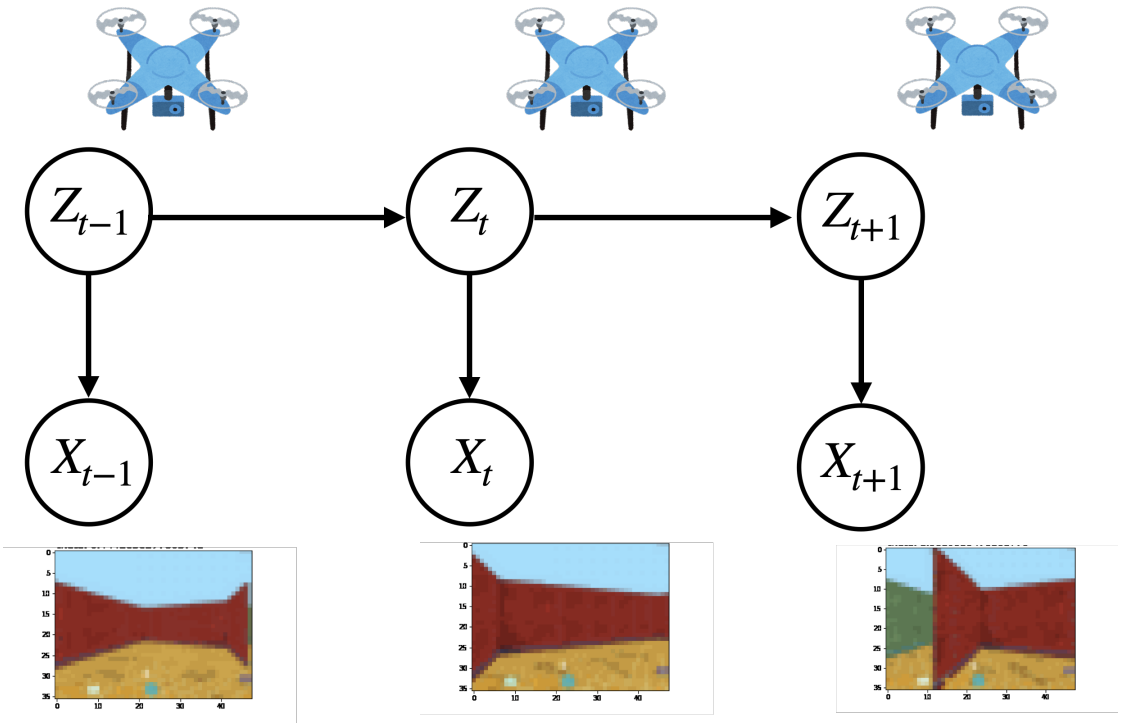
Importance Weighting Approach in Kernel Bayes' Rule

Liyuan Xu [1] Yutian Chen [2]
Arnaud Doucet [2] Arthur Gretton [1]

[1]Gatsby Unit

[2]DeepMind

Drone Localization



- Predict drone location Z_t from camera images $X_1, \ldots, X_t$.
- One approach: Bayes' Filter

Bayes' Filter

$$P(Z_{t+1}|X_{1;:::;t}) = \int P(Z_{t+1}|Z_t)\mathrm{d}P(Z_t|X_{1;:::;t})$$

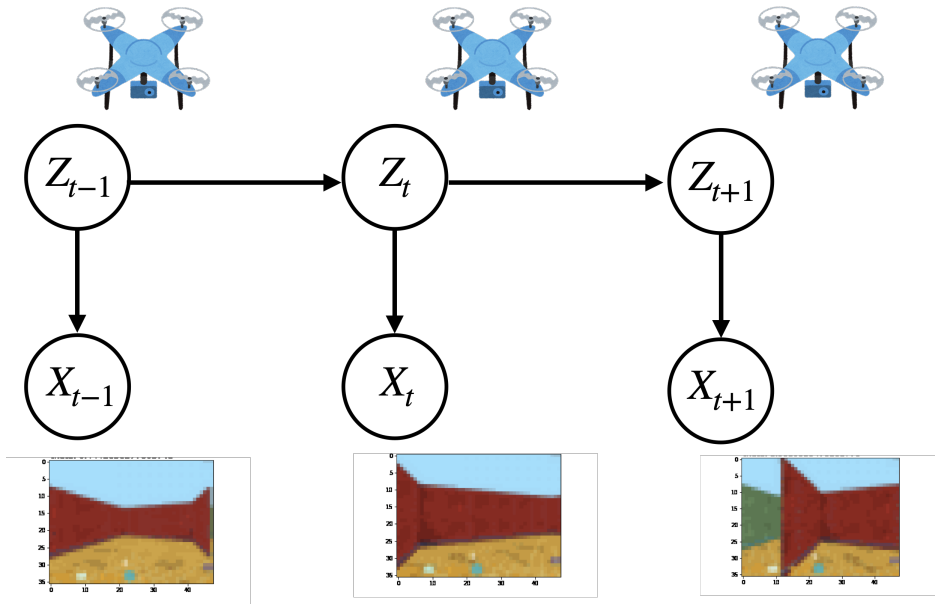
- Known as "sum rule".

