カーネル法入門
7. カーネル法によるノンパラメトリックなベイズ推論

福水健次

統計数理研究所／総合研究大学院大学

大阪大学大学院基礎工学研究科・集中講義

2014 September

Bayesian inference

– Bayes' rule

$$q(x|y) = \frac{p(y|x)\pi(x)}{\int p(y|x)\pi(x)dx}$$

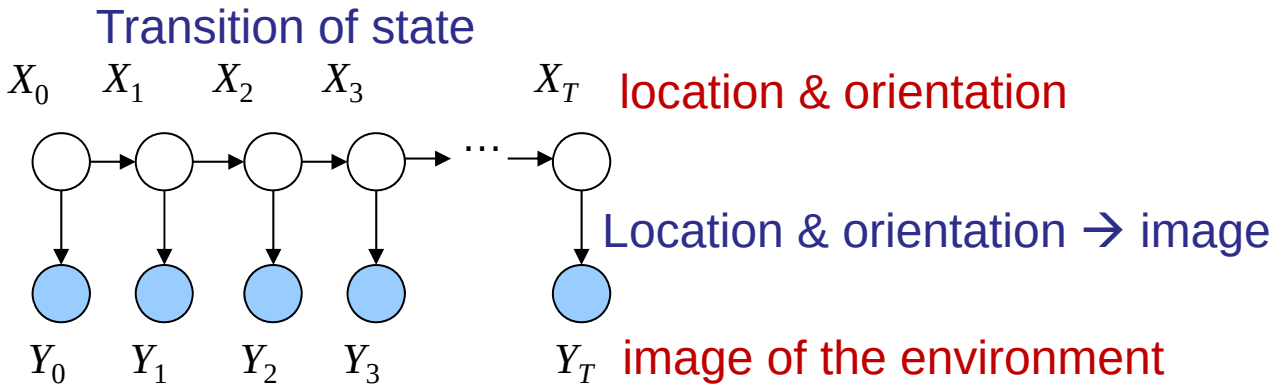
Thomas Bayes (1701-1761)
Portrait?

- PROS
  - Principled and flexible method for statistical inference.
  - Can incorporate prior knowledge.
- CONS
  - Computation: integral is needed
    - » Numerical integration: MCMC, SMC, etc
    - » Approximation: Variational Bayes, belief propagation, expectation propagation, etc.

Motivating Example

Robot localization

- State $X_t \in \mathbf{R}^3$: 2-D coordinate and orientation of a robot
- Observation Y_t : image sequence taken by a camera on the robot.
- Task: estimate the location of a robot from image sequences.
- It is well formulated by a HMM.



– Estimation / prediction with HMM

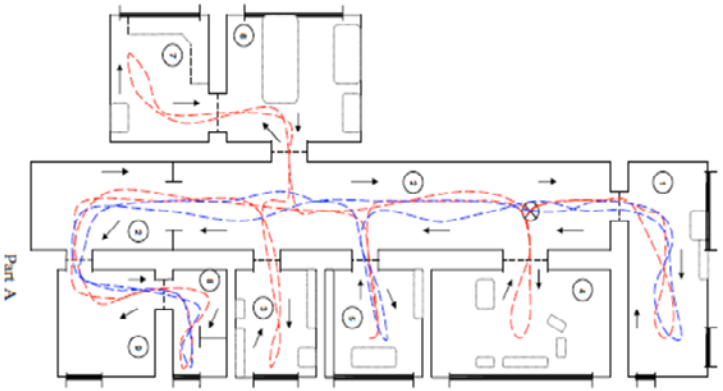
→ Sequential application of Bayes' rule solves the task.

– Nonparametric approach is needed:

Observation model: $p(Y_t|X_t)$ is very difficult to model with a simple parametric model.

→"Nonparametric" implementation of Bayesian inference

c.f. Monte Carlo approach needs to know the density functions.



location & orientation

$p(Y_t|X_t)$

image of the environment

Kernel method for Bayesian inference

- Topic of this chapter
  - A new nonparametric / kernel approach to Bayesian inference
    - Kernel mean approach to represent and manipulate probabilities.
    - “Nonparametric” Bayesian inference
      - No densities are needed, but data is needed.
    - Bayesian inference with matrix computation.
      - Computation is done with Gram matrices.
      - No integral, no approximate inference.

Outline

1. Introduction
2. Representing conditional probabilities
3. Kernel sum rule and chain rule
4. Kernel Bayes rule: kernel methods for Bayesian inference
5. Conclusions

Representing conditional probabilities

Review: kernel mean

X: random variable taking value on a measurable space Ω, $\sim P$.

k: pos.def. kernel on Ω. H: RKHS defined by k.

– kernel mean:

$$m_P := E[\Phi(X)] = E[k(\cdot, X)] = \int k(\cdot, x) dP(x) \quad \in H_k$$

– With a characteristic kernel, kernel mean uniquely identifies the distribution.

- Embedding of the probabilities into the RKHS.
- Examples: Gaussian, Laplace kernel.

Nonparametric inference with kernels

Principle: with characteristic kernels,

Inference on P $\Rightarrow$ Inference on m_P

- Two sample test → $m_P = m_Q$? MMD
- Independence test → $m_{XY} = m_X \otimes m_Y$? HSIC
- Bayesian Inference → this chapter.

Operations for inference

- Operations on probabilities necessary for inference
  - Conditioning: $p_{XY}(x,y) \rightarrow p(y|x)$
  - Sum rule: $p_{XY}(x,y) \rightarrow p_X(x) = \sum_y p_{XY}(x,y)$
  - Product rule: $p_{Y|X}(y|x), p_X(x) \rightarrow p_{XY}(x,y) = p_{Y|X}(y|x)p_X(x)$
    (aka. Chain rule)
  - Bayes' rule: $p(y|x), \pi(x) \rightarrow q(x|y) = \frac{p(y|x)\pi(x)}{\int p(y|x)\pi(x)dx}$

Conditional kernel mean

- Review: conditional mean of Gaussian r.v.

  X, Y: Gaussian random vectors with mean 0 ($\in \mathbf{R}^m, \mathbf{R}^\ell$, resp.)

  – Least square prediction

$$\operatorname*{argmin}_{A \in R^{\ell \times m}} \int \|Y - AX\|^2 dP(X,Y) \;=\; V_{YX} {V_{XX}}^{-1}$$

V_{YX}, V_{XX}: 共分散行列

  – Conditional expectation

$$E[Y|X = x] = V_{YX} {V_{XX}}^{-1} x$$

General random variables

With characteristic kernels, for general X and Y,

– Least square prediction

$$\underset{F:H_X \to H_Y,\ \text{linear}}{\operatorname{argmin}} \int \left\| \Phi_y(Y) - \underbrace{F(X)}_{\langle F, \Phi_x(X) \rangle} \right\|_{H_y}^2 dP(X,Y) = C_{YX} {C_{XX}}^{-1}$$

– Conditional expectation

$$\underbrace{E\left[\Phi_y(Y) \middle| X = x\right]}_{\text{Kernel mean of } p(y|x)} = \underbrace{C_{YX} {C_{XX}}^{-1} \Phi_x(x)}_{\text{Expression by kernel mean and covariance operators. Estimable!}}$$

Empirical estimation

$$E[\Phi(Y)|X = x] = C_{YX}{C_{XX}}^{-1}\Phi_X(x)$$

$$\Rightarrow$$

$$\hat{m}_{Y|X=x} := \hat{E}_{reg}[\Phi(Y)|X = x] = \hat{C}_{YX}\left(\hat{C}_{XX} + \varepsilon_n I\right)^{-1}\Phi_X(x)$$

ε_n: regularization coefficient

– Remark: Even in population operators, regularized inversion is needed if the dimensionality is infinite. Since $Tr[C_{XX}] < \infty$, there are always arbitrarily small eigenvalues.

