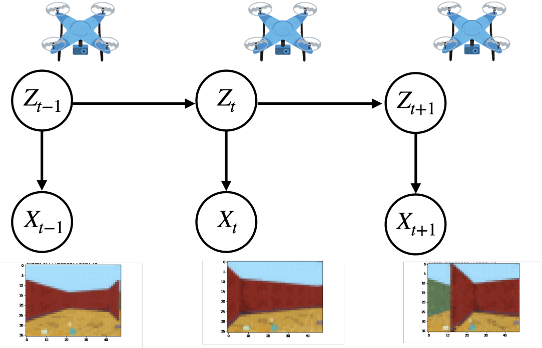
Importance Weighting Approach in Kernel Bayes' Rule

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

[1]Gatsby Unit

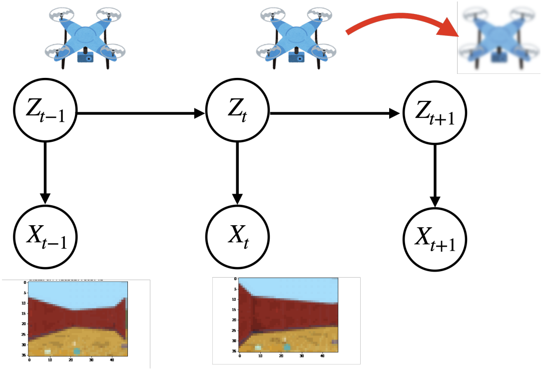
[2]DeepMind

Drone Localization



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

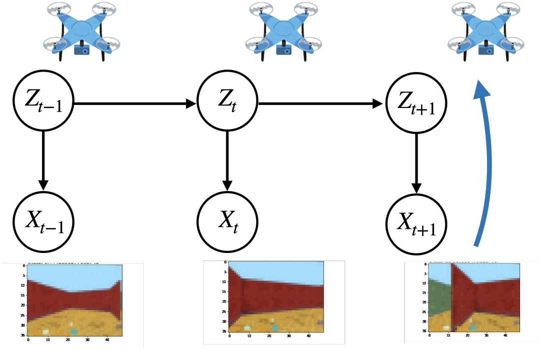
Bayes' Filter



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

- Known as “sum rule”.

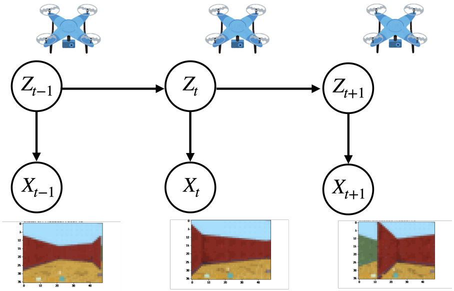
Bayes' Filter



$$P(Z_{t+1}|X_{1,\ldots,t+1}) \propto P(X_{t+1}|Z_{t+1})P(Z_{t+1}|X_{1,\ldots,t+1})$$

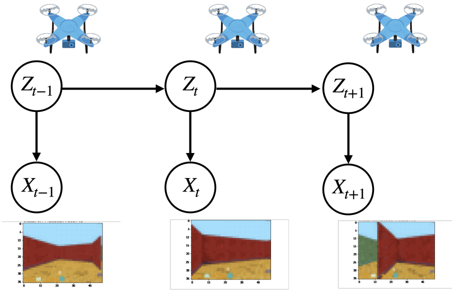
- Bayesian update of prior $P(Z_{t+1}|X_{1,\ldots,t})$ given observation X_{t+1}
- Likelihood function is $P(X_{t+1}|Z_{t+1}) = P(X_t|Z_t)$

Difficulty of Bayes' Filter



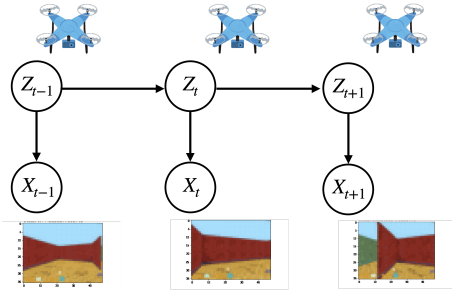
- There is no explicit model of $P(Z_{t+1}|Z_t)$ or $P(X_t|Z_t)$.
  $\rightarrow$ Need to learn them from data $\{X_t, Z_t\}$
- $P(Z_t|X_{1,\dots,t})$ is assumed to be in a specific parametric form.
  $\rightarrow$ Might cause a bias in estimation.
- Desirable to use non-parametric representation of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