Bayes' Filter

$$P(Z_{t+1}|X_{1:\ldots:t+1}) \;/\; P(X_{t+1}|Z_{t+1})P(Z_{t+1}|X_{1:\ldots:t+1})$$

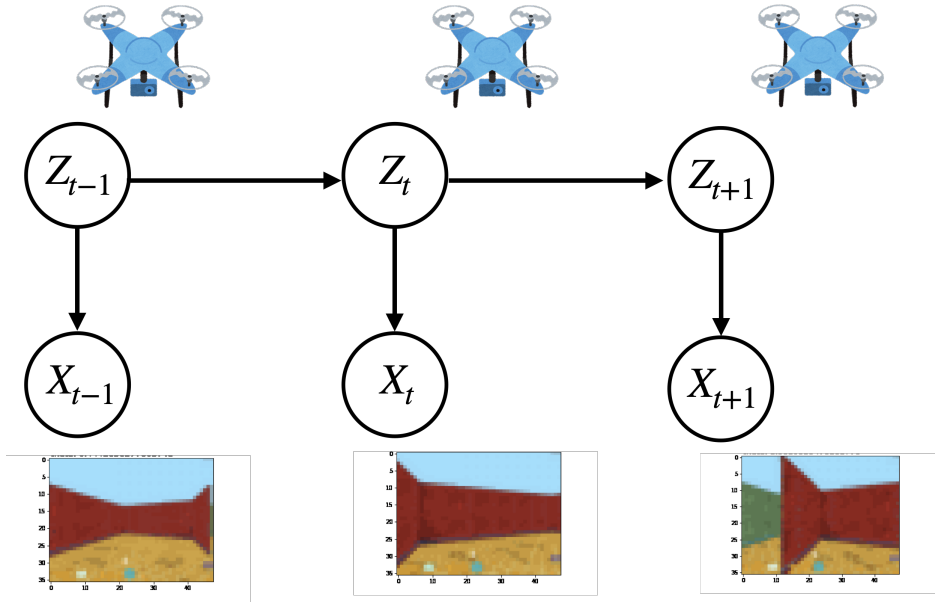
- Bayesian update of prior $P(Z_{t+1}|X_{1:\ldots:t})$ given observation X_{t+1}
- Likelihood function is $P(X_{t+1}|Z_{t+1}) = P(X_t|Z_t)$

Difficulty of Bayes' Filter



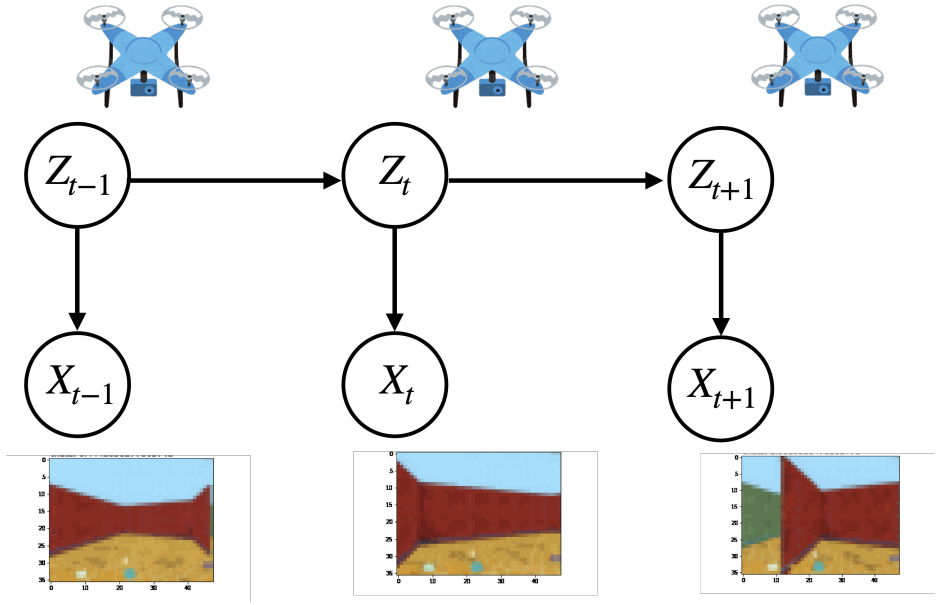
- There is no explicit model of $P(Z_{t+1}|Z_t)$ or $P(X_t|Z_t)$.
  - ! Need to learn them from data $\{X_t; Z_t\}$
- $P(Z_t|X_{1:t})$ is assumed to be in a specific parametric form.
  - ! Might cause a bias in estimation.
- Desirable to use non-parametric representation of distributions.

