Approximate Inference 2: Importance Sampling

(Unnormalized) Importance Sampling

(Unnormalized) Importance Sampling

- Likelihood weighting is a special case of a general approach called importance sampling
- Let ***X*** be a set of variables that takes on values in some space *Val(**X**)*
- Importance sampling is a way to estimate $E_{P(x)}[f(x)]$ ie. the expectation of a function *f(**x**)* relative to some distribution *P(**X**)*, typically called the target distribution

(Unnormalized) Importance Sampling

- Generate samples ***x****[1]*, …, ***x****[M]* from P
- Then estimate:

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

(Unnormalized) Importance Sampling

- Sometimes you might want to generate samples from a different distribution Q (called a proposal distribution or sampling distribution)
- Why?
  - Might be impossible or computationally expensive to sample from P
- Proposal distribution can be arbitrary
  - Require that $Q(\boldsymbol{x}) > 0$ whenever $P(\boldsymbol{x}) > 0$
  - But computational performance of importance sampling depends strongly on how similar Q is to P

(Unnormalized) Importance Sampling

How to use the proposal distribution:

$$\boldsymbol{E}_{Q(\boldsymbol{X})}\left[f(\boldsymbol{X})\frac{P(\boldsymbol{X})}{Q(\boldsymbol{X})}\right] = \sum_{\boldsymbol{x}} Q(\boldsymbol{x}) f(\boldsymbol{x})\frac{P(\boldsymbol{x})}{Q(\boldsymbol{x})}$$

$$= \sum_{\boldsymbol{x}} f(\boldsymbol{x})P(\boldsymbol{x}) = \boldsymbol{E}_{P(\boldsymbol{X})}[f(\boldsymbol{X})]$$

Generate a set of samples $\mathcal{D} = \{\mathbf{x}[1], \ldots, \mathbf{x}[M]\}$ from Q then estimate:

$$\hat{\boldsymbol{E}}_{D}(f) = \frac{1}{M}\sum_{m=1}^{M} f(\boldsymbol{x}[m])\frac{P(\boldsymbol{x}[m])}{Q(\boldsymbol{x}[m])}$$

Unnormalized importance sampling estimator

(Unnormalized) Importance Sampling

This estimator is unbiased:

$$\boldsymbol{E}_D[\hat{\boldsymbol{E}}_D(f)] = \boldsymbol{E}_{Q(\boldsymbol{X})}[f(\boldsymbol{X})\frac{P(\boldsymbol{X})}{Q(\boldsymbol{X})}]$$

$$= \boldsymbol{E}_{Q(\boldsymbol{X})}[f(\boldsymbol{X})w(\boldsymbol{X})] = E_{P(\boldsymbol{X})}[f(\boldsymbol{X})]$$

Normalized Importance Sampling

Normalized Importance Sampling

- Frequently, P is known only up to a normalizing constant Z ie.

$$\tilde{P}(\boldsymbol{X}) = ZP(\boldsymbol{X})$$

- Happens when:
  - We know *P(X,**e**)* but need *P(X|**e**)*
  - We have the unnormalized product of clique potentials for a Markov network

Normalized Importance Sampling

- Define

$$w(\boldsymbol{X}) = \frac{\tilde{P}(\boldsymbol{X})}{Q(\boldsymbol{X})}$$

- The expected value of the *w(**X**)* under *Q(**X**)* is

$$E_{Q(X)}[w(X)] = \sum_x Q(x)\frac{\tilde{P}(x)}{Q(x)} = \sum_x \tilde{P}(x) = Z$$

Normalized Importance Sampling

$$E_{P(X)}[f(\boldsymbol{X})] = \sum_{\boldsymbol{x}} P(\boldsymbol{x}) f(\boldsymbol{x})$$

$$= \sum_{\boldsymbol{x}} Q(\boldsymbol{x}) f(\boldsymbol{x}) \frac{P(\boldsymbol{x})}{Q(\boldsymbol{x})}$$

$$= \frac{1}{Z} \sum_{\boldsymbol{x}} Q(\boldsymbol{x}) f(\boldsymbol{x}) \frac{\tilde{P}(\boldsymbol{x})}{Q(\boldsymbol{x})} \quad (\text{since } \mathrm{P(x)} = \frac{1}{\mathrm{Z}} \tilde{P}(x))$$

$$= \frac{1}{Z} \boldsymbol{E}_{Q(\boldsymbol{X})}[f(\boldsymbol{X}) w(\boldsymbol{X})]$$

$$= \frac{\boldsymbol{E}_{Q(\boldsymbol{X})}[f(\boldsymbol{X}) w(\boldsymbol{X})]}{\boldsymbol{E}_{Q(\boldsymbol{X})}[w(\boldsymbol{X})]}$$

