如何利用贝叶斯采样器处理(拥抱)不确定性

刘斌
南京邮电大学计算机学院

2017-11-02 @ 华东师大中国R会

不确定性(Uncertainty)

概率(Probability)

“不确定性”的来源

- 世界运转的规律(规则）：有可能就是随机的
- 不可知(或尚未可知)的因素
- 观测噪声

物理概率（Physical Probability）

- 频率论（反主观论）：
  - $P(h) = F(h)$
  - 极限频率（limit frequency）

物理概率（Physical Probability）

- 频率论（反主观论）：
  - $P(h) = F(h)$
  - 极限频率（limit frequency）
    - “乔丹命中下一球的概率为50%” 是指 。。。 ？

物理概率（Physical Probability）

- 频率论（反主观论）：
  - $P(h) = F(h)$
  - 极限频率（limit frequency）
    - “乔丹命中下一球的概率为50%” 是指 。。。 ？

此“概率”更适于分析物理实验，不会因主观意见而“让步”

主观概率（Subjective Probability）

- 相信的程度（Degree of belief）
  - 允许融入“主观意见”
  - 假定这个世界是确定性的（所有物理概率要么为0，要么为1），下述说法依然合理：
    - 乔丹命中下一球的概率为50%

连接主观概率与物理概率

- Principal Principle (David Lewis (1941–2001)，20世纪最重要的哲学家之一)：

  $P(h \mid Ch(h) = r) = r$

- 理论本身也是“概率”产生的源头之一

基于概率论进行不确定性推理

如何由“所见所闻”推“所知”：

三类重要的“量”：

- 观测量
- 未知量
- 辅助变量

给定观测量，求未知量的概率

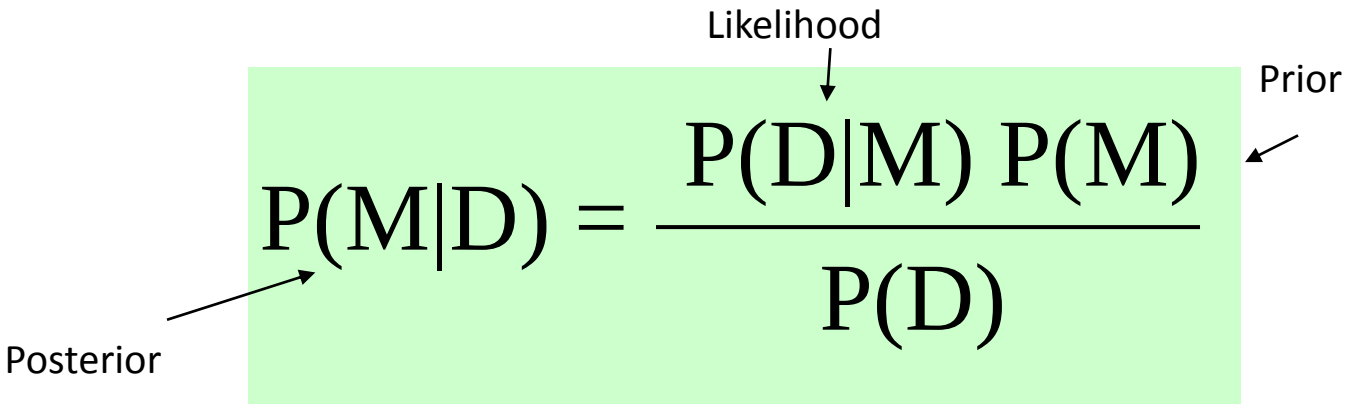
基于概率论进行不确定性推理

"Probability theory is nothing more than common sense reduced to calculation."

- Pierre-Simon Laplace, 1814

基于概率论进行不确定性推理

- 贝叶斯公式：



贝叶斯推理:一个小例子

- 一个现象已持续了一段时间t_{past}. 它一共将会持续多长时间? (也就是t_{total}?)

- 可将此处的“时间” 替换成任意连续取值的其它量（取值范围为0~未知上限）

贝叶斯推理:一个小例子

$$P(t_{total} \mid t_{past}) \propto P(t_{past} \mid t_{total})\, P(t_{total})$$

posterior probability

likelihood

prior

$$\propto \quad 1/t_{total} \qquad 1/t_{total}$$

Assume random sample

$(0 < t_{past} < t_{total})$

"Uninformative" prior

贝叶斯推理:一个小例子

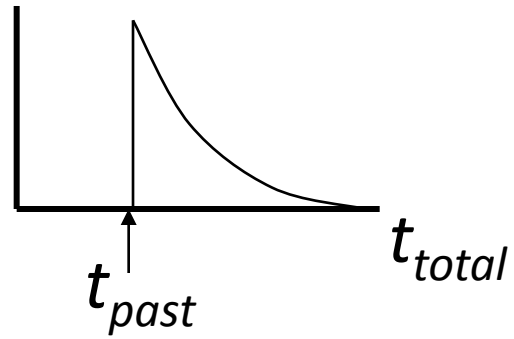
$$P(t_{total} \mid t_{past}) \propto 1/t_{total} \quad 1/t_{total}$$

