Approximate Inference Via Sampling, particle filter
Hidden Markov Model (HMM)
Hidden Markov Model (HMM)

Example:
Example:

But you don’t observe location (dark room)
Example:

But you don’t observe location (dark room)

You hear how close the bot is!
Example:

But you don’t observe location (dark room)

You hear how close the bot is!

What you hear: + noise
Example:
But you don’t observe location (dark room)
You hear how close the bot is!

What you hear: + noise

Can you catch the Bot? In time?
**Hidden Markov Model (HMM)**

\[ S_1 \rightarrow S_2 \rightarrow S_3 \]

\[ X_1 \rightarrow X_2 \rightarrow X_3 \]

- \( X_t \)'s are what you hear (observation)
- \( S_t \)'s are the unseen locations (states)

Eg: for \( m \times m \) grid we have, \( K = m^2 \) states

Number of alphabets = \# colors you can observe
Eg: for m x m grid we have, K = m^2 states
HIDDEN MARKOV MODEL (HMM)

Eg: for m x m grid we have, $K = m^2$ states

Transition matrix is $K \times K$ (too large)
Eg: for $m \times m$ grid we have, $K = m^2$ states

Transition matrix is $K \times K$ (too large)

Use sampling to do approximate inference
Number of samples $n << m^4$
Inference Question

• Can we compute (efficiently and approximately)

\[ P(S_t|x_1, \ldots, x_{t-1}) \]

• We can’t afford too much time to compute since we need to move the bot in time

• We can perform inference via sampling
Who is more likely to win the game?
Who is more likely to win the game?
Compute sum of exact probabilities of all possible sequence of moves leading to Player 1’s victory
Who is more likely to win the game?

Throw dice and simulate multiple games, see who wins more often
• Draw $n$ samples from the sampling distribution
Inference Via Sampling

- Draw $n$ samples from the sampling distribution
- Compute approximate probabilities by computing empirical frequencies
Inference Via Sampling

- Draw \( n \) samples from the sampling distribution
- Compute approximate probabilities by computing empirical frequencies
- Why sampling?
• Draw n samples from the sampling distribution

• Compute approximate probabilities by computing empirical frequencies

• Why sampling?

  • Getting multiple samples often faster than computing exact probabilities
Inference Via Sampling

- Draw \( n \) samples from the sampling distribution
- Compute approximate probabilities by computing empirical frequencies
- Why sampling?
  - Getting multiple samples often faster than computing exact probabilities
  - Inference is key step in learning
Inference Via Sampling

- Law of large numbers: empirical distribution using large samples approximates the true distribution

- Some approaches:
  - Rejection sampling: sample all the variables, retain only ones that match evidence
  - Importance sampling: Sample from a different distribution but then apply correction while computing empirical marginals
  - Gibbs sampling: iteratively sample from distributions closer and closer to the true one
• Getting a sample from HMM given parameters is easy!

• Its accounting for observations (conditionally sampling given observations) that is hard!

• Can we use the fact that sampling from HMM is easy inference?
HIDDEN MARKOV MODEL (HMM)
HIDDEN MARKOV MODEL (HMM)
HIDDEN MARKOV MODEL (HMM)
Hidden Markov Model (HMM)
Hidden Markov Model (HMM)
Hidden Markov Model (HMM)
Hidden Markov Model (HMM)
HIDDEN MARKOV MODEL (HMM)
Hidden Markov Model (HMM)
Hidden Markov Model (HMM)
Hidden Markov Model (HMM)
HIDDEN MARKOV MODEL (HMM)
Hidden Markov Model (HMM)
Hidden Markov Model (HMM)
HIDDEN MARKOV MODEL (HMM)

Eg: say observations were
**Hidden Markov Model (HMM)**

Eg: say observations were

Rejection sampling: Reject samples that don’t match observations
Eg: say observations were

Rejection sampling: Reject samples that don’t match observations

We can do this sequentially!
Eg: say observations were
Eg: say observations were
Eg: say observations were
Eg: say observations were
Eg: say observations were
Eg: say observations were
Hidden Markov Model (HMM)

Eg: say observations were
Hidden Markov Model (HMM)

Eg: say observations were
Eg: say observations were
Eg: say observations were
Eg: say observations were

Hidden Markov Model (HMM)
Hidden Markov Model (HMM)

Eg: say observations were multiple samples simultaneously.
Eg: say observations were

Multiple samples simultaneously.