Normalized Importance Sampling

With M samples $\mathcal{D} = \{\mathbf{x}[1], \ldots, \mathbf{x}[M]\}$ from Q, we can estimate:

$$\hat{\boldsymbol{E}}_{D}(f) = \frac{\sum_{m=1}^{M} f(\boldsymbol{x}[m]) w(\boldsymbol{x}[m])}{\sum_{m=1}^{M} w(\boldsymbol{x}[m])}$$

This is called the normalized importance sampling estimator or weighted importance sampling estimator

Normalized Importance Sampling

- Normalized importance sampling estimator is biased
- But has smaller variance than the unnormalized estimator
- Normalized estimator often used instead of unnormalized estimator, even when P is known and can be sampled from effectively

Importance Sampling for Bayesian Networks

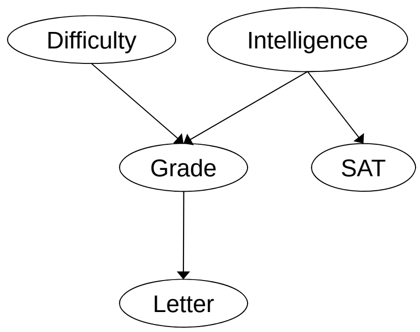
Importance Sampling for Bayesian Networks

Student Example

I	P(I)
low	0.7
high	0.3

D	P(D)
low	0.6
high	0.4

D	I	G	P(G|D,I)
low	low	C	0.3
low	low	B	0.4
low	low	A	0.3
low	high	C	0.02
low	high	B	0.08
low	high	A	0.9
high	low	C	0.7
high	low	B	0.25
high	low	A	0.05
high	high	C	0.2
high	high	B	0.3
high	high	A	0.5



I	S	P(S|I)
low	low	0.95
low	high	0.05
high	low	0.2
high	high	0.8

G	L	P(L|G)
C	weak	0.99
C	strong	0.01
B	weak	0.4
B	strong	0.6
A	weak	0.1
A	strong	0.9

Importance Sampling for Bayesian Networks

- What proposal distribution do we use?
- Suppose we want an event *Grade=B* either as a query or as evidence
  - Easy to sample *P(Letter* | *Grade = B)*
  - Difficult to account for *Grade=B*'s influence on *Difficulty*, *Intelligence* and *SAT*
- In general:
  - Want to account for effect of the event on the descendants
  - But avoid accounting for its effects on the nondescendants

Importance Sampling for Bayesian Networks

- Let B be a network, and $Z_1 = z_1, \ldots, Z_k = z_k$, abbreviated $\boldsymbol{Z}=\boldsymbol{z}$, an instantiation of variables.
- We define the mutilated network $B_{\boldsymbol{Z}=\boldsymbol{z}}$ as follows:
  - Each node $Z_i \in \boldsymbol{Z}$ has no parents in $B_{\boldsymbol{Z}=\boldsymbol{z}}$
  - The CPD of Z_i in $B_{\boldsymbol{Z}=\boldsymbol{z}}$ gives probability 1 to $Z_i = z_i$ and probability 0 to all other values $z_i' \in Val(Z_i)$
  - The parents and CPDs of all other nodes $X \notin \boldsymbol{Z}$ are unchanged

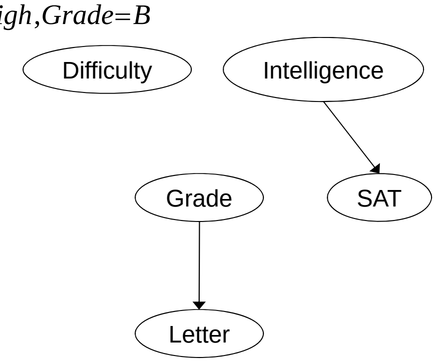
Importance Sampling for Bayesian Networks

Mutilated Network:

$$\beta^{student}_{Intelligence=high,Grade=B}$$

D	P(D)
low	0.6
high	0.4

G	P(G|D,I)
C	0
B	1
A	0



I	P(I)
low	0
high	1

I	S	P(S|I)
low	low	0.95
low	high	0.05
high	low	0.2
high	high	0.8

G	L	P(L|G)
C	weak	0.99
C	strong	0.01
B	weak	0.4
B	strong	0.6
A	weak	0.1
A	strong	0.9

Importance Sampling for Bayesian Networks

- Proposition 12.2: Let ξ be a sample generated by Likelihood Weighting and w be its weight. Then the distribution over ξ is as defined by the network $B_{\mathbf{Z=z}}$, and

$$w(\xi) = \frac{P_B(\xi)}{P_{B_{Z=z}}(\xi)}$$

- (Informally) Importance sampling using a mutilated network as a proposal distribution is equivalent to Likelihood Weighting with $P_B(X,\mathbf{z})$ and proposal distribution Q induced by the mutilated network $B_{\mathbf{E=e}}$.