posterior probability

Random sampling

"Uninformative" prior

$P(t_{total} \mid t_{past})$



贝叶斯推理:一个小例子

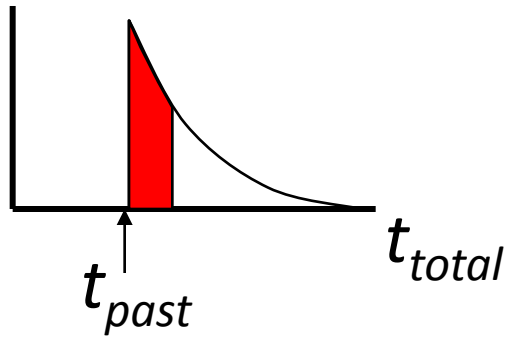
$$P(t_{total} \mid t_{past}) \propto 1/t_{total} \quad 1/t_{total}$$

posterior probability

Random sampling

"Uninformative" prior

$P(t_{total} \mid t_{past})$



对t_{total}的最佳猜测: 满足$P(t_{total} > t \mid t_{past}) = 0.5$的$t$

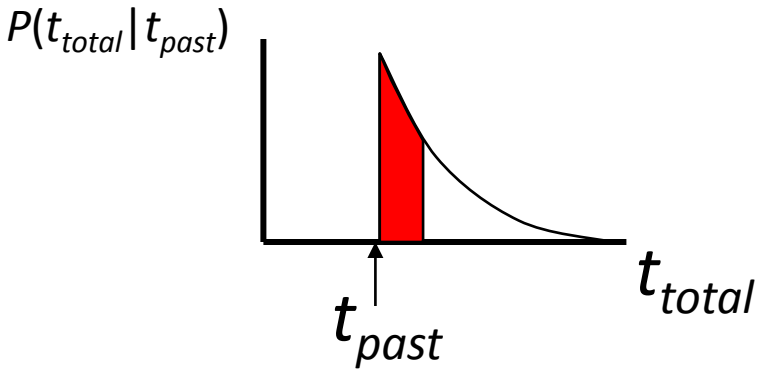
贝叶斯推理:一个小例子

$$P(t_{total} \mid t_{past}) \propto 1/t_{total} \quad 1/t_{total}$$

posterior probability

Random sampling

"Uninformative" prior



对t_{total}的最佳猜测: 满足$P(t_{total} > t \mid t_{past}) = 0.5$的$t$

即 $t_{total} = 2t_{past}$

贝叶斯推理:一个小例子

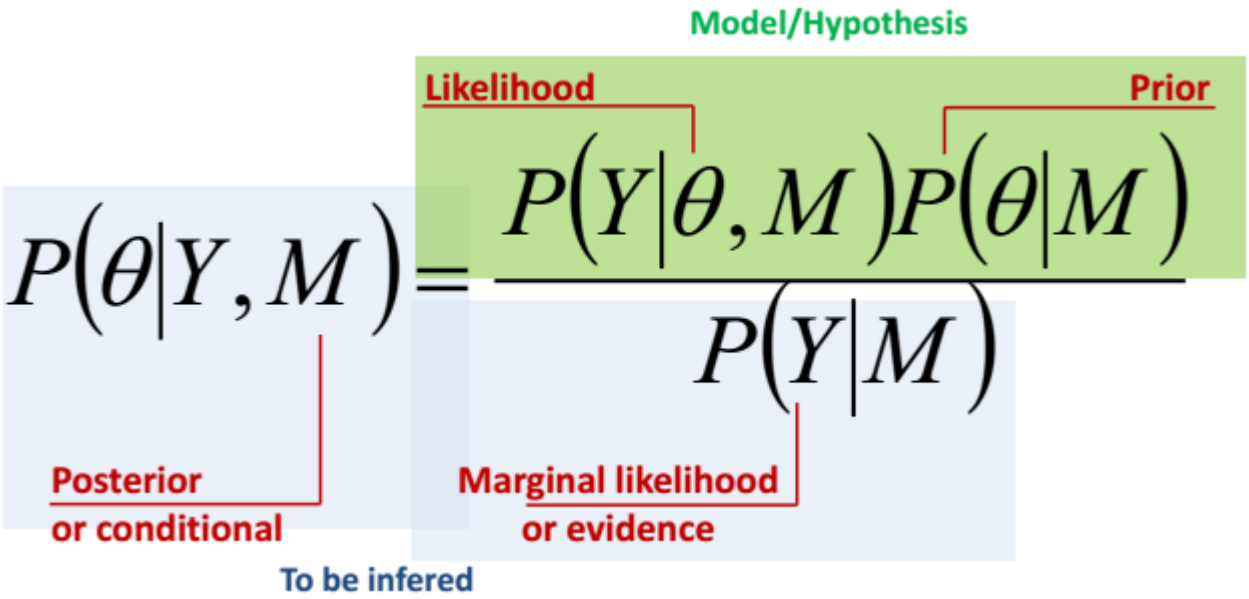
- 当你得知一部电影的票房已达5亿元，其最终票房会是多少?

