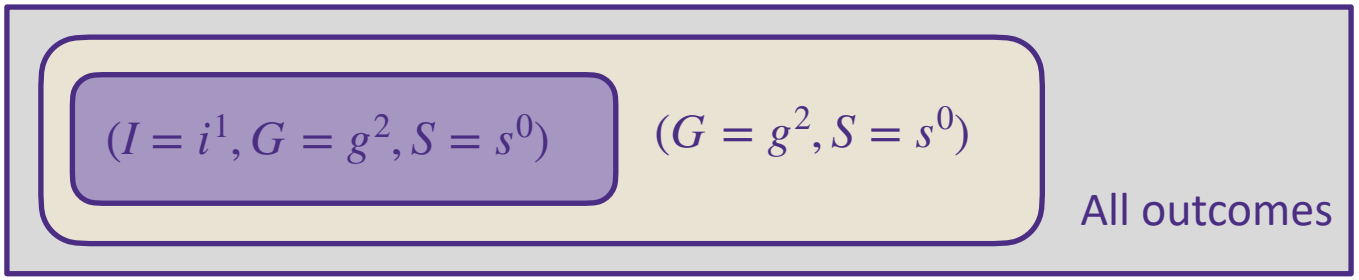
Goal: compute queries like $P(I = i^1 \mid G = g^2, S = s^0)$ or in words "the probability of high intelligence given medium grade and low SAT?"

$$P(I = i^1 \mid G = g^2, S = s^0) = \frac{P(I = i^1, G = g^2, S = s^0)}{P(G = g^2, S = s^0)}$$

Draw M samples $\{(L_k, G_k, S_k, D_k, I_k)\}_{k=1}^{M}$ using forward sampling and

Set $\chi = \{k : G_k = g^2, S_k = s^0\}$ and output

$$\widehat{P}(I = i^1 \mid G = g^2, S = s^0) = \frac{1}{|\chi|} \sum_{k \in \chi} \mathbf{1}\{I_k = i^1, G_k = g^2, S_k = s^0\}$$



Rejection sampling

Goal: compute queries like $P(I = i^1 \mid G = g^2, S = s^0)$ or in words "the probability of high intelligence given medium grade and low SAT?"

$$P(I = i^1 \mid G = g^2, S = s^0) = \frac{P(I = i^1, G = g^2, S = s^0)}{P(G = g^2, S = s^0)}$$

Draw M samples $\{(L_k, G_k, S_k, D_k, I_k)\}_{k=1}^{M}$ using forward sampling and

Set $\chi = \{k : G_k = g^2, S_k = s^0\}$ and output

$$\widehat{P}(I = i^1 \mid G = g^2, S = s^0) = \frac{1}{|\chi|} \sum_{k \in \chi} \mathbf{1}\{I_k = i^1, G_k = g^2, S_k = s^0\}$$

$$\mathbb{E}\left[\widehat{P}(I = i^1 \mid G = g^2, S = s^0)\right] = \mathbb{E}\left[\frac{1}{|\chi|} \sum_{k \in \chi} \mathbf{1}\{I_k = i^1, G_k = g^2, S_k = s^0\}\right]$$

$$= \sum_{n=1}^{\infty} \mathbb{E}\left[\frac{1}{n} \sum_{k \in \chi} \mathbf{1}\{I_k = i^1, G_k = g^2, S_k = s^0\} \,\middle|\, |\chi| = n\right] \mathbb{P}(|\chi| = n)$$

$$= \sum_{n=1}^{\infty} \frac{n P(I = i^1 \mid G = g^2, S = s^0)}{n} \mathbb{P}(|\chi| = n)$$

Unbiased estimator of $P(I = i^1 \mid G = g^2, S = s^0)$!

Rejection sampling

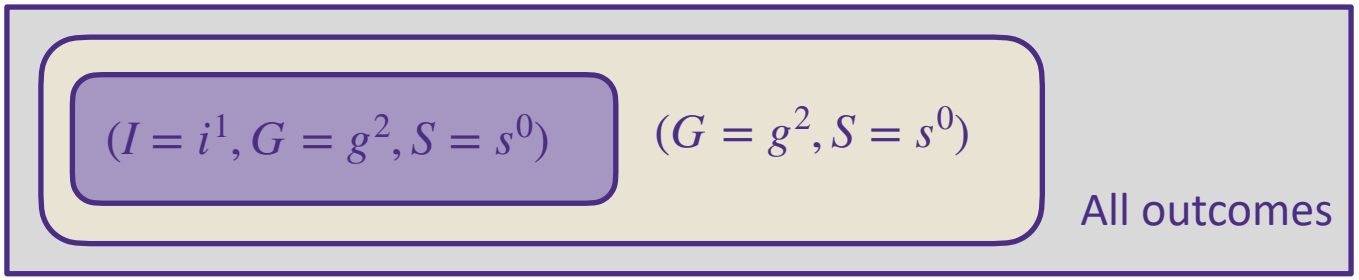
Goal: compute queries like $P(I = i^1 \mid G = g^2, S = s^0)$ or in words "the probability of high intelligence given medium grade and low SAT?"

$$P(I = i^1 \mid G = g^2, S = s^0) = \frac{P(I = i^1, G = g^2, S = s^0)}{P(G = g^2, S = s^0)}$$

Draw M samples $\{(L_k, G_k, S_k, D_k, I_k)\}_{k=1}^{M}$ using forward sampling and

Set $\chi = \{k : G_k = g^2, S_k = s^0\}$ and output

$$\widehat{P}(I = i^1 \mid G = g^2, S = s^0) = \frac{1}{|\chi|} \sum_{k \in \chi} \mathbf{1}\{I_k = i^1, G_k = g^2, S_k = s^0\}$$



How big goes M need to be to get an accurate estimate?