Likelihood Weighting Revisited

Likelihood Weighting Revisited

Two versions of likelihood weighting

1. Ratio Likelihood Weighting
2. Normalized Likelihood Weighting

Likelihood Weighting Revisited

Ratio Likelihood Weighting

$$P(\boldsymbol{y} \mid \boldsymbol{e}) = \frac{P(\boldsymbol{y}, \boldsymbol{e})}{P(\boldsymbol{e})}$$

Use unnormalized importance sampling:

1. For numerator – use LW to generate M samples with ***Y***=***y***, ***E***=***e*** as the event
2. For denominator – use LW to generate M' samples with ***E***=***e*** *as the event*

$$\hat{P}_D(\boldsymbol{y} \mid \boldsymbol{e}) = \frac{\hat{P}_D(\boldsymbol{y}, \boldsymbol{e})}{\hat{P}_{D'}(\boldsymbol{e})} = \frac{\frac{1}{M}\sum_{m=1}^{M} w[m]}{\frac{1}{M'}\sum_{m=1}^{M'} w'[m]}$$

Likelihood Weighting Revisited

Normalized Likelihood Weighting

- Ratio Likelihood Weighting estimates a single query *P(**y**|**e**)* from a set of samples (ie. it sets ***Y***=***y*** when sampling)
- Sometimes we want to evaluate a set of queries *P(**y**|**e**)*
- Use normalized likelihood weighting with

  $$\tilde{P}(\mathcal{X}) = P_B(\mathcal{X}, \boldsymbol{e})$$

- Estimate the expectation of a function f:

  $$f(\xi) = \boldsymbol{I}\{\xi\langle \boldsymbol{Y} \rangle = \boldsymbol{y}\}$$

Likelihood Weighting Revisited

Quality of importance sampling depends on how close the proposal distribution Q is to the target distribution P.

Consider the two extremes:

1. All evidence at the roots:
   - Proposal distribution is the posterior
   - Evidence affects samples all along the way and all samples have the same weight P(e)

Likelihood Weighting Revisited

2. All evidence at the leaves:
   - Proposal distribution is the prior distribution $P_{\mathcal{B}}(X)$
   - Evidence doesn't affect samples, weights have to compensate. LW will only work well if prior is similar to the posterior

Likelihood Weighting Revisited

$$\underbrace{P(X)}_{\text{Prior}} = \sum_e P(e)\underbrace{P(X \mid e)}_{\text{Posterior}}$$

- If *P(**e**)* is high, then the posterior *P(X|**e**)* plays a large role and is close to the prior *P(X)*
- If *P(**e**)* is low, then the posterior *P(X|**e**)* plays a small role and the prior *P(X)* will likely look very different

Likelihood Weighting Revisited

Summary

Ratio Likelihood Weighting

- Computes P(***y***|***e***) for a specific ***y*** (ie. values for ***y*** are set)
- Uses unnormalized importance sampling for both numerator and denominator in P(**y**,**e**)/P(**e**)
- Needs a new set of samples for each query **y**
- Lower variance in estimator
- Can bound # of samples required for a good estimate (but under strong conditions)

Likelihood Weighting Revisited

Summary

Normalized Likelihood Weighting

- Samples an assignment for ***Y***, which introduces additional variance
- Allows multiple queries ***y*** using the same set of samples (conditioned on evidence ***e***)

Likelihood Weighting Revisted

Problems with Likelihood Weighting:

- If there are a lot of evidence variables $P(\boldsymbol{Y} \mid \boldsymbol{E_1} = \boldsymbol{e_1}, \ldots, \boldsymbol{E_k} = \boldsymbol{e_k})$:
  – Many samples will have ε weight
  – Weighted estimate dominated by a small fraction of samples that have $> \varepsilon$
- If evidence variables occur in the leaves, the samples drawn will not be affected much by the evidence