定理（一致性）

$E[k(Y,\tilde{Y})|X=x,\tilde{X}=\tilde{x}]$が$(x,\tilde{x})$の関数としてRange$(C_{XX})\otimes$ Range(C_{XX}) に属すると仮定する. このとき $\varepsilon_n = n^{-1/5}$に対し,

$$\left\|\hat{C}_{YX}\left(\hat{C}_{XX}+\varepsilon_n I\right)^{-1}\Phi_X(x)-E\left[\Phi_y(Y)\middle|X=x\right]\right\|_{Hy}=O_p\left(n^{-\frac{1}{5}}\right).$$

命題（グラム行列表現）

$$\hat{C}_{YX}\left(\hat{C}_{XX}+\varepsilon_n I\right)^{-1}\Phi_X(x)=\sum_{i=1}^{n} w_i k_y(\cdot,Y_i)$$

$$w=(G_X+n\varepsilon_n I_n)^{-1}\mathbf{k}_{\mathrm{X}}(x)$$

$$\mathbf{k}_X(x)=\left(k_X(x,X_1),\ldots,k_X(x,X_n)\right)^T\in\mathbf{R}^n,$$

■ Proof （Gram行列表現）

Let $\xi = \left(\hat{C}_{XX} + \varepsilon_n I\right)^{-1} \Phi_X(x)$, and

decompose it as $\xi = \sum_j \alpha_j \Phi_X\left(X_j\right) + h_\perp$, $h_\perp$ is orthogonal to $\mathrm{Span}\{\Phi_X(X_i)\}_{i=1}^n$.

$$\Phi_X(x) = \left(\hat{C}_{XX} + \varepsilon_n I\right)\left(\sum_j \alpha_j \Phi_X\left(X_j\right) + h_\perp\right)$$
$$= \frac{1}{n}\sum_{i,j} \Phi_X(X_i) k\left(X_i, X_j\right)\alpha_j + \varepsilon_n \sum_j \alpha_j \Phi_X\left(X_j\right) + \varepsilon_n h_\perp .$$

Take inner product with $\Phi_X(X_\ell)$. Then,

$$\mathbf{k}_X(x) = \frac{1}{n} G_X^2 \alpha + \varepsilon_n G_X \alpha = \left(\frac{1}{n} G_X + \varepsilon_n I_n\right) G_X \alpha. \qquad \cdots (*)$$

The target is

$$\hat{C}_{YX}\xi = \frac{1}{n}\sum_i \Phi_Y(Y_i) k_X\left(X_i, X_j\right)\alpha_j$$
$$= \frac{1}{n}\mathbf{k}_Y^T(\cdot) G_X \alpha = \frac{1}{n}\mathbf{k}_Y^T(\cdot)\left(\frac{1}{n} G_X + \varepsilon_n I_n\right)^{-1} \mathbf{k}_X(x)$$
$$(*)$$

– 一致性の証明は略

条件付カーネル平均の使い方

– ノンパラメトリック回帰

$$\hat{E}[g(Y)|X=x] = \langle \hat{m}_{Y|X=x}, g \rangle_{H_Y}$$
$$= \mathbf{k}_X^T(x)(G_X + n\varepsilon_n I_n)^{-1}\mathbf{g}$$

$$\mathbf{g} = \big(g(Y_1), \ldots, g(Y_n)\big)^T \in \mathbf{R}^n$$

カーネルリッジ回帰

– 条件付独立性への応用 (Fukumizu et al. JMLR 2004, AoS 2009, NIPS 2010)

– ベイズ推論

Note: 一致性に対しては, カーネル（バンド幅）は固定, 正則化係数のみ $\varepsilon_n \to 0$ とする.

c.f. 平滑化カーネル.

Addendum: Kernel ridge regression

– (Linear) ridge regression

$$\min_{a \in \mathbf{R}^m} \sum_{i=1}^{n} \|Y_i - a^T X_i\|^2 + \lambda \|a\|^2$$

$$\hat{y}(x) = \hat{a}_\lambda^T x = Y^T X (X^T X + \lambda I_n)^{-1} x$$

Popularly used when $X^T X$ is (almost) degenerate.

– Kernel ridge regression

$$\min_{f \in H_X} \sum_{i=1}^{n} \|Y_i - f(X_i)\|^2 + \lambda \|f\|_{H_X}^2$$

$$\hat{f}(x) = Y^T (G_X + n\lambda I_n)^{-1} \mathbf{k}_X^T(x)$$

Nonparametric regression with kernel

Nonparametric regression: theoretical comparison

Assume Y is 1 dim., and kernel is used only for X

→ $$\hat{E}_{reg}[Y|X=x] \coloneqq \mathbf{k}_X^T(x)(G_X + n\varepsilon_n I_n)^{-1}Y$$

Same as Gaussian process / kernel ridge regression

– Consistency 1 (Eberts & Steinwart 2011)

If k_X is Gaussian, and $E[Y|X] \in W_2^\alpha(P_X)$, (under some technical assumptions) for any $\rho > 0$,

$$E\left|\hat{E}[Y|X] - E[Y|X]\right|^2 = O_p\left(n^{-\frac{2\alpha}{2\alpha+m}+\rho}\right) \qquad (n \to \infty)$$

Note: $O_p\left(n^{-\frac{2\alpha}{2\alpha+m}}\right)$ is the optimal rate for a linear estimator (Stone 1982).

* $W_2^\alpha(P_X)$: Sobolev space of order α.

– Consistency 2 (case: $E[Y|X] \in H_X$)

Suppose $E[Y|X] \in R(C_{XX}^{\beta})$ with $\beta \geq 0$, Then, with a characteristic kernel k_X,

$$E\left|\hat{E}[Y|X] - E[Y|X]\right|^2 = O_p\left(n^{-\min\left\{\frac{1}{2}, \frac{2\beta+1}{2\beta+3}\right\}}\right)$$

$$\left\|\hat{E}[Y|X] - E[Y|X]\right\|_{H_X}^2 = O_p\left(n^{-\min\left\{\frac{1}{2}, \frac{\beta}{\beta+1}\right\}}\right)$$

- The rates do not depend on m (dim of X), since the analysis can be done within the RKHS.
- $\|\cdot\|_{H_X}$ is stronger than $\|\cdot\|_{sup}$. Thus,

$$\sup_x\left|\hat{E}[Y|X=x] - E[Y|X=x]\right| = O_p\left(n^{-\min\left\{\frac{1}{4}, \frac{\beta}{2\beta+2}\right\}}\right)$$

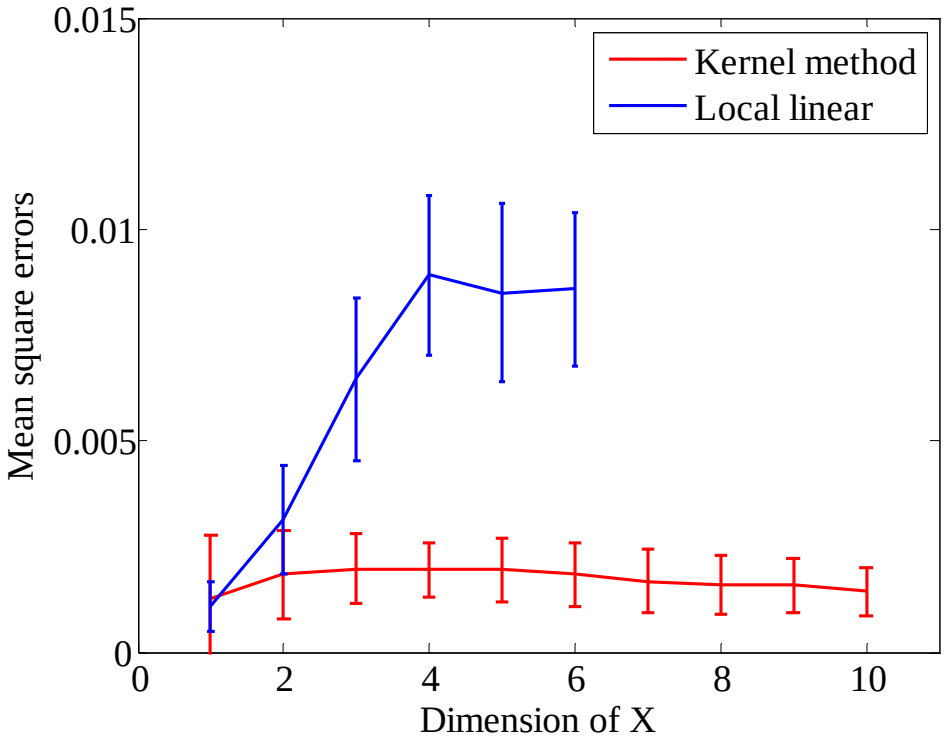
Nonparametric regression: experimental comparison

$Y = 1/(1.5 + ||X||^2) + Z, \qquad X \sim N(0, I_d),\ Z \sim N(0, 0.1^2)$

- Kernel ridge regression
  Gaussian kernel
- Local linear regression
  Epanechnikov kernel
  ('locfit' in R is used)

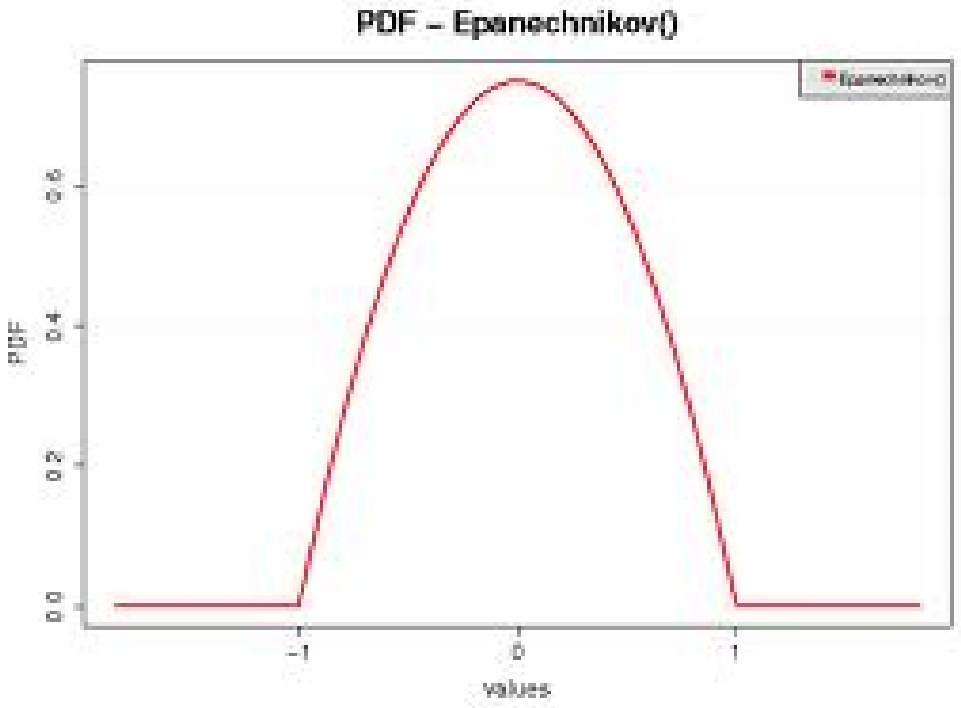
n = 100, 500 runs

Bandwidth parameters are chosen by CV.