Rejection sampling

Goal: compute queries like $P(I = i^1 \mid G = g^2, S = s^0)$ or in words "the probability of high intelligence given medium grade and low SAT?"

$$P(I = i^1 \mid G = g^2, S = s^0) = \frac{P(I = i^1, G = g^2, S = s^0)}{P(G = g^2, S = s^0)}$$

Draw M samples $\{(L_k, G_k, S_k, D_k, I_k)\}_{k=1}^{M}$ using forward sampling and

Set $\chi = \{k : G_k = g^2, S_k = s^0\}$ and output

$$\widehat{P}(I = i^1 \mid G = g^2, S = s^0) = \frac{1}{|\chi|} \sum_{k \in \chi} \mathbf{1}\{I_k = i^1, G_k = g^2, S_k = s^0\}$$

Rejection sampling takeaways:

- Very simple to implement
- May require an enormous amount of samples if the conditional statement is rare.
    - Consider P(disease | symptoms). Any precise set of symptoms is going to be rare.

Inference in Bayesian Networks

Goal: compute queries like $P(I = i^1 \mid G = g^2, S = s^0)$ or in words "the probability of high intelligence given medium grade and low SAT?"

$$P(I = i^1 \mid G = g^2, S = s^0) = \frac{P(I = i^1, G = g^2, S = s^0)}{P(G = g^2, S = s^0)}$$

Approaches:

- **Exact inference by variable elimination**
  - Straightforward, but can be computationally prohibitive
- **Variational inference**
  - Approximate, biased
- **Sampling strategies**
  - Rejection sampling 
  - Importance weighted sampling
  - MCMC

Importance sampling

Fix any function $f : \mathcal{X} \to [0,1]$ and suppose we wish to estimate $\mu = \mathbb{E}_P[f(X)]$

If I draw $X_1, \ldots, X_M \sim P$ and define $\widehat{\mu}_P := \frac{1}{M} \sum_{i=1}^{M} f(X_i)$ then $\mathbb{E}_P\left[\widehat{\mu}_P\right] = \mathbb{E}_P[f(X)]$

Importance sampling

Fix any function $f : \mathcal{X} \to [0,1]$ and suppose we wish to estimate $\mu = \mathbb{E}_P[f(X)]$

If I draw $X_1, \ldots, X_M \sim P$ and define $\hat{\mu}_P := \frac{1}{M}\sum_{i=1}^{M} f(X_i)$ then $\mathbb{E}_P\left[\hat{\mu}_P\right] = \mathbb{E}_P[f(X)]$

Importance sampling with candidate distribution Q:

If I draw $Y_1, \ldots, Y_M \sim Q$ and define $\hat{\mu}_Q := \frac{1}{M}\sum_{i=1}^{M} f(Y_i)\frac{P(Y_i)}{Q(Y_i)}$ then $\mathbb{E}_Q\left[\hat{\mu}_Q\right] = \mathbb{E}_P[f(X)]$

$$\mathbb{E}_Q\left[\frac{1}{M}\sum_{i=1}^{M} f(Y_i)\frac{P(Y_i)}{Q(Y_i)}\right] =$$

Importance sampling

Fix any function $f : \mathcal{X} \to [0,1]$ and suppose we wish to estimate $\mu = \mathbb{E}_P[f(X)]$

If I draw $X_1, \ldots, X_M \sim P$ and define $\hat{\mu}_P := \frac{1}{M}\sum_{i=1}^{M} f(X_i)$ then $\mathbb{E}_P\left[\hat{\mu}_P\right] = \mathbb{E}_P[f(X)]$

Importance sampling with candidate distribution Q:

If I draw $Y_1, \ldots, Y_M \sim Q$ and define $\hat{\mu}_Q := \frac{1}{M}\sum_{i=1}^{M} f(Y_i)\frac{P(Y_i)}{Q(Y_i)}$ then $\mathbb{E}_Q\left[\hat{\mu}_Q\right] = \mathbb{E}_P[f(X)]$

$$\mathbb{E}_Q\left[\frac{1}{M}\sum_{i=1}^{M} f(Y_i)\frac{P(Y_i)}{Q(Y_i)}\right] = \mathbb{E}_Q\left[f(Y_1)\frac{P(Y_1)}{Q(Y_1)}\right]$$

$$= \sum_x Q(x)\cdot f(x)\frac{P(x)}{Q(x)}$$

$$= \sum_x P(x)\cdot f(x)$$

$$= \mathbb{E}_P[f(X)]$$

Importance sampling

Fix any function $f : \mathcal{X} \to [0,1]$ and suppose we wish to estimate $\mu = \mathbb{E}_P[f(X)]$

If I draw $X_1, \ldots, X_M \sim P$ and define $\hat{\mu}_P := \frac{1}{M}\sum_{i=1}^{M} f(X_i)$ then $\mathbb{E}_P\left[\hat{\mu}_P\right] = \mathbb{E}_P[f(X)]$

Importance sampling with candidate distribution Q:

If I draw $Y_1, \ldots, Y_M \sim Q$ and define $\hat{\mu}_Q := \frac{1}{M}\sum_{i=1}^{M} f(Y_i)\frac{P(Y_i)}{Q(Y_i)}$ then $\mathbb{E}_Q\left[\hat{\mu}_Q\right] = \mathbb{E}_P[f(X)]$