贝叶斯推理:一个小例子

- 当你得知一部电影的票房已达5亿元，其最终票房会是多少?
  - “10亿元” 是个合理估计

利用贝叶斯应对不确定性

- 模型M确定，模型参数不确定



利用贝叶斯应对不确定性

- 模型不确定，模型参数也不确定

$$P\left(\theta \mid Y, M_i\right) = \frac{P\left(Y \mid \theta, M_i\right) P\left(\theta \mid M_i\right)}{P\left(Y \mid M_i\right)}$$

利用贝叶斯应对不确定性

- 模型不确定，模型参数也不确定

$$P(\theta|Y,M_i)=\frac{P(Y|\theta,M_i)P(\theta|M_i)}{P(Y|M_i)}$$

$$P(\theta|Y)=\sum_i P(\theta|M_i,Y)P(M_i|Y)$$

利用贝叶斯应对不确定性

- 模型不确定，模型参数也不确定

$$P(\theta|Y,M_i)=\frac{P(Y|\theta,M_i)P(\theta|M_i)}{P(Y|M_i)}$$

$$P(\theta|Y)=\sum_i P(\theta|M_i,Y)P(M_i|Y)$$

$$P(M_i|Y)=\frac{P(Y|M_i)P(M_i)}{\sum_i P(Y|M_i)P(M_i)}$$

利用贝叶斯应对不确定性

- 模型不确定，模型参数也不确定

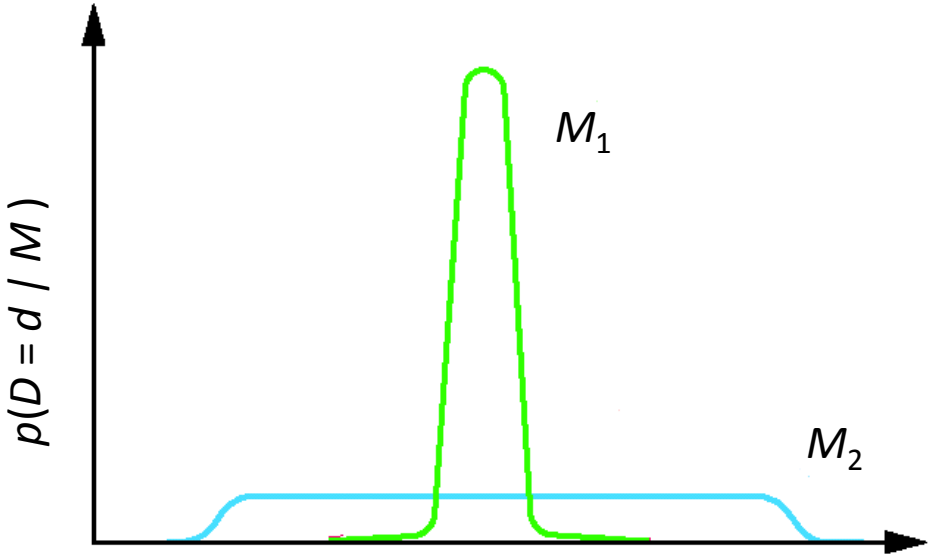
$$P(\theta|Y,M_i)=\frac{P(Y|\theta,M_i)P(\theta|M_i)}{P(Y|M_i)}$$

$$P(\theta|Y)=\sum_i P(\theta|M_i,Y)P(M_i|Y)$$

$$P(M_i|Y)=\frac{P(Y|M_i)P(M_i)}{\sum_i P(Y|M_i)P(M_i)}$$

$$P(Y|M_i)=\int P(Y|\theta,M_i)P(\theta|M_i)d\theta$$

奥卡姆剃刀

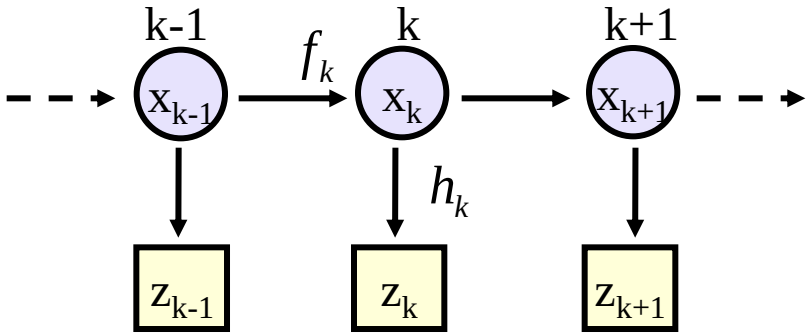


(MacKay, 2003; Ghahramani tutorials)

For any model M, $\sum_{\text{all } d \in D} p(D = d \mid M) = 1$

Law of "conservation of belief": A model that can predict many possible data sets must assign each of them low probability.

动态不确定性



Stochastic diffusion

State equation: $x_k = f_k(x_{k-1}, v_k)$

x_k state vector at time instant k

f_k state transition function, $f_k : R^{N_x} \times R^{N_v} \rightarrow R^{N_x}$

v_k i.i.d process noise

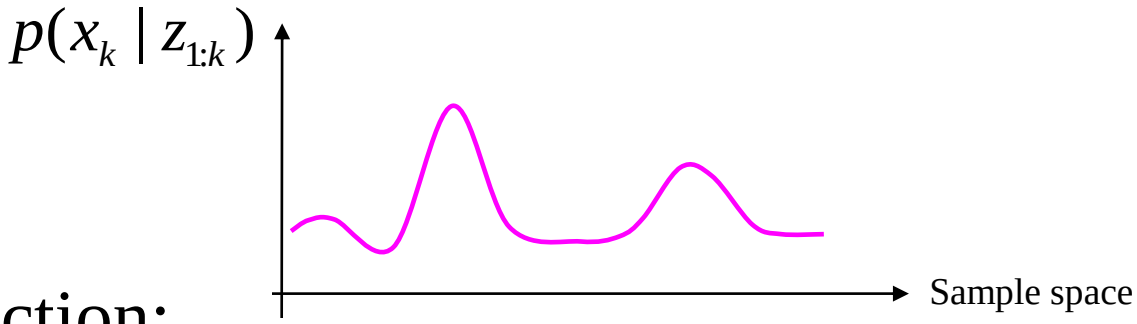
Observation equation: $z_k = h_k(x_k, w_k)$

z_k observations at time instant k

h_k observation function, $h_k : R^{N_x} \times R^{N_w} \rightarrow R^{N_z}$

w_k i.i.d measurement noise

迭代贝叶斯滤波



- Prediction:

$$p(x_k \mid z_{1:k-1}) = \int p(x_k \mid x_{k-1}) p(x_{k-1} \mid z_{1:k-1}) dx_{k-1} \quad (1)$$