– Epanechnikov kernel

$$K(u) = \frac{3}{4}(1 - u^2) I_{[-1,1]}(u)$$



Kernel Sum and Product Rules

Kernel Sum Rule

Sum rule

$$q_Y(y) = \int p(y|x)\pi(x)dx$$

Kernel Sum Rule

- Input: π のカーネル平均，共分散作用素 C_{YX}, C_{XX}
- output: q_Y のカーネル平均

条件付確率 $p(y|x)$ は共分散作用素 C_{YX}, C_{XX}の形で与える.
あるいは，サンプル $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P$ の形で与える. ここでP は条件付確率が $p(y|x)$ となる同時分布.

■ Intuition

$$m_{q_y} = C_{YX} C_{XX}^{-1} m_\pi$$

- $C_{XX}^{-1} m_\pi = \frac{\pi}{p_X}$ $\quad$ ($\frac{\pi}{p_X} \in H_x$を仮定せよ)

  $C_{XX} f = \int k_x(\cdot, x) f(x) p_X(x) dx$

  $m_\pi = \int k_x(\cdot, x) \pi(x) dx$

  より従う.

- $C_{YX} C_{XX}^{-1} m_\pi = C_{YX} \left(\frac{\pi}{p_X} \right)$

  $= \int \int k_y(\cdot, y) \frac{\pi}{p_X}(x) p(x, y) dx dy$

  $= \int k_y(\cdot, y) \int p(y|x) \pi(x) dx dy$

  $= m_{q_y}$.

■ 推定量

$$\hat{m}_{q_y} := \hat{C}_{YX}\left(\hat{C}_{XX} + \varepsilon_n I\right)^{-1}\hat{m}_{\pi}$$

– Gram行列表現

$\hat{m}_{\pi} = \sum_{i=1}^{\ell} \alpha_i k_x\left(\cdot, \tilde{X}_i\right), \quad (X_1, Y_1), \ldots, (X_n, Y_n) \sim P$

$$\hat{m}_{q_y} = \sum_{i=1}^{n} \beta_i k_y(\cdot, Y_i),$$

$$\beta = (G_X + n\varepsilon_n I_n)^{-1} G_{X\tilde{X}}\alpha, \qquad G_{X\tilde{X}} = \left(k\left(X_i, \tilde{X}_j\right)\right)_{ij}$$

■ 一致性

<u>定理</u>　$\|\hat{m}_{\pi} - m_{\pi}\|_{H_x} = O_p(n^{-\alpha}), \quad \frac{\pi}{p_X} \in R\left(C_{XX}^{\beta}\right)$ $(\beta \geq 0)$とするとき,

$$\left\|\hat{m}_{q_y} - m_{q_y}\right\|_{H_y} = O_p\left(n^{-\min\left\{\frac{2}{3}\alpha, \frac{2\beta+1}{2\beta+2}\alpha\right\}}\right), \qquad (n \to \infty)$$

Kernel Chain Rule

– Chain rule: $q(x,y) = p(y|x)\pi(x)$

– カーネル化: $m_q = C_{(YX)X} C_{XX}^{-1} m_\pi$ $\qquad C_{(YX)}: H_x \to H_y \otimes H_x$

$$\hat{m}_q := \hat{C}_{(YX)X}\left(\hat{C}_{XX} + \varepsilon_n I\right)^{-1} \hat{m}_\pi$$

– Gram行列表現:

Input: $\hat{m}_\pi = \sum_{i=1}^{\ell} \alpha_i \Phi\left(\tilde{X}_i\right), (X_1, Y_1), \ldots, (X_n, Y_n) \sim P_{XY}$

$\hat{m}_q = \sum_{i=1}^{n} \beta_i \Phi(Y_i) \otimes \Phi(X_i)$, $\qquad \beta = (G_X + n\varepsilon_n I_n)^{-1} G_{X\tilde{X}} \alpha.$

注意: 係数 β は, kernel sum rule と全く同一

– Intuition:

$C_{(YX)X}: H_X \to H_Y \otimes H_X, \qquad E\left[\left(\Phi(Y) \otimes \Phi(X)\right) \otimes \Phi(X)\right]$

確率変数 $(X, (X, Y))$ を考える.

条件付確率の密度関数: $p(y, x|x') = p(y|x)\delta(x - x')$

Kernel Sum Rule より

$$C_{(YX)X} C_{XX}^{-1} m_\pi = \int\int\int \Phi(y) \otimes \Phi(x)\, p(y, x|x')\pi(x')dydxdx'$$

$$= \int\int\int \Phi(y) \otimes \Phi(x)\, p(y|x)\delta(x - x')\pi(x')dydxdx'$$

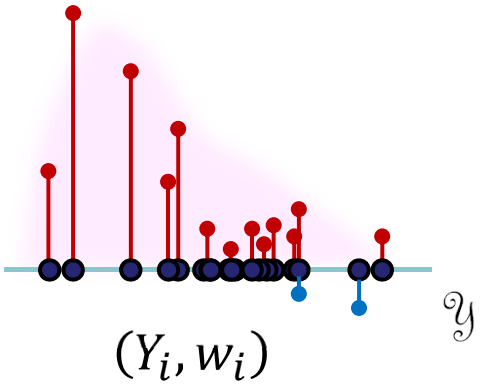
$$= \int\int \Phi(y) \otimes \Phi(x) p(y|x)\pi(x)dydx$$

$$= \ m_q$$

重み付サンプルとしての解釈

– Kernel Sum / Chain Rule

$$\hat{m}_q(\cdot) = \sum_{i=1}^{n} w_i k_X(\cdot, Y_i)$$



離散測度

$$\sum_{i=1}^{n} w_i \delta_{Y_i}$$

の標本カーネル平均と同一.

重みは負の値を取り得るため，確率測度とは限らない.
「符号付き測度(signed measure)」のカーネル平均.

しかし，重み付サンプルとしての解釈が可能.

- 和は1に近づく. $\sum_{i=1}^{n} w_i \to 1 \ (n \to \infty)$
- 期待値計算可能 $\sum_{i=1}^{n} w_i f(Y_i) \to \int f(y)q(y)dy \quad (n \to \infty)$.

定理

Let $\widehat{m}_{q_y}(\cdot) = \sum_{i=1}^{n} w_i k_X(\cdot, Y_i)$ is the estimator of Kernel Sum Rule.

If $k_y(y,y) \in L^2(P_Y)$ and $\frac{\pi}{p_X} \in R\left(C_{XX}^{\beta}\right)$ $(0 \leq \beta \leq 1)$, for any P_Y-square integrable function f, with $\varepsilon_n = n^{-\max\left\{\frac{1}{4}, \frac{1}{2\beta+2}\right\}}$,

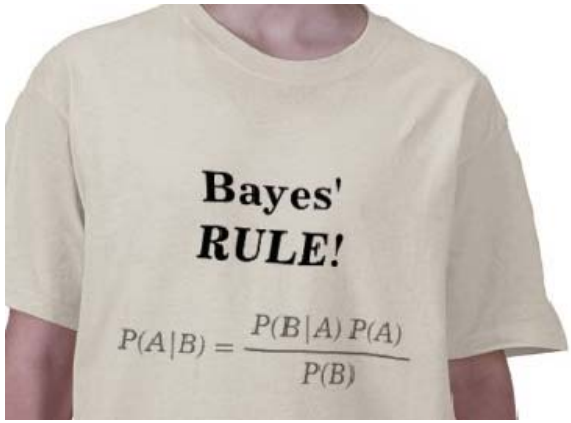
$$\sum_{i=1}^{n} w_i f(Y_i) \to \int f(y) q(y) dy, \qquad (n \to \infty)$$

with the convergence rate $O_p\left(n^{-\min\left\{\frac{1}{4}, \frac{\beta}{2\beta+2}\right\}}\right)$.

e.g.