Difficulty of Bayes' Filter



- There is no explicit model of $P(Z_{t+1}|Z_t)$ or $P(X_t|Z_t)$.
  ! Need to learn them from data $\{X_t; Z_t\}$
- $P(Z_t|X_{1:t})$ is assumed to be in a specific parametric form.
  ! Might cause a bias in estimation.
- Desirable to use non-parametric representation of distributions.

Difficulty of Bayes' Filter



- There is no explicit model of $P(Z_{t+1}|Z_t)$ or $P(X_t|Z_t)$.
  ! Need to learn them from data $\{X_t; Z_t\}$
- $P(Z_t|X_{1:t})$ is assumed to be in a specific parametric form.
  ! Might cause a bias in estimation.
- Desirable to use non-parametric representation of distributions.

RKHS Embeddings

- Define kernel $k(x; x^0)$ and accompanied feature map (x).

$$k(x; x^0) = \quad (x); \ (x^0) \ ; \quad f(x) = hf; \ (x)i$$

- Mean embedding $_P$ of distribution P is defined as

$$_P = \mathrm{E}_P[\ (X)];$$

- It can be generalized to conditional distribution $P_{X\,jZ}$

$$_{P_{X\,jZ}}(z) = \mathrm{E}_P[\ (X)jZ = z]$$

RKHS Embeddings

- Define kernel $k(x; x^0)$ and accompanied feature map (x).

$$k(x; x^0) = \quad (x); \ (x^0) \ ; \quad f(x) = hf; \ (x)i$$

- Mean embedding $_P$ of distribution P is defined as

$$_P = \mathrm{E}_P [\ (X)];$$

- It can be generalized to conditional distribution $P_{X|Z}$

$$_{P_{X|Z}}(z) = \mathrm{E}_P [\ (X)|Z = z]$$

RKHS Embeddings

- Define kernel $k(x; x')$ and accompanied feature map $\phi(x)$.

$$k(x; x') = \langle \phi(x); \phi(x') \rangle ; \quad f(x) = \langle f; \phi(x) \rangle$$

- Mean embedding μ_P of distribution P is defined as

$$\mu_P = \mathrm{E}_P[\phi(X)];$$

- It can be generalized to conditional distribution $P_{X|Z}$

$$\mu_{P_{X|Z}}(z) = \mathrm{E}_P[\phi(X)|Z = z]$$

Advantage of RKHS Embeddings

- Embedding $_P$ can uniquely determine the distribution.
- Embeddings can be non-parametrically estimated from $f X_i; Z_i g_{i=1}^n$ as

$$_P = \frac{1}{n} \sum_{i=1} (x_i); \qquad _{P_{X|Z}}(z) = \frac{1}{n} \sum_{i=1} w_i(z) \ (x_i)$$

  for some weighting function $w_i(z)$.
- Kernel Bayes' Filter:

  Represent distributions $P(Z_t | X_{1,\ldots,t})$ using embeddings

Advantage of RKHS Embeddings

- Embedding $_P$ can uniquely determine the distribution.
- Embeddings can be non-parametrically estimated from $f X_i; Z_i g_{i=1}^n$ as

$$_P = \frac{1}{n} \sum_{i=1}^{n} (x_i); \qquad _{P_{X|Z}}(z) = \frac{1}{n} \sum_{i=1}^{n} w_i(z) \ (x_i)$$

  for some weighting function $w_i(z)$.
- Kernel Bayes' Filter:

  Represent distributions $P(Z_t | X_{1,\ldots,t})$ using embeddings

Advantage of RKHS Embeddings

- Embedding $_P$ can uniquely determine the distribution.
- Embeddings can be non-parametrically estimated from $f X_i; Z_i g_{i=1}^n$ as

$$_P = \frac{1}{n} \sum_{i=1}^{n} (x_i); \qquad _{P_{X|Z}}(z) = \frac{1}{n} \sum_{i=1}^{n} w_i(z) \ (x_i)$$

  for some weighting function $w_i(z)$.
- Kernel Bayes' Filter:

  Represent distributions $P(Z_t | X_{1;\dots;t})$ using embeddings

Kernel Bayes' Filter

$$P_{Z_t|X_{1,\dots,t}} \xrightarrow{\text{Data: } \{Z_t, Z_{t+1}\}} P_{Z_{t+1}|X_{1,\dots,t}}$$

- Use kernel sum rule [Song et al. 2011] for $P_{Z_{t+1}|X_{1,\dots,X_t}}$.