- Update:

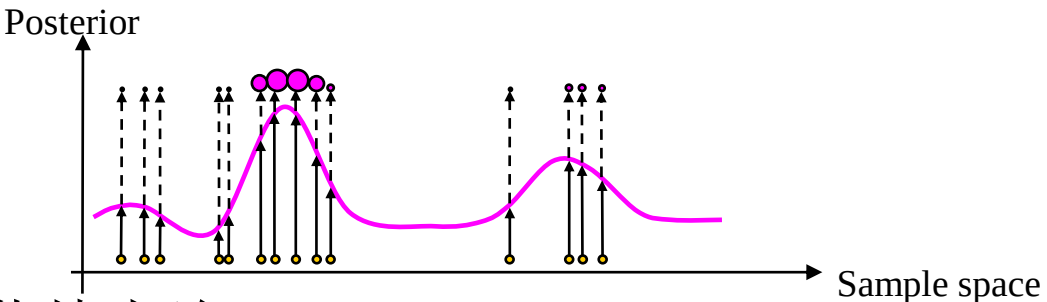
$$p(x_k \mid z_{1:k}) = \frac{p(z_k \mid x_k) p(x_k \mid z_{1:k-1})}{p(z_k \mid z_{1:k-1})} \quad (2)$$

$$p(z_k \mid z_{1:k-1}) = \int p(z_k \mid x_k) p(x_k \mid z_{1:k-1}) dx_k$$

贝叶斯采样之序列蒙特卡洛

- 很多变种，同一理念:

> 利用随机抽取的加权样本集（粒子）表示后验概率密度函数



- 随机抽取= 蒙特卡洛 (Monte Carlo , MC)
- 当样本数目变得极大时 –该样本集成为真正后验概率密度函数的等价表示

贝叶斯采样之序列蒙特卡洛

- Compared to other methods
  - Can represent any arbitrary distribution
    - multimodal support
  - Keep track of many hypotheses as there are particles
  - **Approximate representation of complex model rather than exact representation of simplified model**

- The basic building-block: *Importance Sampling*

重要性采样(Importance Sampling)

- Evaluate complex integrals using probabilistic techniques
- Assume we are trying to estimate a complicated integral of a function *f* over some domain *D*: $F = \int_D f(\vec{x}) d\vec{x}$

- Also assume there exists some PDF *p* defined over *D*

重要性采样

- Then

$$F = \int_D f(\vec{x})d\vec{x} = \int_D \frac{f(\vec{x})}{p(\vec{x})} p(\vec{x})d\vec{x}$$

- But

$$\int_D \frac{f(\vec{x})}{p(\vec{x})} p(\vec{x})d\vec{x} = E\left[\frac{f(\vec{x})}{p(\vec{x})}\right], x \sim p$$

- This is true for <u>any</u> PDF p over D!

重要性采样

- Now, if we have i.i.d random samples $\vec{x}_1,...,\vec{x}_N$ sampled from p, then we can approximate $E\left[\frac{f(\vec{x})}{p(\vec{x})}\right]$ by

$$F_N = \frac{1}{N}\sum_{i=1}^{N}\frac{f(\vec{x}_i)}{p(\vec{x}_i)}$$

- Guaranteed by law of large numbers:

$$N \to \infty, F_N \overset{a.s}{\to} E\left[\frac{f(\vec{x})}{p(\vec{x})}\right] = F$$

重要性采样

- What about $p(\vec{x}) = 0$?
- If p is very small, f / p can be arbitrarily large, 'damaging' the average
  - Design p such that f / p is bounded
  - Rule of thumb: take p similar to f as possible

Importance weights

***Importance* or *proposal* density**

- The effect: get more samples in 'important' areas of f, i.e. where f is large

Sequential Importance Sampling (SIS)

$$E(f(X)) = \int_X f(x_{0:k}) p(x_{0:k} \mid z_{1:k}) dx_{0:k}$$

$$= \int_X f(x_{0:k}) \frac{p(x_{0:k} \mid z_{1:k})}{q(x_{0:k} \mid z_{1:k})} q(x_{0:k} \mid z_{1:k}) dx_{0:k}$$

- We characterize the posterior pdf using a set of samples (particles) and their weights

$$\{x_{0:k}^i, w_k^i\}_{i=1}^N$$

- Then the joint posterior density at time k is approximated by

$$p(x_{0:k} \mid z_{1:k}) \approx \sum_{i=1}^N w_k^i \delta(x_{0:k} - x_{0:k}^i)$$

SIS

- We draw the samples from the importance density $q(x_{0:k} \mid z_{1:k})$ with importance weights

$$w_k^i \propto \frac{p(x_{0:k} \mid z_{1:k})}{q(x_{0:k} \mid z_{1:k})}$$

- Sequential update (after some calculation…)