- $f(y) = I_B(y)$: $\sum_{Y_i \in B} w_i \to$ probability $Q_y(B)$.
- 特に, $\sum_{i=1}^{n} w_i \to 1$ as $n \to \infty$.
- $f(y) = y^r$: $\sum_i w_i Y_i^r \to r$-th moment Q_y.

Kernel method for Bayesian inference



Kernel realization of Bayes' rule

- Bayes' rule

$$q(x|y) = \frac{p(y|x)\pi(x)}{q(y)}, \qquad q(y) = \int p(y|x)\pi(x)dx.$$

Π: prior with p. d. f π

$p(y|x)$: conditional probability (likelihood).

- Kernel realization:

Goal: estimate the kernel mean of the posterior

$$m_{post|y_*} := \int k_X(\cdot, x)q(x|y_*)dx$$

given

- m_Π: kernel mean of prior Π,
- C_{XX}, C_{YX}: covariance operators for $(X, Y) \sim P$,

where P is the joint probability to give $p(y|x)$ by conditioning.

Kernel Bayes' Rule: overview

- Input
  - Prior: consistent estimator of kernel mean m_Π

$$\hat{m}_\Pi = \sum_{j=1}^{\ell} \gamma_j \Phi_X(U_j)$$

$(U_1, \gamma_1), \ldots, (U_\ell, \gamma_\ell)$: weighted sample expression

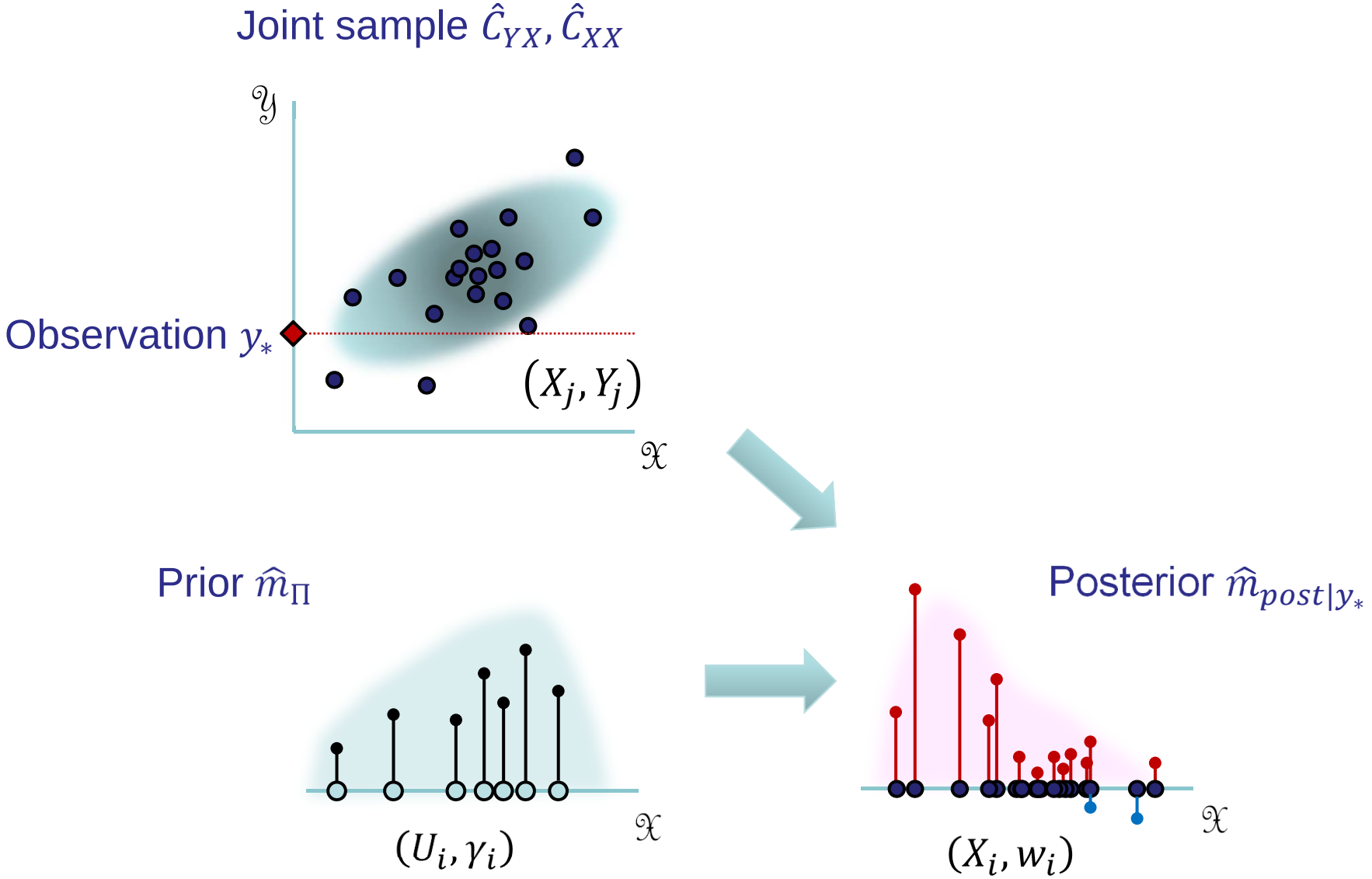
  - Conditional probability: covariance of a JOINT distribution.

$(X_1, Y_1), \ldots, (X_n, Y_n)$: (joint) sample ~ $p(x, y)$.

$p(x, y)$: conditioning gives $p(y|x)$

  - Observation: y_*
- Output
  - Posterior: $\hat{m}_{post|y_*}(\cdot) = \sum_{i=1}^{n} w_i(y_*) \Phi_X(X_i)$

weighted sample expression


Joint sample $\hat{C}_{YX}, \hat{C}_{XX}$
$\mathcal{Y}$
Observation y_*
(X_j, Y_j)
$\mathcal{X}$
Prior $\hat{m}_\Pi$
(U_i, γ_i)
$\mathcal{X}$
Posterior $\hat{m}_{post|y_*}$
(X_i, w_i)
$\mathcal{X}$

Derivation of KBR

Bayes' rule revisited

Bayes' rule consists of two steps:

1. Product: $q(x,y) = p(y|x)\pi(x)$
2. Conditioning: $q(x|y_*) = \frac{q(x,y_*)}{q(y_*)}$

We already know how to compute them!

Denote $(Z,W)\sim Q$

Note: $C_{ZW} \cong m_Q$

Product rule → $\hat{C}_{ZW} = \hat{C}_{(YX)X}\left(\hat{C}_{XX} + \varepsilon_n I\right)^{-1}\hat{m}_\pi$

Sum rule → $\hat{C}_{WW} = \hat{C}_{(YY)X}\left(\hat{C}_{XX} + \varepsilon_n I\right)^{-1}\hat{m}_\pi$

Condition → $\hat{m}_{post|y_*} = \hat{C}_{ZW}\left(\hat{C}_{WW} + \delta_n I\right)^{-1}\mathbf{k}_Y(y_*)$

Kernel Bayes' Rule

(Fukumizu, Song, Gretton JMLR2014)

Input: $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P$ (to give cond. probability).

$\hat{m}_\Pi = \sum_{j=1}^{\ell} \gamma_j \Phi_X(U_j)$ (prior) a consistent estimator of m_Π.

1. [Product]

   Compute $\Lambda = \mathrm{Diag}[(G_X/n + \varepsilon_n I_n)^{-1} G_{XU} \gamma\ ]$

2. [Conditioning]

   Compute $R_{x|y} = \Lambda G_Y \left((\Lambda G_Y)^2 + \delta_n I_n\right)^{-1} \Lambda.$

   * ε_n, δ_n: regularization coefficients

Output: estimator for kernel mean of posterior given observation y_*

$$\hat{m}_{post|y_*}(\cdot) = \mathbf{k}_X(\cdot)^T R_{x|y} \mathbf{k}_Y(y_*) = \sum_{i=1}^{n} w_i(y_*) k_X(\cdot, X_i)$$

Inference with KBR

- Weighted sample expression

$$\hat{m}_{post|y_*}(\cdot) = \sum_{i=1}^{n} w_i(y_*) k_X(\cdot, X_i)$$

Equivalent to the kernel mean of

$$\sum_{i=1}^{n} w_i(y_*) \delta_{X_i} \qquad (\delta_x\text{: Dirac's delta})$$

which is a signed measure (not necessarily a probability).

Some weights may be negative.

– $\sum_{i=1}^{n} w_i(y_*) \to 1$ in probability as $n \to \infty$.

How to use?

– Expectation: if $\frac{\pi}{p_X} \in \overline{\mathrm{Range}(C_{XX})}$ and $f \in L^2(P_X)$ satisfies $\int f(x)p(y|x)\pi(x)dx \in \mathrm{Range}(C_{YY})$,

$$\sum_{i=1}^{n} w_i(y_*)f(X_i) \to \int f(x)q(x|y_*)dx, \qquad (n \to \infty). \quad \text{(consistent)}$$

e.g.