Difficulty of Bayes' Filter



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

RKHS Embeddings

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

$$k(x, x') = \left\langle \phi(x), \phi(x') \right\rangle, \quad f(x) = \langle f, \phi(x) \rangle$$

- Mean embedding μ_P of distribution P is defined as

$$\mu_P = \mathbb{E}_P \left[\phi(X) \right],$$

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

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

RKHS Embeddings

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

$$k(x, x') = \left\langle \phi(x), \phi(x') \right\rangle, \quad f(x) = \langle f, \phi(x) \rangle$$

- Mean embedding μ_P of distribution P is defined as

$$\mu_P = \mathbb{E}_P \left[\phi(X) \right],$$

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

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

RKHS Embeddings

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

$$k(x, x') = \left\langle \phi(x), \phi(x') \right\rangle, \quad f(x) = \langle f, \phi(x) \rangle$$

- Mean embedding μ_P of distribution P is defined as

$$\mu_P = \mathbb{E}_P \left[\phi(X) \right],$$

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

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

Advantage of RKHS Embeddings

- Embedding μ_P can uniquely determine the distribution.
- Embeddings can be non-parametrically estimated from $\{X_i, Z_i\}_{i=1}^n$ as

$$\mu_P = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i), \quad \mu_{P_{X|Z}}(z) = \frac{1}{n}\sum_{i=1}^{n} w_i(z)\phi(x_i)$$

  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 μ_P can uniquely determine the distribution.
- Embeddings can be non-parametrically estimated from $\{X_i, Z_i\}_{i=1}^n$ as

$$\mu_P = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i), \quad \mu_{P_{X|Z}}(z) = \frac{1}{n}\sum_{i=1}^{n} w_i(z)\phi(x_i)$$

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

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

Advantage of RKHS Embeddings

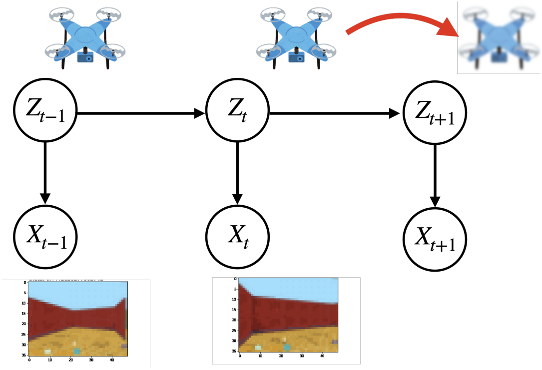
- Embedding μ_P can uniquely determine the distribution.
- Embeddings can be non-parametrically estimated from $\{X_i, Z_i\}_{i=1}^n$ as

$$\mu_P = \frac{1}{n}\sum_{i=1}^n \phi(x_i), \quad \mu_{P_{X|Z}}(z) = \frac{1}{n}\sum_{i=1}^n w_i(z)\phi(x_i)$$

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

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

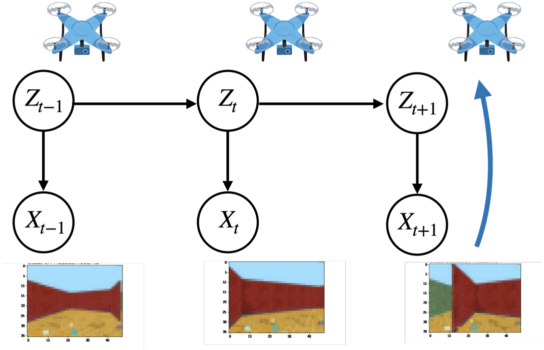
Kernel Bayes' Filter



$$\mu_{P_{Z_t|X_{1,\ldots,t}}} \xrightarrow{\text{Data: } \{Z_t, Z_{t+1}\}} \mu_{P_{Z_{\color{red}{t+1}}|X_{1,\ldots,t}}}$$

- Use kernel sum rule [Song et al. 2011] for $\mu_{P_{Z_{\color{red}{t+1}}|X_1,\ldots,X_t}}$.