Moreover, as $M \to \infty$ we have $\frac{1}{M}\sum_{i=1}^{M} f(Y_i)\frac{P(Y_i)}{Q(Y_i)} \sim \mathcal{N}\left(\mathbb{E}_P[f(X)], \sigma_Q^2/M\right)$ where

$$\sigma_Q^2 = \mathbb{E}_Q[\left(f(Y)\frac{P(Y)}{Q(Y)}\right)^2] - \mathbb{E}_Q[f(Y)\frac{P(Y)}{Q(Y)}]^2$$

$$= \mathbb{E}_P[f(X)^2\frac{P(X)}{Q(X)}] - \mathbb{E}_P[f(X)]^2$$

σ_Q^2 is minimizes when

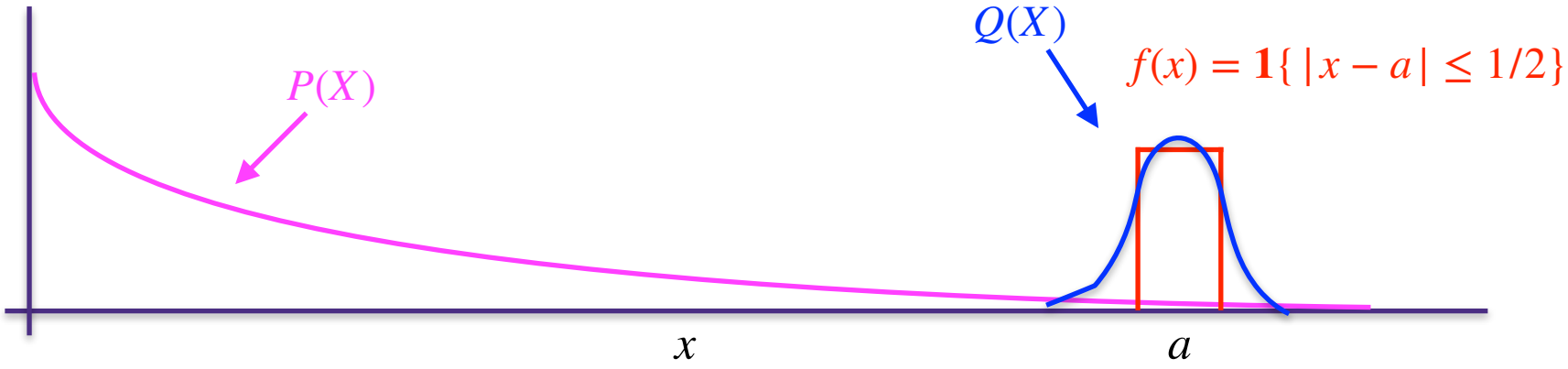
$Q(x) \propto |f(x)|P(x)$

Importance sampling

Importance sampling with candidate distribution Q:

If I draw $Y_1, \ldots, Y_M \sim Q$ and define $\widehat{\mu}_Q := \frac{1}{M}\sum_{i=1}^{M} f(Y_i)\frac{P(Y_i)}{Q(Y_i)}$ then $\mathbb{E}_Q\left[\widehat{\mu}_Q\right] = \mathbb{E}_P[f(X)]$

Example: rare event sampling

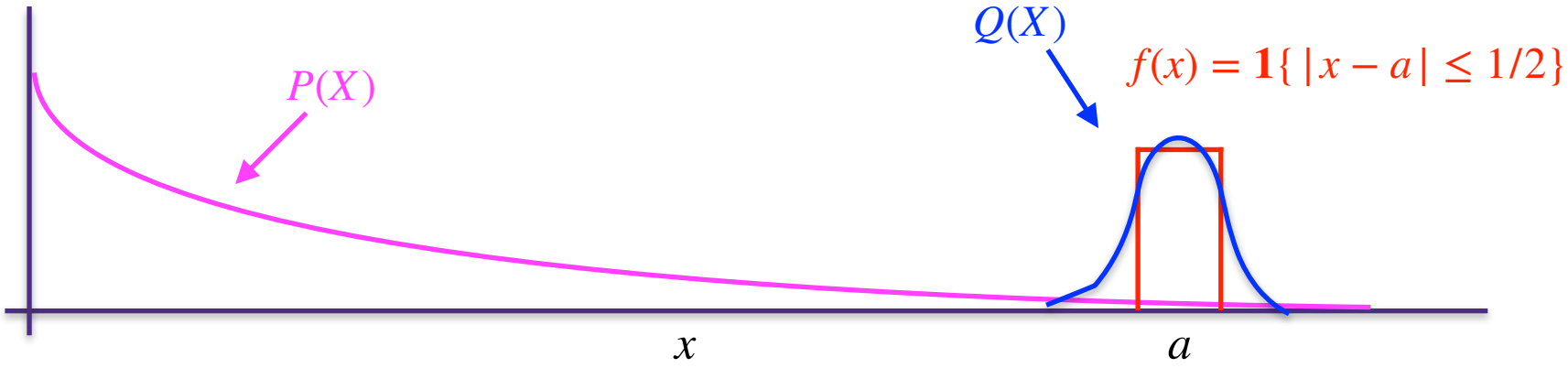


Importance sampling

Importance sampling with candidate distribution Q:

If I draw $Y_1, \ldots, Y_M \sim Q$ and define $\widehat{\mu}_Q := \frac{1}{M}\sum_{i=1}^{M} f(Y_i)\frac{P(Y_i)}{Q(Y_i)}$ then $\mathbb{E}_Q\left[\widehat{\mu}_Q\right] = \mathbb{E}_P[f(X)]$

Example: rare event sampling



$$\mathbb{E}_P[f(X)] = \int_x P(x)\mathbf{1}\{\,|x-a| \le 1/2\}dx \approx P(a)$$

$\sigma_Q^2 \approx P(a)^2 \implies$ sampling from Q requires $M \ge \epsilon^{-2}$ for relative error $\frac{\widehat{\mu}_Q - \mu}{\mu} \le \epsilon$

$\sigma_P^2 \approx P(a) \implies$ sampling from P requires $M \ge \frac{\epsilon^{-2}}{P(a)}$ for relative error $\frac{\widehat{\mu}_P - \mu}{\mu} \le \epsilon$

Self-normalized Importance sampling

Fix any function $f : \mathcal{X} \to [0,1]$ and suppose we wish to estimate $\mu = \mathbb{E}_P[f(X)]$

If I draw $X_1, \ldots, X_M \sim P$ and define $\widehat{\mu}_P := \frac{1}{M}\sum_{i=1}^{M} f(X_i)$ then $\mathbb{E}_P\left[\widehat{\mu}_P\right] = \mathbb{E}_P[f(X)]$

Importance sampling with candidate distribution Q:

If I draw $Y_1, \ldots, Y_M \sim Q$ and define $\widehat{\mu}_Q := \frac{1}{M}\sum_{i=1}^{M} f(Y_i)\frac{P(Y_i)}{Q(Y_i)}$ then $\mathbb{E}_Q\left[\widehat{\mu}_Q\right] = \mathbb{E}_P[f(X)]$

Self-normalized Importance sampling with candidate distribution Q:

If I draw $Y_1, \ldots, Y_M \sim Q$ and define $\widehat{\mu}_Q^{sn} := \frac{\sum_{i=1}^{M} f(Y_i)\frac{P(Y_i)}{Q(Y_i)}}{\sum_{i=1}^{M}\frac{P(Y_i)}{Q(Y_i)}}$ then $\mathbb{E}_Q\left[\widehat{\mu}_Q\right] = \mathbb{E}_P[f(X)]$

Biased $\mathbb{E}_Q[\widehat{\mu}_Q^{sn}] \neq \mu$ but is asymptotically consistent since $\mathbb{E}_Q[\frac{P(Y_i)}{Q(Y_i)}] = 1$

Self-normalized Importance sampling

Fix any function $f : \mathcal{X} \to [0,1]$ and suppose we wish to estimate $\mu = \mathbb{E}_P[f(X)]$

If I draw $X_1, \ldots, X_M \sim P$ and define $\hat{\mu}_P := \frac{1}{M}\sum_{i=1}^{M} f(X_i)$ then $\mathbb{E}_P\left[\hat{\mu}_P\right] = \mathbb{E}_P[f(X)]$

Importance sampling with candidate distribution Q:

If I draw $Y_1, \ldots, Y_M \sim Q$ and define $\hat{\mu}_Q := \frac{1}{M}\sum_{i=1}^{M} f(Y_i)\frac{P(Y_i)}{Q(Y_i)}$ then $\mathbb{E}_Q\left[\hat{\mu}_Q\right] = \mathbb{E}_P[f(X)]$

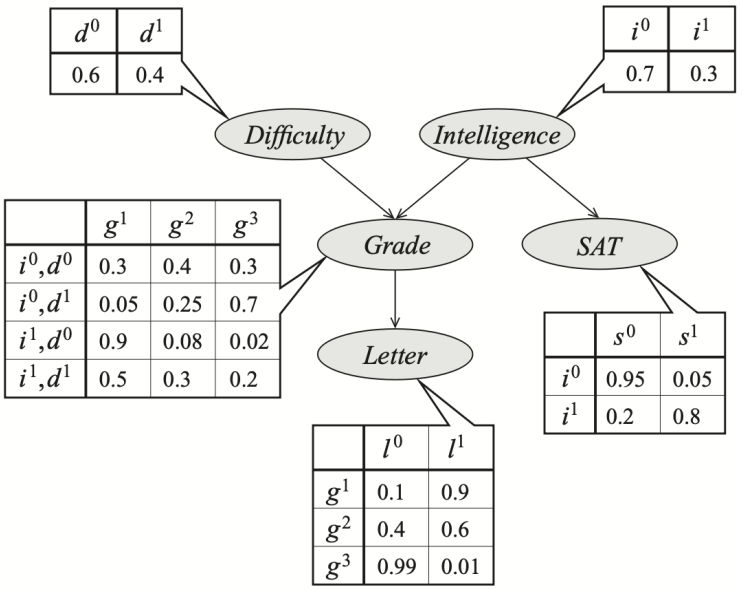
Self-normalized Importance sampling with candidate distribution Q:

If I draw $Y_1, \ldots, Y_M \sim Q$ and define $\hat{\mu}_Q^{sn} := \frac{\sum_{i=1}^{M} f(Y_i)\frac{P(Y_i)}{Q(Y_i)}}{\sum_{i=1}^{M}\frac{P(Y_i)}{Q(Y_i)}}$ then $\mathbb{E}_Q\left[\hat{\mu}_Q\right] = \mathbb{E}_P[f(X)]$