Particle update

$$x_k^i \sim q(x_k \mid x_{k-1}^i, z_k)$$

Weight update

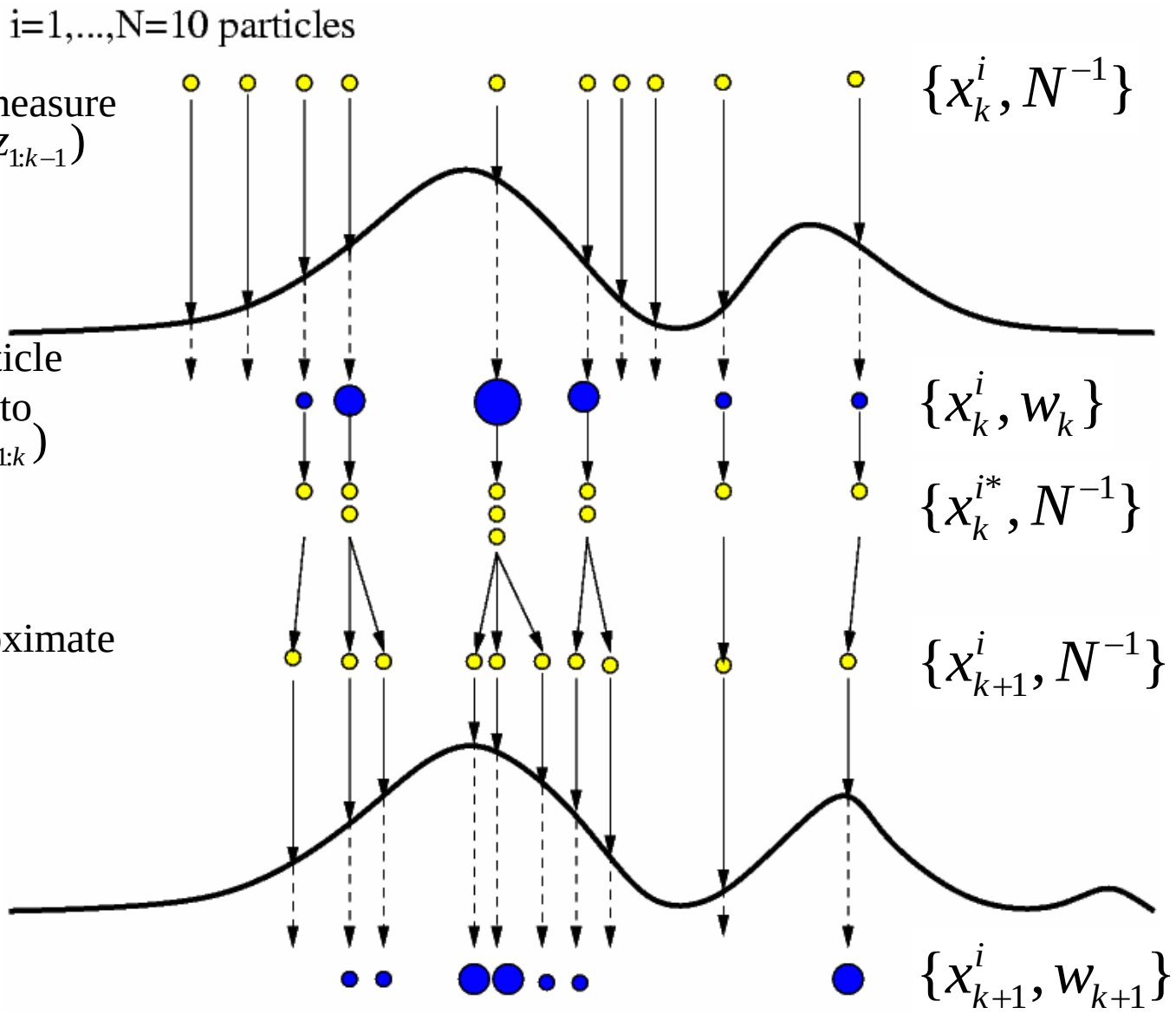
$$w_k^i = w_{k-1}^i \frac{p(z_k \mid x_k^i)\, p(x_k^i \mid x_{k-1}^i)}{q(x_k^i \mid x_{k-1}^i, z_k)}$$

SIS

$$\left[\{x_k^i, w_k^i\}_{i=1}^N\right] = \mathrm{SIS}\left[\{x_{k-1}^i, w_{k-1}^i\}_{i=1}^N, z_k\right]$$

- FOR i=1:N
  - Draw $x_k^i \sim q(x_k \mid x_{k-1}^i, z_k)$
  - Update weights $w_k^i = w_{k-1}^i \dfrac{p(z_k \mid x_k^i)\, p(x_k^i \mid x_{k-1}^i)}{q(x_k^i \mid x_{k-1}^i, z_k)}$
- END
- Normalize weights

序列蒙特卡洛



详细内容请参见:

- *Liu*, *B.*, Robust Particle Filter by Dynamic Averaging of Multiple Noise Models, Proc. of the 42nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017), pp.4034-4038, 2017.
- Dai, Y., *Liu*, *B.*, Robust video object tracking via Bayesian model averaging based feature fusion, Optical Engineering, vol.55, no.8, pp.083102(1-11), 2016.
- *Liu*, *B.*, Adaptive Annealed Importance Sampling for Multimodal Posterior Exploration and Model Selection with Application to Extrasolar Planet Detection, The Astrophysical Journal Supplement Series, vol. 213, no. 14, pp. 1-16, 2014.
- *Liu*, *B.*, Instantaneous Frequency Tracking under Model Uncertainty via Dynamic Model Averaging and Particle Filtering, IEEE Trans. on Wireless Communications, vol.10, no.6, pp.1810-1819,2011.