Kernel Bayes' Filter



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

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

Kernel Bayes' Rule [Fukumizu+ 2013]

Given

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

Outputs posterior embedding

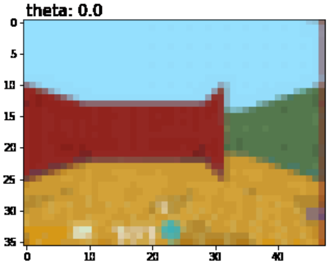
$$\mu_Q(x) = \int \phi(z) \frac{P(x|z)\pi(z)}{\int P(x|z)\pi(z)\mathrm{d}z} \mathrm{d}z$$

Contribution

Proposed a novel instance of kernel Bayes' rule.

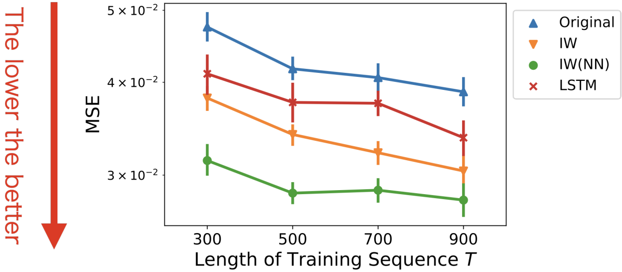
- Based on importance weighting.
- Achieves superior numerical stability to existing work [Fukumizu+ 2013].
- Admits the use of neural network feature in kernel Bayes' rule.

DeepMind Lab Experiment



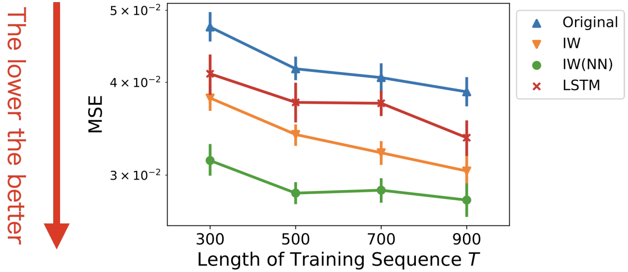
- A drone is rotating in a maze.
- Latent Z_t: True angle of the drone.
- Observation X_t: The image observed at the noisy version of Z_t
- Task: Predict Z_t from $X_1, \ldots, X_t$

DeepMind Lab Experiment



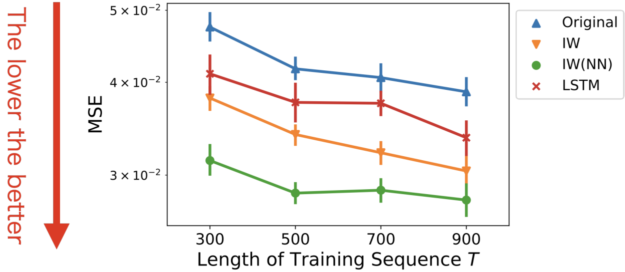
- `Original`: Existing kernel Bayes rule [Fukumizu+ 2013]
- `LSTM`: Directly regress Z_t from $X_{t-5,\ldots,t}$
- `IW`: New kernel Bayes' rule with RKHS feature
- `IW(NN)`: New kernel Bayes' rule with adaptive feature

DeepMind Lab Experiment



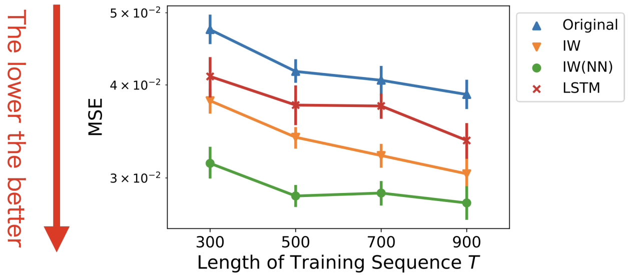
- `Original`: Existing kernel Bayes rule [Fukumizu+ 2013]
- `LSTM`: Directly regress Z_t from $X_{t-5,\dots,t}$
- `IW`: New kernel Bayes' rule with RKHS feature
- `IW(NN)`: New kernel Bayes' rule with adaptive feature

DeepMind Lab Experiment



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