If we only know P up to a normalizing constant such that we have $\widetilde{P}(x) = ZP(x)$, we can use $\widetilde{P}$ to obtain the same estimator:

$$\hat{\mu}_Q^{sn} = \frac{\sum_{i=1}^{M} f(Y_i)\frac{\widetilde{P}(Y_i)}{Q(Y_i)}}{\sum_{i=1}^{M}\frac{\widetilde{P}(Y_i)}{Q(Y_i)}} = \frac{\sum_{i=1}^{M} f(Y_i)\frac{P(Y_i)}{Q(Y_i)}}{\sum_{i=1}^{M}\frac{P(Y_i)}{Q(Y_i)}}$$

Bayesian Network Inference with Importance Sampling



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

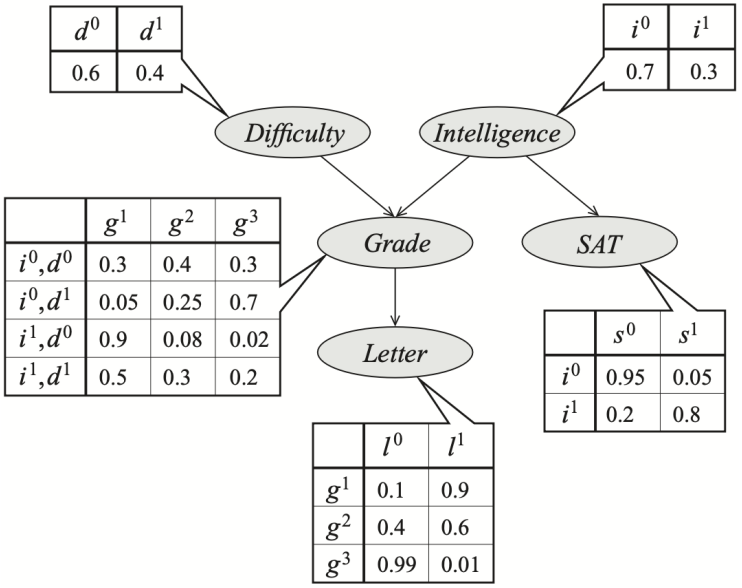
	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

$$P(X) := P(L, G, S, D, I)$$
$$= P(L \mid G)P(G \mid D, I)P(S \mid I)P(D)P(I)$$

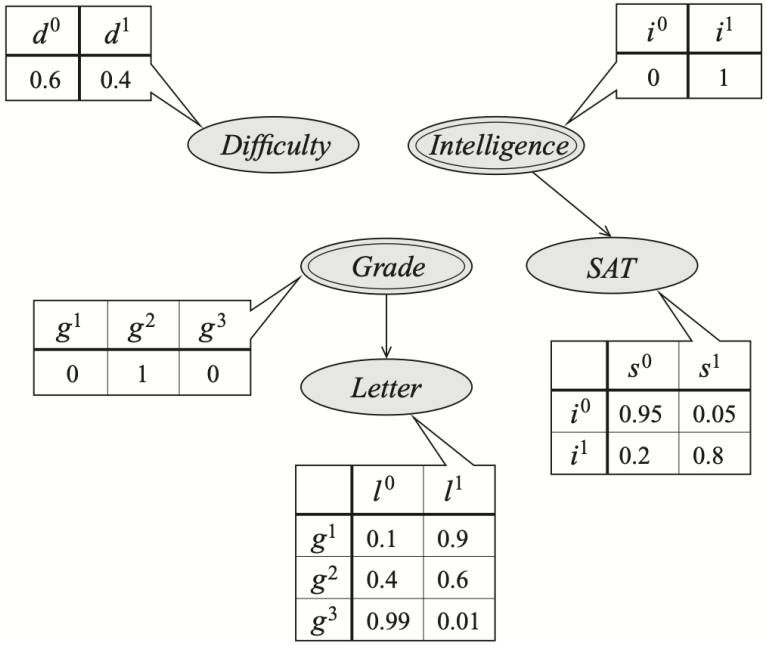
If $X := (L, G, S, D, I)$ and $P(X)$ denotes a Bayesian network then as we saw earlier, sampling from conditionals of P directly is awkward and inefficient. Can we define a convenient Q and use important sampling?

Bayesian Network Inference with Importance Sampling

Original network $\mathcal{B}$



"Mutilated" network $\mathcal{B}_{I=i^1,G=g^2}$



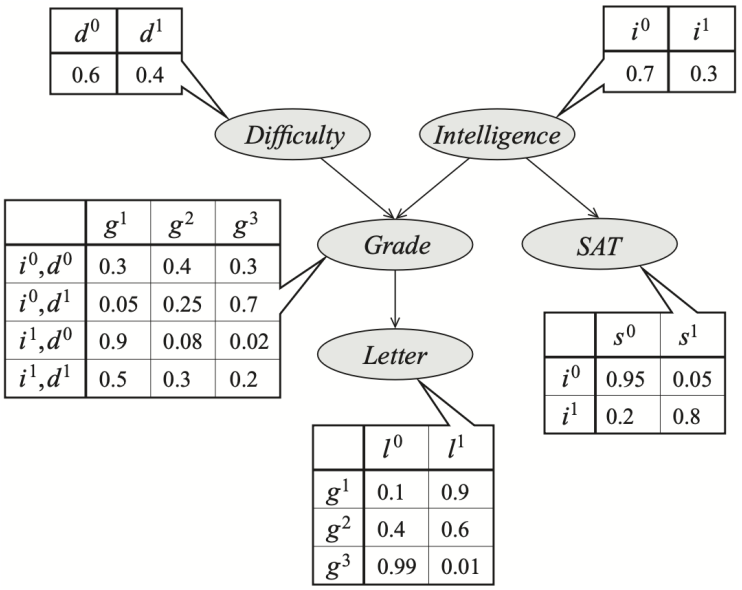
$$P(X) := P(L,G,S,D,I)$$
$$= P(L|G)P(G|D,I)P(S|I)P(D)P(I)$$