Problem: Most samples rejected
We really want to draw from distribution $P$.

But we can only draw from distribution $Q$ easily

Trick:

1. Draw $x_1, \ldots, x_n \sim Q$
2. Re-weight each sample $x_t$ by $P(X = x_t) / Q(X = x_t)$
Why does it work?

$$\mathbb{E}_{X \sim p}[f(X)] = \sum_x P(X = x)f(x)$$
Why does it work?

\[ \mathbb{E}_{X \sim P}[f(X)] = \sum_x P(X = x)f(x) \]

\[ = \sum_x Q(X = x) \left( \frac{P(X = x)}{Q(X = x)} f(x) \right) \]
**Importance Sampling**

**Why does it work?**

\[
\mathbb{E}_{X \sim P}[f(X)] = \sum_x P(X = x)f(x)
\]

\[
= \sum_x Q(X = x) \left( \frac{P(X = x)}{Q(X = x)} f(x) \right)
\]

\[
= \mathbb{E}_{X \sim Q} \left[ \frac{P(X)}{Q(X)} f(X) \right]
\]

Example:

\[f(X) = 1_{\{X \in \text{Set}\}}\]

Then

\[\mathbb{E}_{X \sim P}[f(X)] = P(X \in \text{Set}) \approx \frac{1}{n} \sum_{t=1}^{n} P(X = x_t) \frac{Q(X = x_t)}{Q(X = x_t)} f(x_t)\]
**Importance Sampling**

- Why does it work?

\[
\mathbb{E}_{X \sim P}[f(X)] = \sum_x P(X = x) f(x) \\
= \sum_x Q(X = x) \left( \frac{P(X = x)}{Q(X = x)} f(x) \right) \\
= \mathbb{E}_{X \sim Q} \left[ \frac{P(X)}{Q(X)} f(X) \right] \\
\approx \frac{1}{n} \sum_{t=1}^{n} \frac{P(X = x_t)}{Q(X = x_t)} f(x_t)
\]

Example:

\[ f(X) = \begin{cases} 1 & \text{if } X \in \text{Set} \\ 0 & \text{otherwise} \end{cases}, \text{ then } \mathbb{E}_{X \sim P}[f(X)] = P(X \in \text{Set} \)
Why does it work?

\[
\mathbb{E}_{X \sim P}[f(X)] = \sum_x P(X = x)f(x)
\]

\[
= \sum_x Q(X = x) \left( \frac{P(X = x)}{Q(X = x)} f(x) \right)
\]

\[
= \mathbb{E}_{X \sim Q} \left[ \frac{P(X)}{Q(X)} f(X) \right]
\]

\[
\approx \frac{1}{n} \sum_{t=1}^{n} \frac{P(X = x_t)}{Q(X = x_t)} f(x_t)
\]

Example: \( f(X) = 1\{X \in \text{Set}\} \), then \( \mathbb{E}_{X \sim P}[f(X)] = P(X \in \text{Set}) \)
Why does it work?

\[
\mathbb{E}_{X \sim P}[f(X)] = \sum_x P(X = x)f(x)
\]

\[
= \sum_x Q(X = x) \left( \frac{P(X = x)}{Q(X = x)}f(x) \right)
\]

\[
= \mathbb{E}_{X \sim Q} \left[ \frac{P(X)}{Q(X)}f(X) \right]
\]

\[
\approx \frac{1}{n} \sum_{t=1}^n \frac{P(X = x_t)}{Q(X = x_t)}f(x_t)
\]

Example: \(f(X) = 1\{X \in \text{Set}\}\), then \(\mathbb{E}_{X \sim P}[f(X)] = P(X \in \text{Set})\)

Hence, using importance weighted sampling,

\[
P(X \in \text{Set}) \approx \frac{1}{n} \sum_{t=1}^n 1\{x_t \in \text{Set}\} \frac{P(X=x_t)}{Q(X=x_t)}
\]
Importance Sampling

Why does it work?

\[ E \sim P \left[ f(X) \right] = \sum_{x} x P(X = x) f(x) = \sum_{x} x Q(X = x) f(x) P(X = x) Q(X = x) f(x) \approx \frac{1}{n} \sum_{t=1}^{n} P(X = x_t) Q(X = x_t) f(x_t) \]