- $f(x) = I_B(x)$: $\sum_{X_i \in B} w_i \to$ posterior prob. of set B.
- $f(x) = x^r$: $\sum_i w_i X_i^r \to r$-th moment of posterior.

(More general discussions in Kanagawa and Fukumizu, AISTATS 2014)

– Point estimation (quasi-MAP):

$$\hat{x} = \mathrm{argmin}_x \left\| \hat{m}_{post|y_*} - \Phi_X(x) \right\|_{H_X}$$

Solved numerically

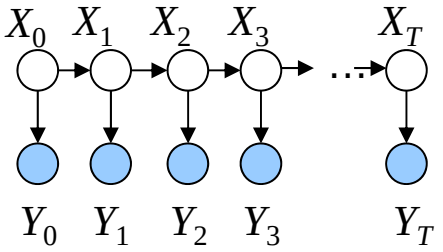
– Completely nonparametric way of computing Bayes rule.

No parametric models are needed, but data or samples are used to express the probabilistic relations.

Examples:

1. Nonparametric HMM

$$p(X,Y) = p(X_0, Y_0) \prod_{t=1}^{T} p(Y_t|X_t) q(X_t|X_{t-1})$$

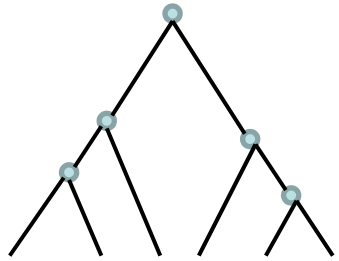


$p(Y_t|X_t)$ and /or $q(X_t|X_{t-1})$ are unknown, but data are available.

2. Explicit form of likelihood $p(y|x)$ or prior π is unavailable, but sampling is possible.

c.f. Approximate Bayesian Computation (ABC)

(Kernel ABC: Nakagome, Mano, Fukumizu 2013)



Convergence rate

Theorem (Fukumizu, Song, Gretton 2014)

Let $f \in H_X$, $(Z, W) \sim Q$ with p.d.f. $p(y|x)\pi(x)$.

Assumptions:

- $\|\widehat{m}_\Pi - m_\Pi\|_{H_X} = O_p(n^{-\alpha})$ for some $0 < \alpha \leq 1/2$.
- $\pi(x)/p_X(x) \in \mathrm{Range}\left(C_{XX}^{1/2}\right)$ for some $\beta \geq 0$.
- $E[f(Z)|W = \cdot] \in \mathrm{Range}(C_{XX}^2)$ for some $\nu \geq 0$.

Then, with $\varepsilon_n = n^{-2\alpha/3}$ and $\delta_n = n^{-8\alpha/27}$, for any y_*,

$$\sum_{i=1}^{n} w_i(y_*) f(X_i) - E[f(Z)|W = y_*] = O_p(n^{-\frac{8\alpha}{27}}) \qquad (n \to \infty).$$

- Remark: the rate depending on the smoothness of the functions π/p_X and $E[f(Z)|W = \cdot]$ is also available.
- If $\alpha = 1/2$, the rate is $n^{-4/27}$.

Choice of kernel and hyperparameter

- Parameters to be chosen
  - Kernel (parameters in kernel)
  - Regularization parameter
- Cross-validation is recommended, if possible.
  - Straightforward in supervised setting (incl. nonparam. HMM).
  - Make a relevant supervised problem and apply.

  Supports
  - CV has been used successfully for SVM.
  - The rate $O_p(n^{-\frac{2\alpha}{2\alpha+m}+\rho})$ for the regression is attained with parameter choice by validation (Eberts & Steinwart 2011).
- Heuristics:

  $\sigma = \text{Median}\{\|X_i - X_j\| \mid i \neq j\}$ for σ in Gaussian kernel.

Basic experiments

– Comparison with KDE + Importance Weighting

- $(X_i, Y_i) \sim N(\left(0_{d/2}, \mathbf{1}_{d/2}\right)^T, V), \quad i = 1, \dots, N$
  $V \sim A^T A + 2I_d, \; A \sim N(0, I_d)$

- Prior Π: $U_j \sim N(0; 0.5 * V_X), \quad j = 1, \dots, L$

- KDE + Importance Weighting

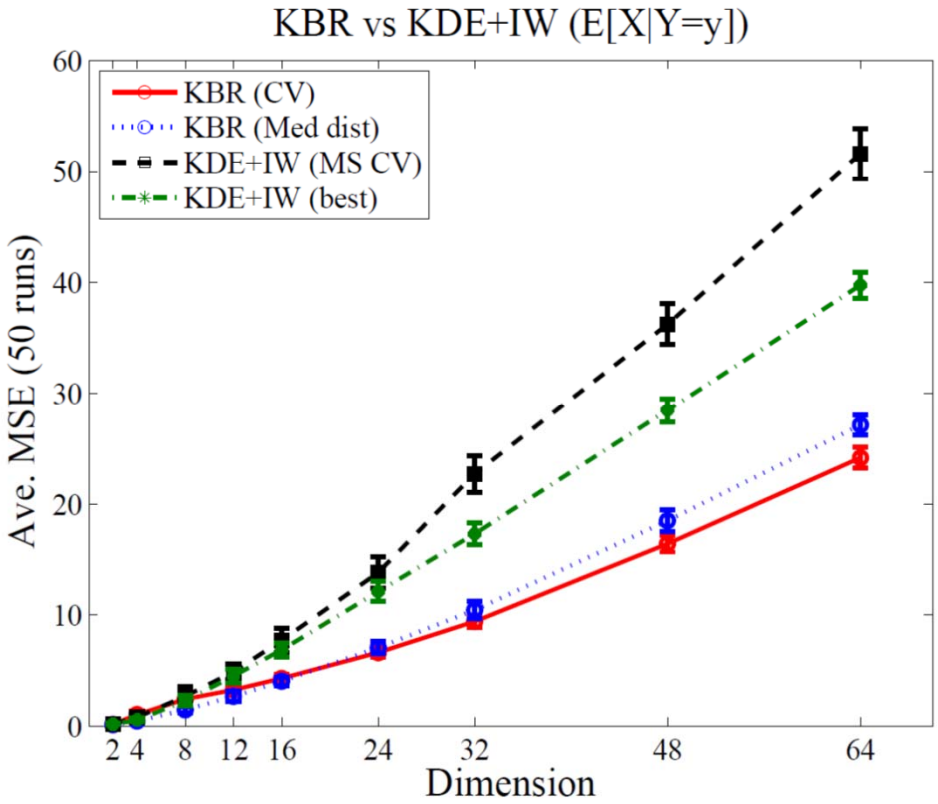
  KDE: $\hat{p}(y|x) = \frac{\sum_i K_{h_X}^X (x - X_i) K_{h_Y}^Y (y - Y_i)}{\sum_j K_{h_X}^X (x - X_j)}$,

  IW: $q(x|y)$ is estimated by weighted sample (U_j, γ_j),

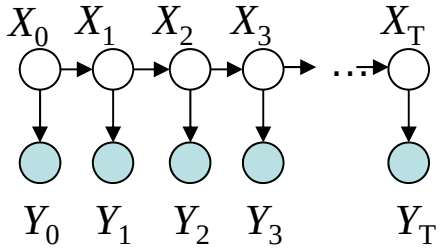
  $$\gamma_j = \frac{\hat{p}(y|U_j)}{\sum_{\ell=1}^{L} \hat{p}(y|U_\ell)}.$$

- N (sample size for (X, Y)) $= L$ (sample size for prior) $= 200$.
- $d = 2, \ldots, 64$

- Gaussian kernels are used for both methods.
- Bandwidth parameters are selected with CV for both methods.



Example: KBR for nonparametric HMM



- Assume:
  $p(y_t|x_t)$ and/or $q(x_t|x_{t-1})$ is not known.
  But, data $(X_t, Y_t)_{t=0}^{T}$ is available
  in training phase.

  Examples:
  - Measurement of hidden states is expensive,
  - Hidden states are measured with time delay.

- Testing phase (e.g., filtering, e.g.):

  given $\tilde{y}_0, \ldots, \tilde{y}_t$, estimate hidden state x_s.

  →KBR point estimator: $\operatorname{argmin}_{x_s} \left\| \hat{m}_{x_s|\tilde{y}_0,\ldots,\tilde{y}_t} - \Phi(x) \right\|_{H_X}$

- General sequential inference uses Bayes' rule ➔ KBR applied.