$$Q_{I=i^1,G=g^2}(Y) = P(L|G=g^2)P(S|I=i^1)P(D)$$

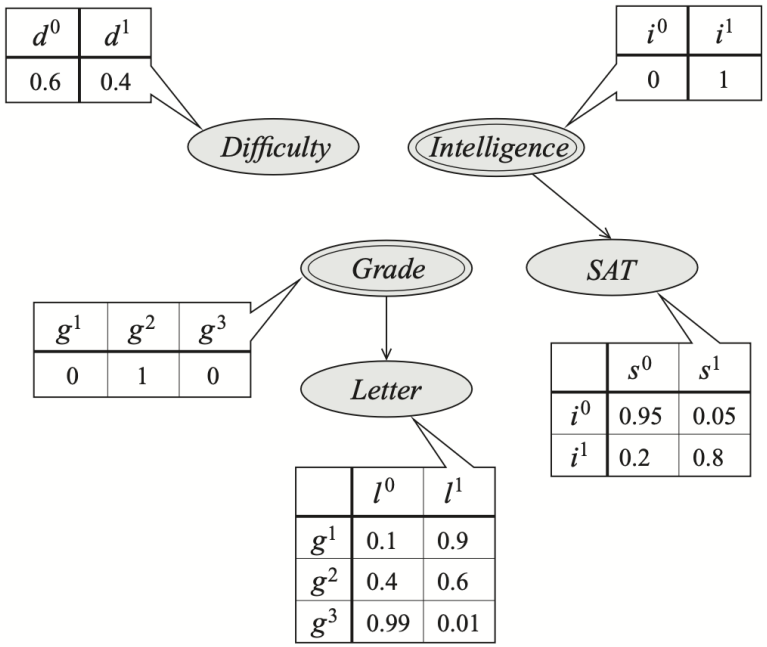
If $X := (L,G,S,D,I)$ and $P(X)$ denotes a Bayesian network then as we saw earlier, sampling from conditionals of P directly is awkward and inefficient. Can we define a convenient Q and use important sampling?

Bayesian Network Inference with Importance Sampling

Original network $\mathscr{B}$



"Mutilated" network $\mathscr{B}_{I=i^1,G=g^2}$



$$P(X) := P(L,G,S,D,I)$$
$$= P(L|G)P(G|D,I)P(S|I)P(D)P(I)$$

$$Q_{I=i^1,G=g^2}(Y) = P(L|G=g^2)P(S|I=i^1)P(D)$$

Note that $P(I=i^1, G=g^2) = \mathbb{E}_P[\mathbf{1}\{I=i^1, G=g^2\}]$

$$= \mathbb{E}_Q[\frac{1}{M}\sum_{k=1}^{M}\mathbf{1}\{I_k=i^1, G_k=g^2\}\frac{P(Y_k)}{Q(Y_k)}]$$

If $\{I=i^1, G=g^2\}$ is a rare event, this could require substantial fewer samples!

Bayesian Network Inference with Importance Sampling

How do we compute $\frac{P(Y)}{Q(Y)}$? Let $Y = (L = l, g = g^2, S = s, D = d, I = i^1)$

Then $$\frac{P(Y)}{Q(Y)} = \frac{P(L = l, g = g^2, S = s, D = d, I = i^1)}{Q_{I=i^1, G=g^2}(L = l, g = g^2, S = s, D = d, I = i^1)}$$

$$P(X) := P(L, G, S, D, I)$$
$$= P(L|G)P(G|D, I)P(S|I)P(D)P(I)$$

$$Q_{I=i^1, G=g^2}(Y) = P(L|G = g^2)P(S|I = i^1)P(D)$$

Note that $P(I = i^1, G = g^2) = \mathbb{E}_P[\mathbf{1}\{I = i^1, G = g^2\}]$

$$= \mathbb{E}_Q[\frac{1}{M}\sum_{k=1}^{M} \mathbf{1}\{I_k = i^1, G_k = g^2\}\frac{P(Y_k)}{Q(Y_k)}]$$

If $\{I = i^1, G = g^2\}$ is a rare event, this could require substantial fewer samples!

Bayesian Network Inference with Importance Sampling

How do we compute $\frac{P(Y)}{Q(Y)}$? Let $Y = (L = l, g = g^2, S = s, D = d, I = i^1)$

$$\text{Then } \frac{P(Y)}{Q(Y)} = \frac{P(L = l, g = g^2, S = s, D = d, I = i^1)}{Q_{I=i^1, G=g^2}(L = l, g = g^2, S = s, D = d, I = i^1)}$$

$$= \frac{P(G = g^2 \mid D = d, I = i^1)P(I = i^1)}{1}$$

Note: all but the "mutilated" terms cancel.

$$P(X) := P(L, G, S, D, I)$$
$$= P(L \mid G)P(G \mid D, I)P(S \mid I)P(D)P(I)$$

$$Q_{I=i^1, G=g^2}(Y) = P(L \mid G = g^2)P(S \mid I = i^1)P(D)$$

Note that $P(I = i^1, G = g^2) = \mathbb{E}_P[\mathbf{1}\{I = i^1, G = g^2\}]$

$$= \mathbb{E}_Q[\frac{1}{M}\sum_{k=1}^{M} \mathbf{1}\{I_k = i^1, G_k = g^2\}\frac{P(Y_k)}{Q(Y_k)}]$$

If $\{I = i^1, G = g^2\}$ is a rare event, this could require substantial fewer samples!