“拥抱”不确定性？

$$\max_{x \in \chi} f(x)$$

χ denotes a non-empty solution space in $\mathbb{R}^n$

$f: \chi \rightarrow \mathbb{R}$ is a continuous real-valued function

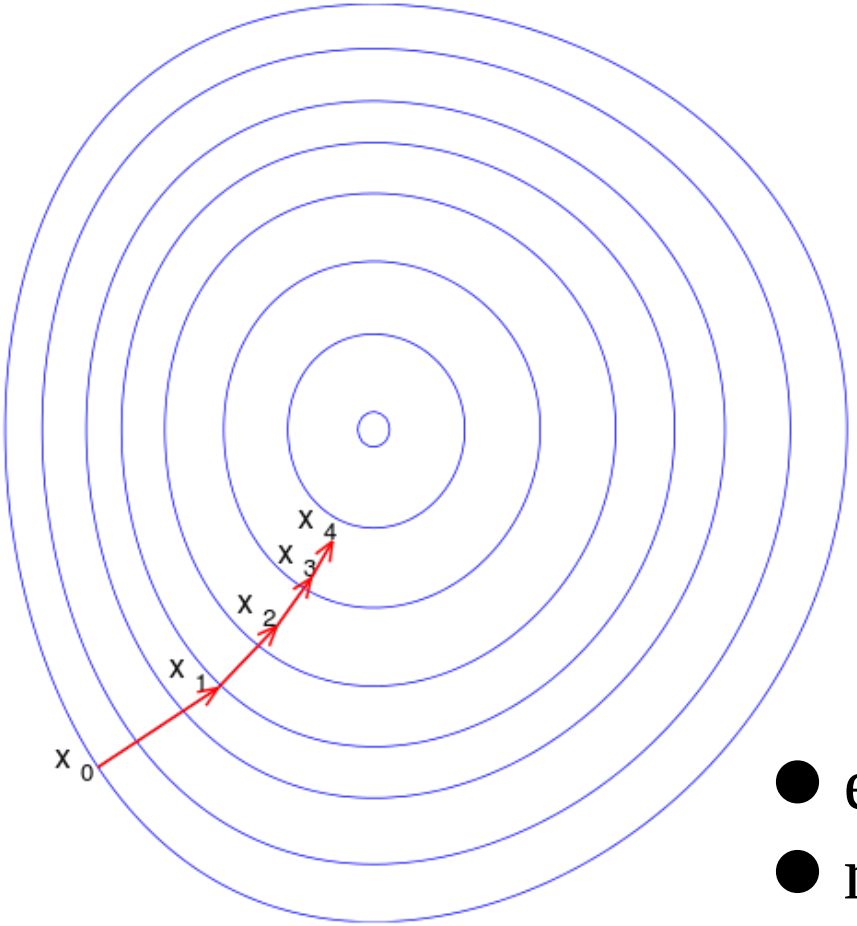
f is bounded on χ, which means $\exists f_l > -\infty$,

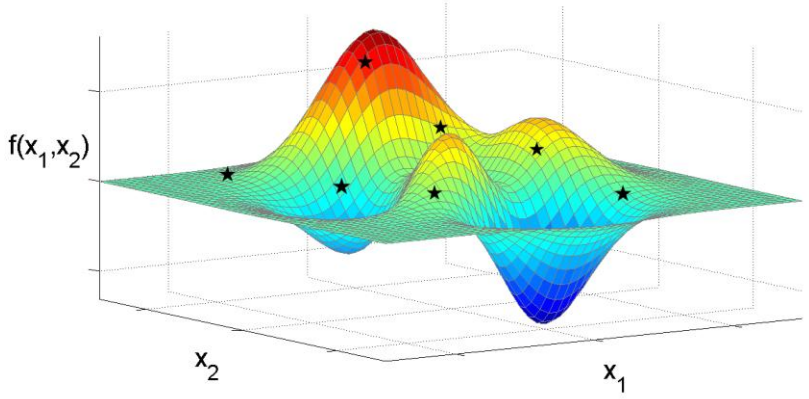
$f_u < \infty$ such that $f_l \leq f(x) \leq f_u, \forall x \in \chi$

denote the maximal function value as f^*,

there exists a x^* such that $f(x) \leq f^* \triangleq f(x^*), \forall x \in \chi$

Gradient descent





- easy to trap into local optimum
- requires gradient information

利用贝叶斯采样器搜索全局最优点

- Specify / design a series of target pdfs
- Specify / design a series of proposal pdfs
- Run SMC and assess the global optimum from the yielded samples / particles