Numerical examples

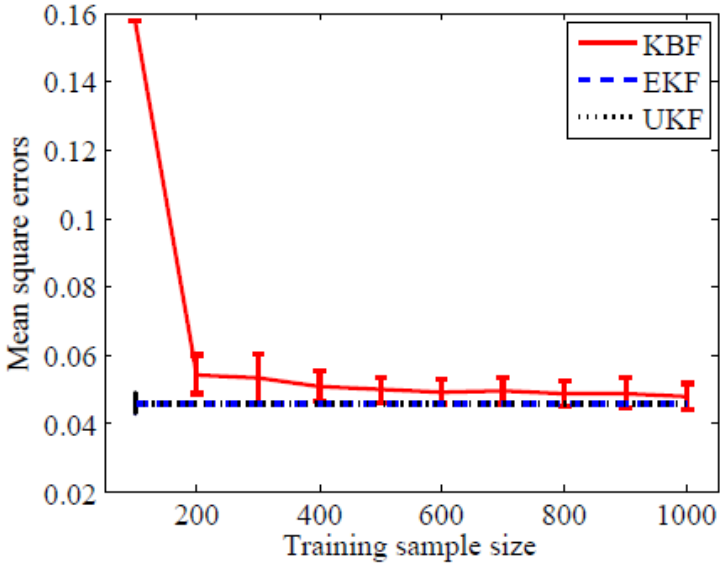
(a) Noisy rotation

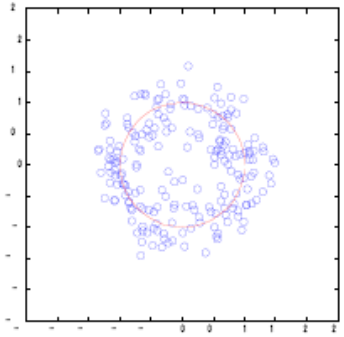
$$\begin{pmatrix} u_t \\ v_t \end{pmatrix} = \begin{pmatrix} \cos(\theta_t) \\ \sin(\theta_t) \end{pmatrix} + Z_t, \qquad \theta_{t+1} = \arctan\left(\frac{v_t}{u_t}\right) + 0.3,$$

$$Y_t = (u_t, v_t)^T + W_t,$$

$$Z_t, W_t \sim N(0, 0.04 I_2)\ (i.i.d.)$$

Filtering with the point estimator by KBR.





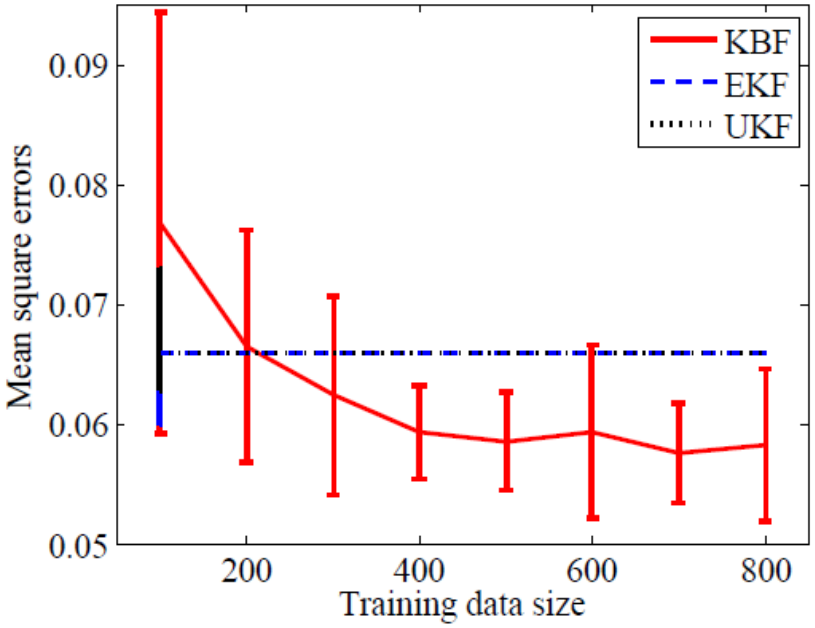
KBR does NOT know the dynamics, while the EKF and UKF use it.

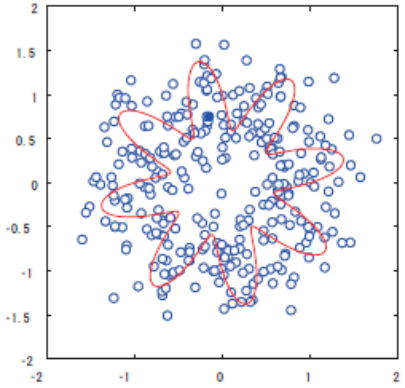
(b) Noisy oscillation

$$\begin{pmatrix} u_t \\ v_t \end{pmatrix} = (1 + 0.4\sin(8\theta_t)\begin{pmatrix} \cos(\theta_t) \\ \sin(\theta_t) \end{pmatrix} + Z_t, \qquad \theta_{t+1} = \arctan\left(\frac{v_t}{u_t}\right) + 0.4,$$

$$Y_t = (u_t, v_t)^T + W_t,$$

$$Z_t, W_t \sim N(0, 0.04I_2)\ (i.i.d.)$$





Camera angles

- Hidden X_t: angles of a video camera located at a corner of a room.
- Observed Y_t: movie frame of a room + additive Gaussian noise.
- X_t: 3600 downsampled frames of 20 x 20 RGB pixels (1200 dim.).
- The first 1800 frames for training, and the second half for testing.

noise	KBR (Trace)	Kalman filter(Q)
$\sigma^2 = 10^{-4}$	$0.15 \pm < 0.01$	0.56 ± 0.02
$\sigma^2 = 10^{-3}$	0.21 ± 0.01	0.54 ± 0.02

Average MSE for camera angles (10 runs)

To represent SO(3) model, $\mathrm{Tr}[AB^{-1}]$ for KBR, and quaternion expression for Kalman filter are used .

Robot localization

COLD (COsy Localization Dataset, IJRR 2009)

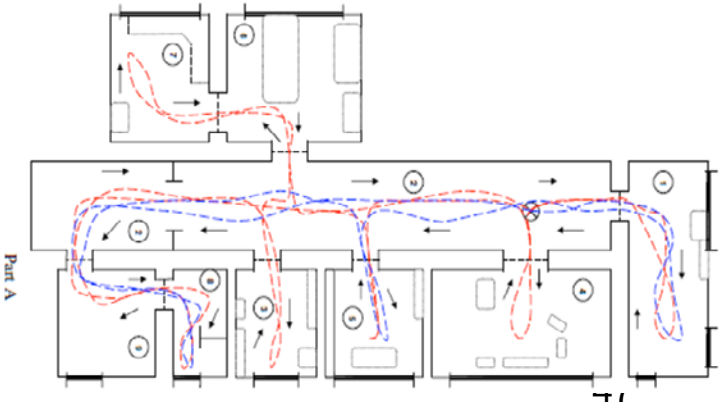
State $X_t \in \mathbf{R}^3$: 2-D coordinate and orientation of a robot

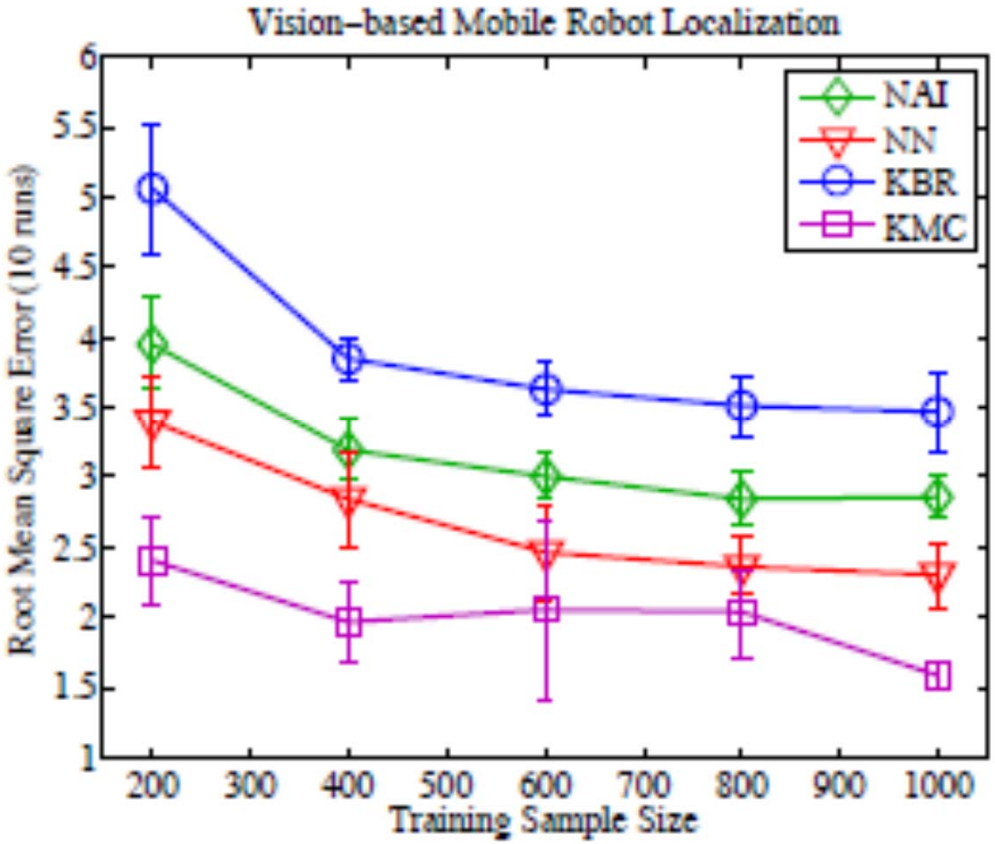
Observation Y_t : image sequence（SIFT feature, 4200dim）