Bayesian Network Inference with Importance Sampling

Algorithm 12.2 Likelihood-weighted particle generation

Procedure LW-Sample (
  $\mathcal{B}$, // Bayesian network over $\mathcal{X}$
  $\boldsymbol{Z} = \boldsymbol{z}$ // Event in the network
)

1 Let $X_1, \ldots, X_n$ be a topological ordering of $\mathcal{X}$
2 $w \leftarrow 1$
3 **for** $i = 1, \ldots, n$
4 $\boldsymbol{u}_i \leftarrow \boldsymbol{x}\langle \mathrm{Pa}_{X_i} \rangle$ // Assignment to Pa_{X_i} in $x_1, \ldots, x_{i-1}$
5 **if** $X_i \notin \boldsymbol{Z}$ **then**
6 Sample x_i from $P(X_i \mid \boldsymbol{u}_i)$
7 **else**
8 $x_i \leftarrow \boldsymbol{z}\langle X_i \rangle$ // Assignment to X_i in $\boldsymbol{z}$
9 $w \leftarrow w \cdot P(x_i \mid \boldsymbol{u}_i)$ // Multiply weight by probability of desired value
10 **return** $(x_1, \ldots, x_n), w$

Theorem: If the above algorithm is run on the mutilated network wrt to a set of set variables **z** and $Q(Y)$ represents its probability, then $w = \dfrac{P(Y)}{Q(Y)}$.

Bayesian Network Inference with Importance Sampling

Algorithm to estimate $P(\chi(X) = z)$:

1. Execute algorithm on previous slide M times with event $\{\chi(X) = z\}$
2. Get $(Y_1, w_1), \ldots, (Y_M, w_M)$ back
3. Set $\widehat{P}(\chi(X) = z) = \frac{1}{M} \sum_{k=1}^{M} \mathbf{1}\{\chi(Y_k) = z\} w_k$

Takeaway: $\widehat{P}(\chi(X) = z)$ is an unbiased estimator of $P(\chi(X) = z)$.
By the properties of importance sampling estimators of the previous slides, if Q is close to P, the critical size of M may not even depend on $P(\chi(X) = z)$, even if it is very rare. But if Q is very far, this could be worse than rejection sampling.

Theorem: If the above algorithm is run on the mutilated network wrt to a set of set variables **z** and $Q(Y)$ represents its probability, then $w = \frac{P(Y)}{Q(Y)}$.

Bayesian Network Inference with Importance Sampling

How do we use estimators like $\widehat{P}(\chi(X) = z) = \frac{1}{M}\sum_{k=1}^{M} \mathbf{1}\{\chi(Y_k) = z\} w_k$ to compute conditional queries like $P(I = i^1 \mid G = g^2, S = s^0)$?

Ratio method

$$P(I = i^1 \mid G = g^2, S = s^0) = \frac{P(I = i^1, G = g^2, S = s^0)}{P(G = g^2, S = s^0)}$$

Use M samples to compute

$$\widehat{P}(I = i^1, G = g^2, S = s^0)$$

Use M' samples to compute

$$\widehat{P}(G = g^2, S = s^0)$$

and take the ratio!

Numerator and denominator unbiased.

Self-normalized method

Collect data $(Y_1, w_1), \ldots, (Y_M, w_M)$ using event $\{G = g^2, S = s^0\}$

Output

$$\frac{\sum_{k=1}^{M} \mathbf{1}\{I_k = i^1, G_k = g^2, S_k = s^0\} w_k}{\sum_{k=1}^{M} w_k}$$

Numerator and denominator unbiased:

$$\mathbb{E}_Q[w_k] = \mathbb{E}_Q[w_k \mathbf{1}\{G_k = g^2, S_k = s^0\}]$$
$$= P(G = g^2, S = s^0)$$

Inference in Bayesian Networks

Goal: compute queries like $P(I = i^1 \mid G = g^2, S = s^0)$ or in words "the probability of high intelligence given medium grade and low SAT?"

$$P(I = i^1 \mid G = g^2, S = s^0) = \frac{P(I = i^1, G = g^2, S = s^0)}{P(G = g^2, S = s^0)}$$

Approaches:

- **Exact inference by variable elimination**
  - Straightforward, but can be computationally prohibitive
- **Variational inference**
  - Approximate, biased
- **Sampling strategies**
  - Rejection sampling 
  - Importance weighted sampling 
  - MCMC

Markov Chain Monte Carlo (MCMC)

Goal: compute queries like $P(I = i^1 \mid G = g^2, S = s^0)$ or in words "the probability of high intelligence given medium grade and low SAT?"

$$P(I = i^1 \mid G = g^2, S = s^0) = \frac{P(I = i^1, G = g^2, S = s^0)}{P(G = g^2, S = s^0)}$$

Sampling from the "mutilated" network will give you sample such that $\{G = g^2, s = s^0\}$. The problem is that this sample is drawn from $Q_{I=i^1, G=g^2}(Y)$ and not $P(I = i^1 \mid G = g^2, S = s^0)$. Importance sampling tries to "correct" the misalignment by weighting the sample appropriately.