Example:

\[ f(X) = 1 \{ X \in \text{Set} \} \]

Then

\[ E \sim P \left[ f(X) \right] = P(X \in \text{Set}) \approx \frac{1}{n} \sum_{t=1}^{n} P(X = x_t) f(x_t) \]
**Importance Sampling**

Why does it work?

\[
E_X \sim P[f(X)] = \mathbb{E}_X \sim P[X = x] f(x) = \mathbb{E}_X \sim Q[f(X)]
\]

\[
E_X \sim Q \approx \frac{1}{n} \sum_{t=1}^{n} P(X = x_t) Q(X = x_t) f(x_t)
\]

Example:

\[
f(X) = 1 \{X \in \text{Set}\}, \text{ then } E_X \sim P[f(X)] = P(X \in \text{Set})
\]

Hence, using importance weighted sampling,

\[
P(X \in \text{Set}) \approx \frac{1}{n} \sum_{t=1}^{n} P(X = x_t)
\]

\[
Q(X = x_t)
\]

\[
f(x_t)
\]
Importance Sampling

$P(1) = 0.9, \quad \forall j \neq 1 \ P(j) = 0.1/5$
**Importance Sampling**

\[ P(1) = 0.9, \quad \forall j \neq 1 \ P(j) = \frac{0.1}{5} \]
**Importance Sampling**

Why does it work?

\[ \mathbb{E}_X \approx \mathbb{E}_X \sim Q \mathbb{E}_X \sim P \]

Example:

\[ f(X) = 1 \{ X \in \text{Set} \}, \text{then} \]

\[ \mathbb{E}_X \sim P [f(X)] = P(X \in \text{Set}) \]

\[ \approx \frac{1}{n} \sum_{t=1}^{n} P(X_t) = 0.1/5 \]

\[ \forall j \neq 1 \quad P(j) = 0.1/5 \]

\[ \forall j \quad Q(j) = 1/6 \]
**Importance Sampling**

**Why does it work?**

\[
X \sim P[f(X)] = \int x \, P(X=x) \, f(x) = \int x \, Q(X=x) \, f(x)
\]

\[
E_X \sim Q[P(X) \approx \frac{1}{n} \sum_{t=1}^{n} P(X=x_t)]
\]

**Example:**

\[f(X) = \begin{cases} 1 & X \in \text{Set} \\ 0 & \text{otherwise} \end{cases}, \text{ then } E_X \sim P[f(X)] = P(X \in \text{Set}) \]

\[
P(1) = 0.9, \quad \forall j \neq 1 \quad P(j) = 0.1/5 \quad \forall j \quad Q(j) = 1/6
\]

Set = \{2, 4, 6\}
**Importance Sampling**

Why does it work?

\[
X \sim P \left[ f(X) \right] = \int_X f(x) P(X = x) dx = \int_X f(x) Q(X = x) dx \approx \frac{1}{n} \sum_{t=1}^{n} P(X = x_t) Q(X = x_t) f(x_t)
\]

Example:

\[f(X) = \{X \in \text{Set}\}, \text{then} \]

\[E_X \sim P \left[ f(X) \right] = P(X \in \text{Set})\]

Hence, using importance weighted sampling,

\[P(X \in \text{Set}) \approx \frac{1}{n} \sum_{t=1}^{n} P(X = x_t) Q(X = x_t) f(x_t)\]

Set = \{2, 4, 6\}

What is \(P(\text{Set})\)?

\[P(1) = 0.9, \quad \forall j \neq 1 P(j) = 0.1/5\]

\[\forall j \quad Q(j) = 1/6\]
**Importance Sampling**

**Why does it work?**

$$P(x) \approx \frac{1}{n} \sum_{t=1}^{n} \mathbb{1}_{\{x_t \in \text{Set}\}} \frac{P(x_t)}{Q(x_t)} = \frac{1}{n} \sum_{t=1}^{n} \mathbb{1}_{\{x_t \in \text{Set}\}} \frac{0.1/5}{1/6}$$

**Example:**

$$f(x) = \begin{cases} 1 & x \in \text{Set} \\ 0 & \text{otherwise} \end{cases}$$