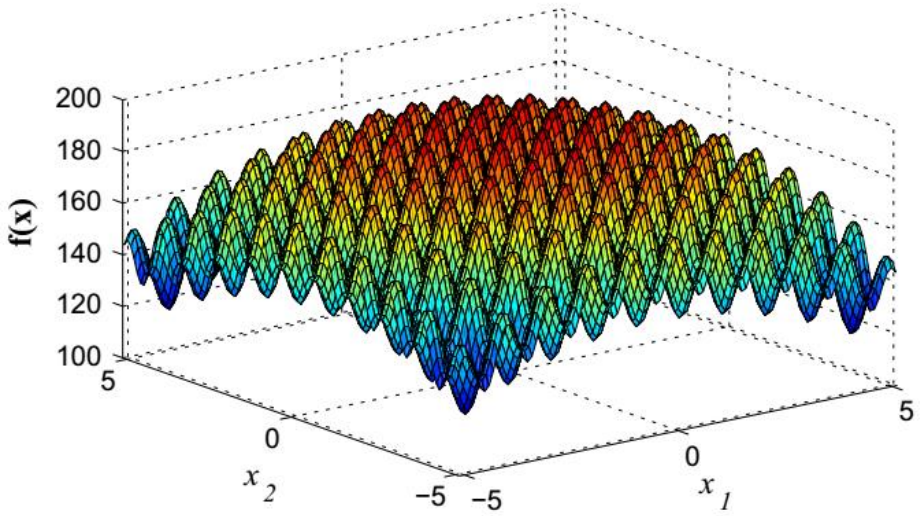
See details in:

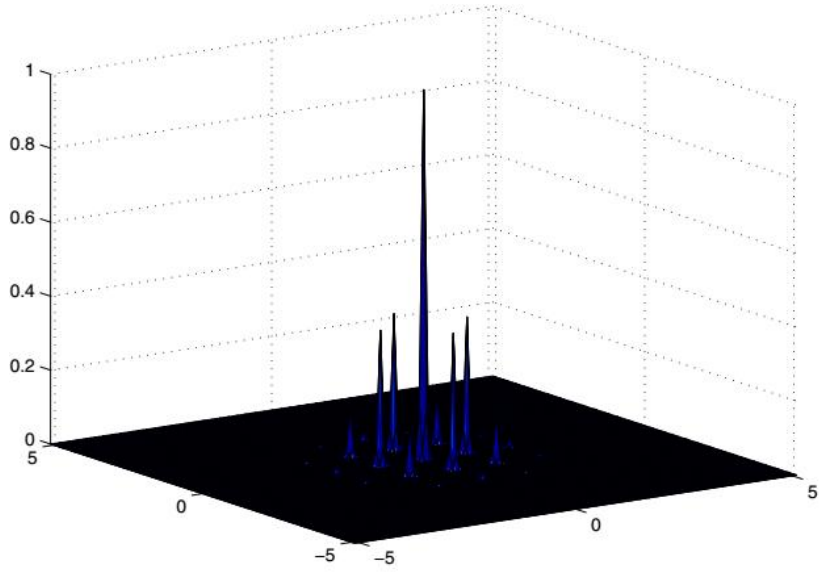
Liu, B., Cheng, S. and Shi, Y., Particle Filter Optimization: A Brief Introduction, Advances in Swarm Intelligence, pp.95-104, 2016.