Kernel Bayes' Filter

$$P_{Z_{t+1}|X_{1,\dots,t}} \xrightarrow[\text{Observation: } X_{t+1}]{\text{Data: } \{Z_t, X_t\}} P_{Z_{t+1}|X_{1,\dots,t+1}}$$

- Bayesian update given the embedding of the prior $P_{Z_{t+1}|X_{1,\dots,t}}$
- This update is called kernel Bayes' rule.

Kernel Bayes' Rule [Fukumizu+ 2013]

Given

- Training data $\{X_i, Z_i\}$ $P(X|Z)P(Z)$
- Embedding of prior (Z)

Outputs posterior embedding

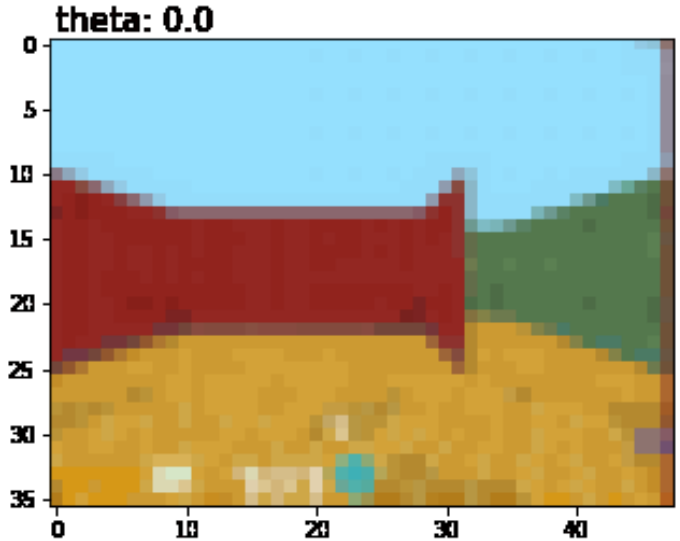
$$_Q(x) = \int (z) \frac{P(x|z) \ (z)}{P(x|z) \ (z)dz} dz$$

Contribution

Proposed a novel instance of kernel Bayes' rule.

- Based on importance weighting.
- Achieves superior numerical stability to existing work [Fukumizu+ 2013].
- Admits the use of neural network feature in kernel Bayes' rule.

DeepMind Lab Experiment



- A drone is rotating in a maze.
- Latent Z_t: True angle of the drone.
- Observation X_t: The image observed at the noisy version of Z_t
- Task: Predict Z_t from $X_1, \cdots, X_t$

DeepMind Lab Experiment

- `Original`: Existing kernel Bayes rule [Fukumizu+ 2013]
- `LSTM`: Directly regress Z_t from $X_{t-5},\ldots,t$
- `IW`: New kernel Bayes' rule with RKHS feature
- `IW(NN)`: New kernel Bayes' rule with adaptive feature

DeepMind Lab Experiment

- `Original`: Existing kernel Bayes rule [Fukumizu+ 2013]
- LSTM: Directly regress Z_t from $X_{t-5:\cdots:t}$
- `IW`: New kernel Bayes' rule with RKHS feature
- `IW(NN)`: New kernel Bayes' rule with adaptive feature

DeepMind Lab Experiment

- `Original`: Existing kernel Bayes rule [Fukumizu+ 2013]
- `LSTM`: Directly regress Z_t from $X_{t-5:\cdots:t}$
- `IW`: New kernel Bayes' rule with RKHS feature
- `IW(NN)`: New kernel Bayes' rule with adaptive feature

DeepMind Lab Experiment

- `Original`: Existing kernel Bayes rule [Fukumizu+ 2013]
- `LSTM`: Directly regress Z_t from $X_{t-5,\ldots,t}$
- `IW`: New kernel Bayes' rule with RKHS feature
- `IW(NN)`: New kernel Bayes' rule with adaptive feature