Then

$$E_X \sim P \left[ f(X) \right] = \frac{1}{n} \sum_{t=1}^{n} \mathbb{1}_{\{x_t \in \text{Set}\}}$$

Hence, using importance weighted sampling,

$$P(x \in \text{Set}) \approx \frac{1}{n} \sum_{t=1}^{n} \mathbb{1}_{\{x_t \in \text{Set}\}}$$

**Set = \{2, 4, 6\}**

**What is P(\text{Set})?**

\[ P(1) = 0.9, \quad \forall j \neq 1 \quad P(j) = 0.1/5 \]

\[ \forall j \quad Q(j) = 1/6 \]
Importance Sampling

Why does it work?

\[
E_X \sim P[f(X)] = \sum_{x} P(X = x) f(x)
\]

\[
= \sum_{x} Q(X = x) \frac{P(X = x) f(x)}{Q(X = x)}
\]

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

\[
\approx \frac{1}{n} \sum_{t=1}^{n} 1\{x_t \in \text{Set}\} \frac{P(x_t)}{Q(x_t)} = \frac{1}{n} \sum_{t=1}^{n} 1\{x_t \in \text{Set}\} \frac{0.1/5}{1/6}
\]

Set = \{2, 4, 6\}

What is P(Set)?

\[
= 0.12 \times \frac{1}{n} \sum_{t=1}^{n} 1\{x_t \in \text{Set}\} \approx 0.12 \times 0.5 = 0.06
\]

\[
P(1) = 0.9, \quad \forall j \neq 1 \quad P(j) = 0.1/5
\]

\[
\forall j \quad Q(j) = 1/6
\]

\[
\text{P(1)} = 0.9, \quad \forall j \neq 1 \quad \text{P}(j) = 0.1/5
\]

\[
\forall j \quad \text{Q}(j) = 1/6
\]

\[
\text{Set} = \{2, 4, 6\}
\]

\[
\text{What is P(\text{Set})?}
\]

\[
= 0.12 \times \frac{1}{n} \sum_{t=1}^{n} 1\{x_t \in \{2, 4, 6\}\} \approx 0.12 \times 0.5 = 0.06
\]
We had the problem of too many rejections because probability of getting our sample to match exactly the observation is very low!
We had the problem of too many rejections because probability of getting our sample to match exactly the observation is very low!

How do we fix this?
We had the problem of too many rejections because probability of getting our sample to match exactly the observation is very low!

How do we fix this?

Fix observations and sample only states from the markov chain!
We had the problem of too many rejections because probability of getting our sample to match exactly the observation is very low!

How do we fix this?

Fix observations and sample only states from the markov chain!

Desired distribution $P$: $P(S_1, \ldots, S_N | X_1 = x_1, \ldots, X_N = x_N)$
We had the problem of too many rejections because probability of getting our sample to match exactly the observation is very low!

How do we fix this?

Fix observations and sample only states from the markov chain!

Desired distribution $P$: $P(S_1, \ldots, S_N | X_1 = x_1, \ldots, X_N = x_N)$

Sampling distribution $Q$: $P(S_1, \ldots, S_N)$
Importance Sampling

For a given sample $s_1, \ldots, s_N$, importance weight given by:

$$
\frac{P(S_1 = s_1, \ldots, S_N = s_n | X_1 = x_1, \ldots, X_N = x_N)}{P(S_1 = s_1, \ldots, S_N = s_N)}
$$
For a given sample $s_1, \ldots, s_N$, importance weight given by:

$$
\frac{P(S_1 = s_1, \ldots, S_N = s_n | X_1 = x_1, \ldots, X_N = x_N)}{P(S_1 = s_1, \ldots, S_N = s_N)}
\frac{P(X_1 = x_1, \ldots, X_N = x_N | S_1 = s_1, \ldots, S_N = s_n)}{P(X_1 = x_1, \ldots, X_N = x_N)}
$$
For a given sample $s_1, \ldots, s_N$, importance weight given by:

\[
\frac{P(S_1 = s_1, \ldots, S_N = s_N \mid X_1 = x_1, \ldots, X_N = x_N)}{P(S_1 = s_1, \ldots, S_N = s_N)}
= \frac{P(X_1 = x_1, \ldots, X_N = x_N \mid S_1 = s_1, \ldots, S_N = s_N)}{P(X_1 = x_1, \ldots, X_N = x_N)}
= \frac{\prod_{t=1}^{N} P(X_t = x_t \mid S_t = s_t)}{P(X_1 = x_1, \ldots, X_N = x_N)} \propto \prod_{t=1}^{N} P(X_t = x_t \mid S_t = s_t)
\]
Eg: say observations were
Hidden Markov Model (HMM)

Eg: say observations were

Importance weighting: weight samples
Eg: say observations were...