MCMC starts with the candidate distribution $\pi^{(0)} := Q$ and then *evolves* the distribution $\pi^{(t)}$ slowly over time until it converges $\pi^{(t)} \to \pi_* := P$. Sampling from this distribution results in a sample from P

Gibbs sampling

Algorithm 12.4 Generating a Gibbs chain trajectory

Procedure Gibbs-Sample (
  $\boldsymbol{X}$ // Set of variables to be sampled
  Φ // Set of factors defining P_Φ
  $P^{(0)}(\boldsymbol{X})$, // Initial state distribution
  T // Number of time steps
)

1 Sample $\boldsymbol{x}^{(0)}$ from $P^{(0)}(\boldsymbol{X})$
2 **for** $t = 1, \ldots, T$
3 $\boldsymbol{x}^{(t)} \leftarrow \boldsymbol{x}^{(t-1)}$
4 **for each** $X_i \in \boldsymbol{X}$
5 Sample $x_i^{(t)}$ from $P_\Phi(X_i \mid \boldsymbol{x}_{-i})$
6 // Change X_i in $\boldsymbol{x}^{(t)}$
7 **return** $\boldsymbol{x}^{(0)}, \ldots, \boldsymbol{x}^{(T)}$

$$P(X_i \mid x_{-i}) = P(X_i \mid \text{MarkovBlanket}(i))$$
$$= P(X_i \mid \text{parents}(i)) \prod_{j \in \text{child}(i)} P(x_j \mid \text{parents}(j))$$

Markov blanket of *Cloudy* is
Sprinkler and *Rain*
Markov blanket of *Rain* is
Cloudy, *Sprinkler*, and *WetGrass*

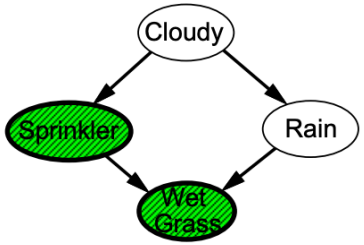


Figure credit: Pedro Domingos

Gibbs sampling

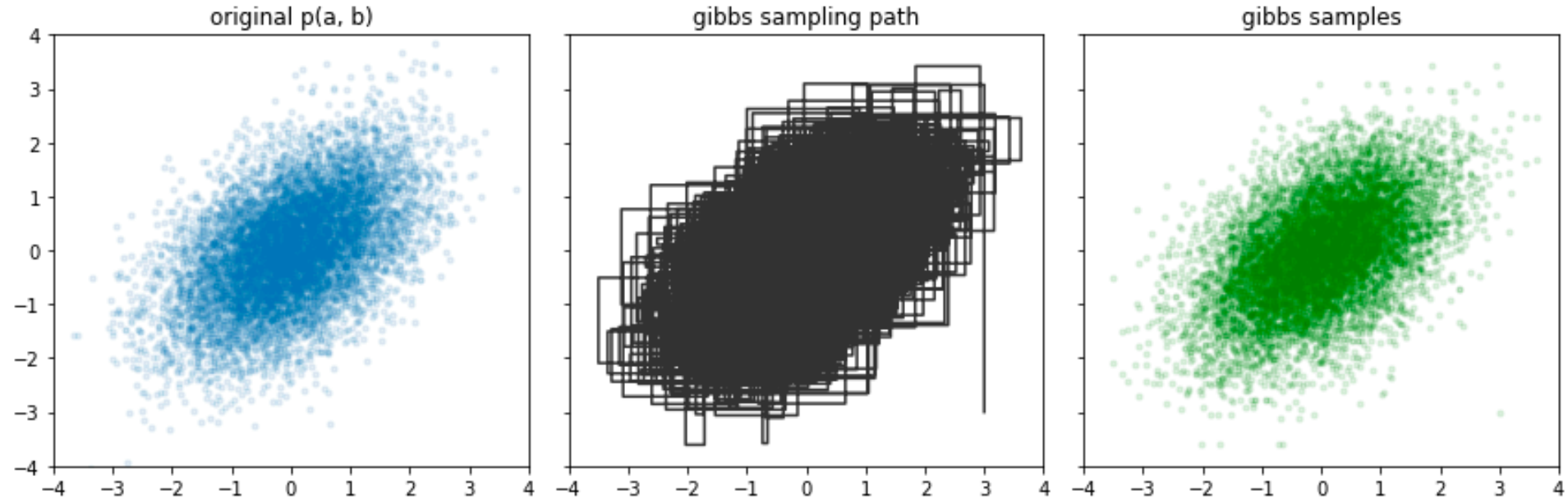


Figure credit: https://jessicastringham.net/2018/05/09/gibbs-sampling/

Markov Chain Monte Carlo (MCMC)

Gibb's sampling is just one example of a MCMC algorithm.

In general, the approach:

- Construct a Markov chain transition kernel $q : \mathcal{X} \to \mathcal{X}$ whose stationary distribution is equal to P
- Start with a realization $x^{(0)} \sim \pi^{(0)}$ drawn from an arbitrary starting distribution $\pi^{(0)}$
- Run Markov chain to evolve $x^{(t)} \mapsto x^{(t+1)}$ (equiv. $\pi^{(t)} \to \pi^{(t+1)}$) and return sample once convergence $\pi^{(t)} \approx P$

Easy to see Gives sampling defines a Markov chain with stationary distribution P.
Is it guaranteed to converge? How fast does it converge? What are other chains?