Training sample $(X_t, Y_t) : t = 1, \ldots, T$

Estimate the location of a robot from image sequences

- Observation: $p(Y_t|X_t)$ difficult to model.
  → KBR
- State transition: linear Gaussian
  Kernel Monte Carlo,
  (Kanagawa, Nishiyama, KF. 2013)



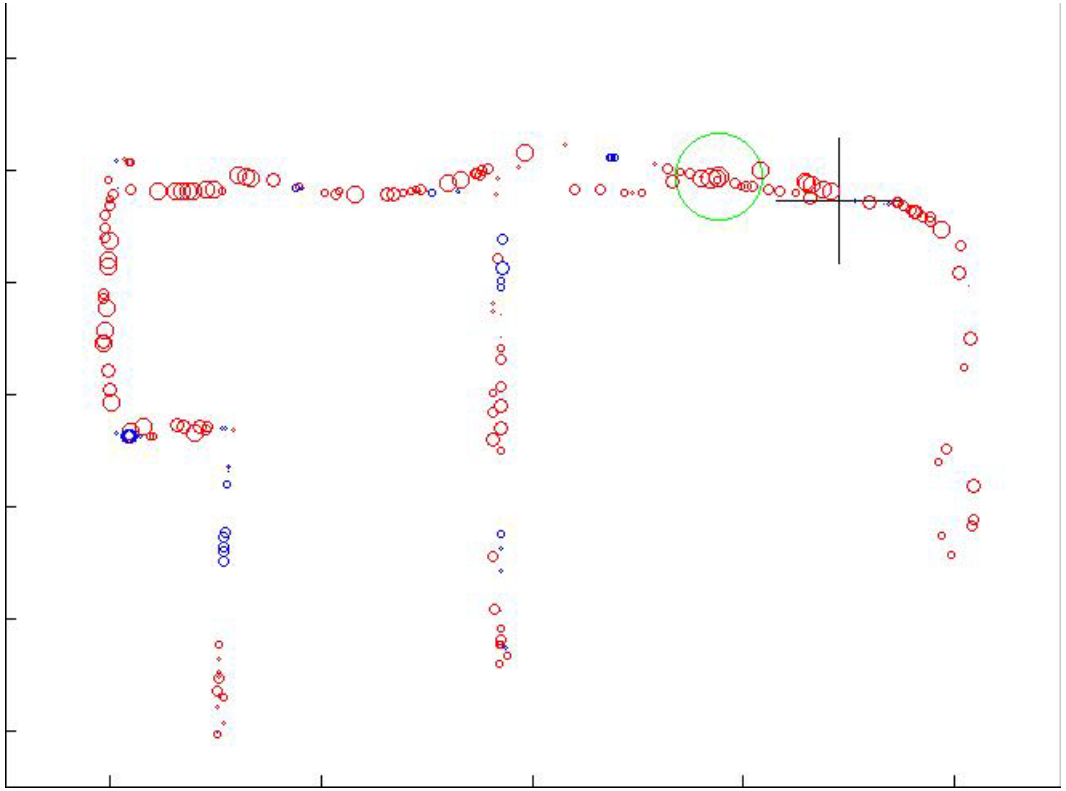


NAI: naïve method (closest image in training data)

NN: PF + K-nearest neighbor (Vlassis, Terwijn, Kröse 2002)

KBR: KBR + KBR

KMC: KBR + Monte Carlo



training sample
= 200

+: true location

◯ : estimate

red (+)/ blue (-) circles: weights on the training sample

Conclusions

- A new nonparametric / kernel approach to Bayesian inference

  - Kernel mean embedding: using positive definite kernels to
  - “Nonparametric” Bayesian inference
    - No densities are needed but data.
  - Bayesian inference with matrix computation.
    - Computation is done with Gram matrices.
    - No integral, no approximate inference.
  - More suitable for high dimensional data than smoothing kernel approach.

Kernel Herding

Kernel Herding

(Cheng, Welling, Smola 2010)

– 目的: カーネル平均 $m_X := E[\Phi(X)]$ を

$$E[\Phi(X)] \approx \frac{1}{T}\sum_{t=1}^{T}\Phi(x_t)$$

の形によって効率的に近似するサンプル $x_1, \ldots, x_T$ を求める.

– 仮定: m_X は知っているとする.

– アルゴリズム:

次の更新則で $x_1, x_2, \ldots$ を生成

$$x_{t+1} = \arg\max_x \langle h_t, \Phi(x)\rangle$$

$$h_{t+1} = h_t + m_X - \Phi(x_{t+1})$$

■ 解釈

$\|\Phi(x)\| = R$（定数）を仮定（平行移動不変なカーネル）.

$\tilde{h}_t := h_t/t$ とおく. $\tilde{h}_0 = m_X$ とする.

$$x_{t+1} = \arg\max_x \langle \tilde{h}_t, \Phi(x) \rangle$$

$$(\mathrm{t}+1)\tilde{h}_{t+1} = t\tilde{h}_t + m_X - \Phi(x_{t+1})$$

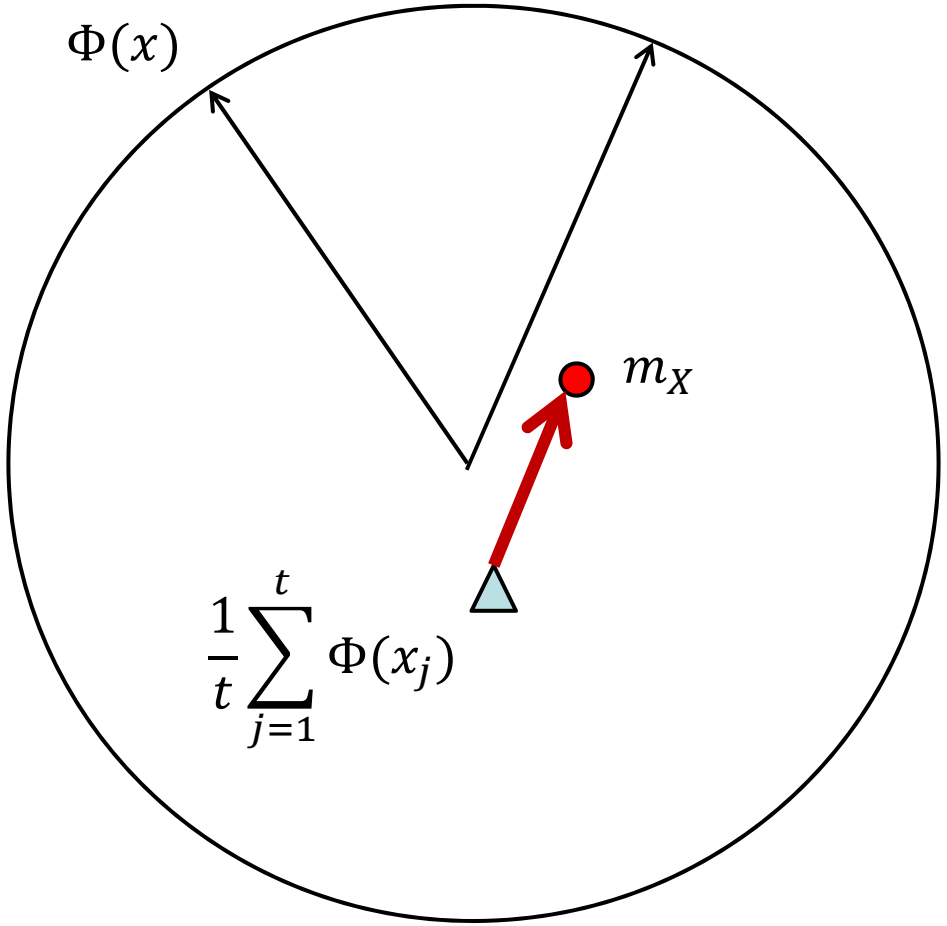
$$x_1 = \arg\max_x m_X(x), \qquad \tilde{h}_1 = m_X - \Phi(x_1)$$