Importance weighting: weight samples
Eg: say observations were

\[ P(\text{pink} | X_1=13) \times P(\text{pink} | X_2=8) \times P(\text{pink} | X_3=9) \times P(\text{pink} | X_5=24) \times P(\text{purple} | X_5=19) \times P(\text{purple} | X_4=14) \]

Importance weighting: weight samples
Use multiple samples and track each one's weights.
Use multiple samples and track each one's weights.
Use multiple samples and track each one's weights.
Use multiple samples and track each one's weights.
Use multiple samples and track each one's weights.
HMM Particle Filter

- Use multiple samples and track each one's weights.
HMM Particle Filter

- Use multiple samples and track each one's weights.
Use multiple samples and track each one's weights.
- Use multiple samples and track each one's weights.
HMM Particle Filter

- Use multiple samples and track each one's weights.

- This is same as 6 separate samples
Use multiple samples and track each one's weights.

- This is same as 6 separate samples
- Instead of tracking each sample's weight, resample according to weights
Use multiple samples and track each one's weights.

- This is same as 6 separate samples
- Instead of tracking each sample's weight, resample according to weights
- Problem: Too many samples have negligible weight!
Instead of tracking each one, resample!
Instead of tracking each one, resample!
Instead of tracking each one, resample!
Instead of tracking each one, resample!

\[ P(\text{pink} | X_1) \]
Instead of tracking each one, resample!
Instead of tracking each one, resample!
HMM Particle Filter

Instead of tracking each one, resample!

\[ P(\text{pink} | X_1) \quad P(\text{pink} | X_2) \]
Instead of tracking each one, resample!
Instead of tracking each one, resample!
Instead of tracking each one, resample!
Instead of tracking each one, resample!
Instead of tracking each one, resample!
Instead of tracking each one, resample!

- On every round, transfer particles from previous states according to transition probability.
Instead of tracking each one, resample!

- On every round, transfer particles from previous states according to transition probability
- Resample particles according to $P(\text{observation}|\text{state})$
HMM Particle Filter

Instead of tracking each one, resample!

- On every round, transfer particles from previous states according to transition probability
- Resample particles according to $P(\text{observation}|\text{state})$
- Use new particles to proceed
Particle Filtering
Particle Filtering

- Without resampling, we carry many particles with very small probabilities
Particle Filtering

• Without resampling, we carry many particles with very small probabilities

  • too many samples needed for a good estimate
Particle Filtering

• Without resampling, we carry many particles with very small probabilities

  • too many samples needed for a good estimate

• By resampling, we got rid of samples with very small probabilities
Particle Filtering

- Without resampling, we carry many particles with very small probabilities
  - too many samples needed for a good estimate
- By resampling, we got rid of samples with very small probabilities
  - Hence fewer samples suffice
• Inference time only depends on number of samples

• Of course more the samples the better accuracy

• Often we don’t need too many samples. Why?
Gibbs Sampling
Gibbs Sampling

• Repeat n times for, n samples,
Gibbs Sampling

• Repeat n times for, n samples,

• Start with arbitrary value for variables
Gibbs Sampling

- Repeat $n$ times for $n$ samples,
  - Start with arbitrary value for variables
  - Replace each variable by new sample from $P(\text{Variable}|\text{all other variables})$
Gibbs Sampling

- Repeat n times for, n samples,
  - Start with arbitrary value for variables
  - Replace each variable by new sample from \( P(\text{Variable}| \text{all other variables}) \)
  - Go over all variables multiple times
Gibbs Sampling

- Repeat n times for n samples,
  - Start with arbitrary value for variables
  - Replace each variable by new sample from $P(\text{Variable} | \text{all other variables})$
  - Go over all variables multiple times
  - Return final sample of the N variables