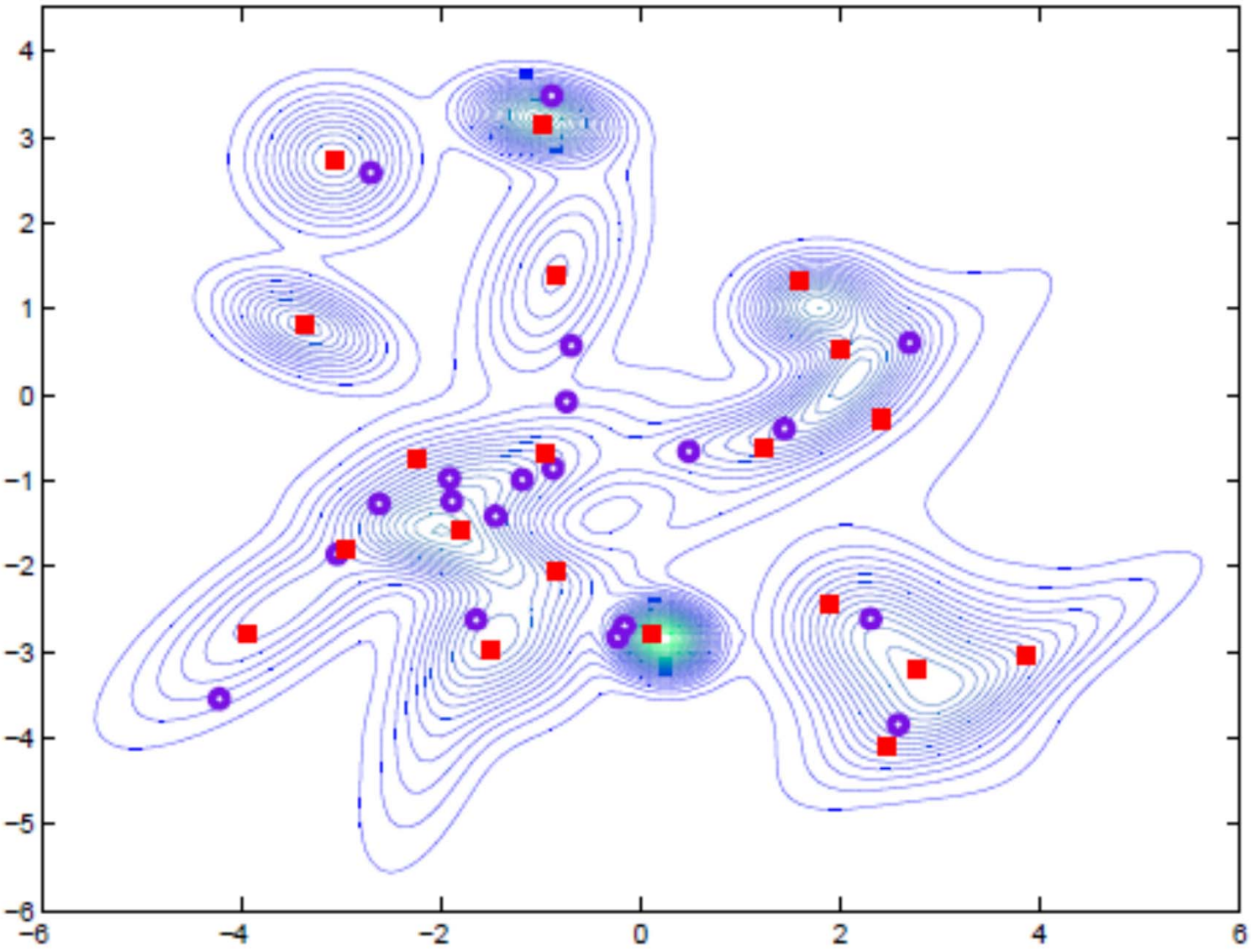
$$x_2 = \arg\max_x \langle \tilde{h}_1, \Phi(x) \rangle, \quad \tilde{h}_2 = m_X - \frac{1}{2}\sum_{j=1}^{2} \Phi(x_j)$$

$$\vdots \qquad\qquad \vdots$$

$$x_{t+1} = \arg\max_x \langle \tilde{h}_t, \Phi(x) \rangle, \qquad \tilde{h}_{t+1} = m_X - \frac{1}{t}\sum_{j=1}^{t} \Phi(x_j)$$

残差の$\Phi(x)$への射影　　近似の残差


$\Phi(x)$
m_X
$\frac{1}{t}\sum_{j=1}^{t}\Phi(x_j)$

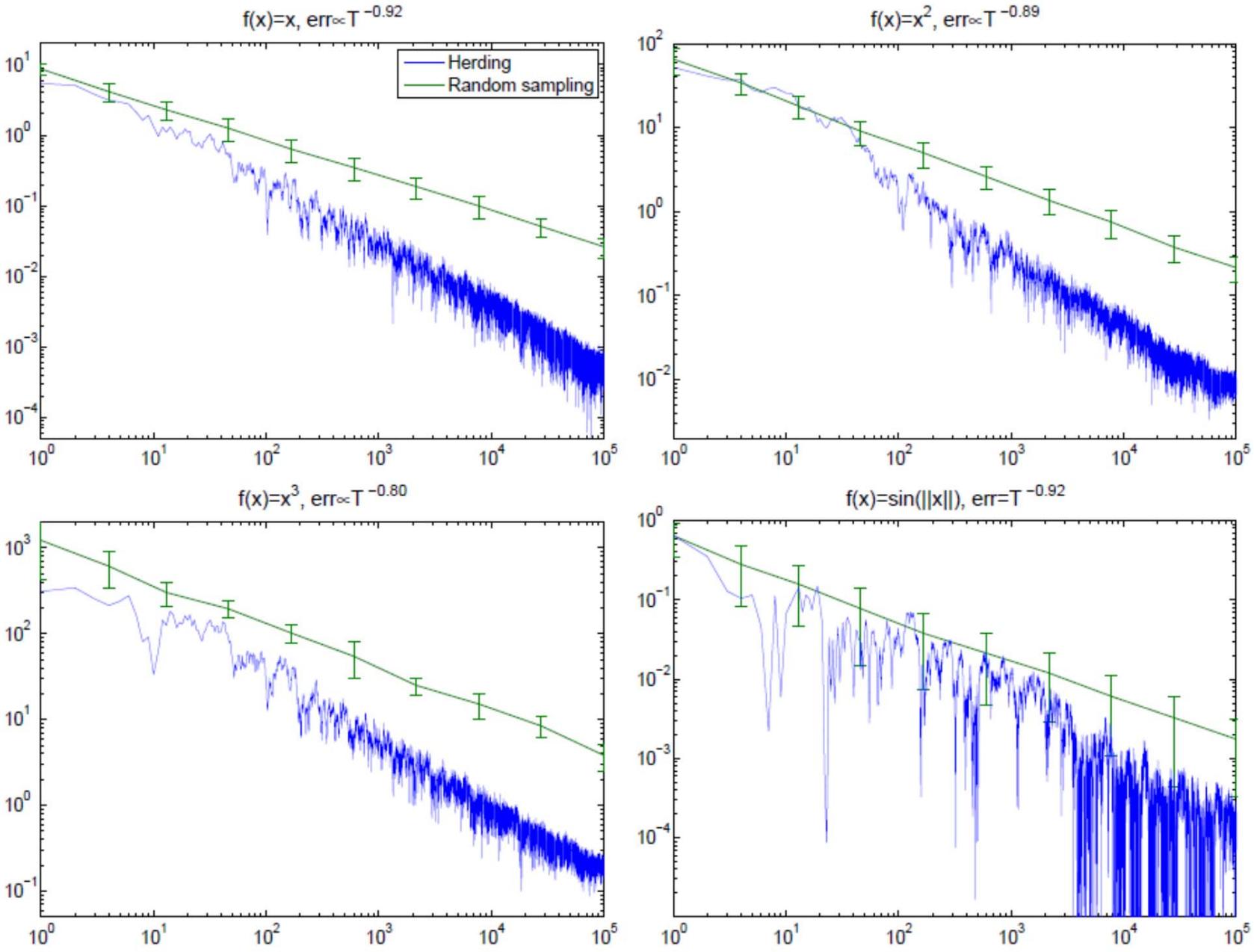


■ Kernel herdingによる最初の20個のサンプル

● X の分布から発生した20個の i.i.d サンプル

– $E[\Phi(X)]$を知らない場合でも，多数のサンプルで$E[\Phi(X)]$が近似されているときに使ってもよい.
少数サンプルで近似すると，グラム行列のサイズを下げることができる
→ カーネル法の計算効率化

– 収束性:
  - 近似誤差 $\left\| m_X - \frac{1}{\mathrm{t}} \sum_{j=1}^{t} \Phi(x_t) \right\|$ は $O(1/t)$ であると予想されている.
  - Optimal な収束レートである.
    c.f. i.i.d. サンプルの場合 $O(1/\sqrt{t})$


f(x)=x, err∝T^-0.92
Herding
Random sampling
f(x)=x^2, err∝T^-0.89
f(x)=x^3, err∝T^-0.80
f(x)=sin(||x||), err=T^-0.92

References

- Fukumizu, K., L. Song, A. Gretton (2014) Kernel Bayes' Rule: Bayesian Inference with Positive Definite Kernels. *Journal of Machine Learning Research. 14:3753−3783.*
- Song, L., Gretton, A., and Fukumizu, K. (2013) Kernel Embeddings of Conditional Distributions. *IEEE Signal Processing Magazine 30(4), 98-111*
- Kanagawa, M., Nishiyama, Y., Gretton, A., Fukumizu. K. (2013) Kernel Monte Carlo Filter. arXiv:1312.4664
- Cheng, Y., M. Welling, A. Smola (2010) Super-samples from Kernel Herding. Proc. Uncertainty in Artificial Intelligence 2010.