构建序列目标分布

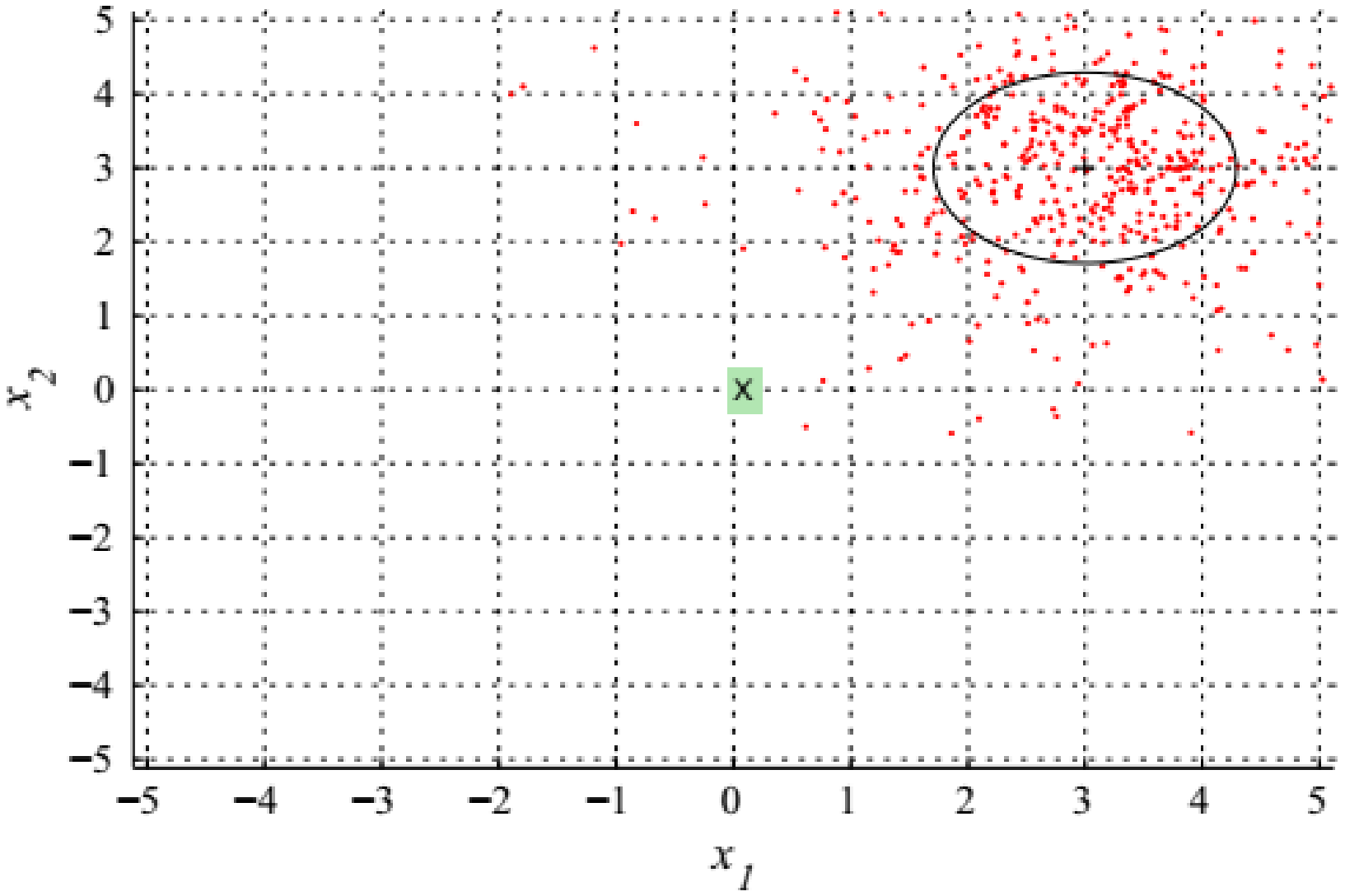
Target pdfs

$$\pi_k(x) \propto f(x)^{\lambda_k}, k = 1, 2, \ldots,$$

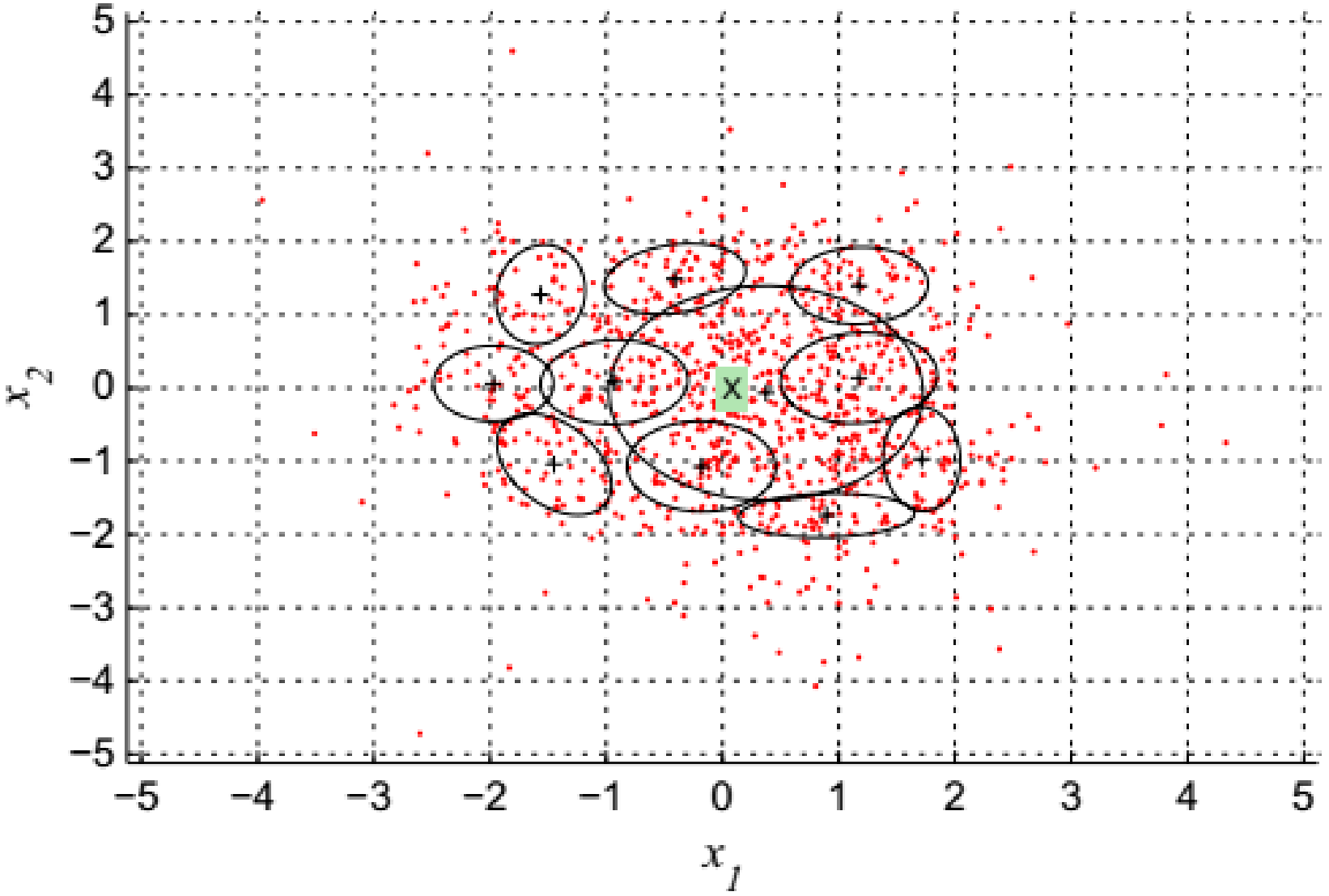




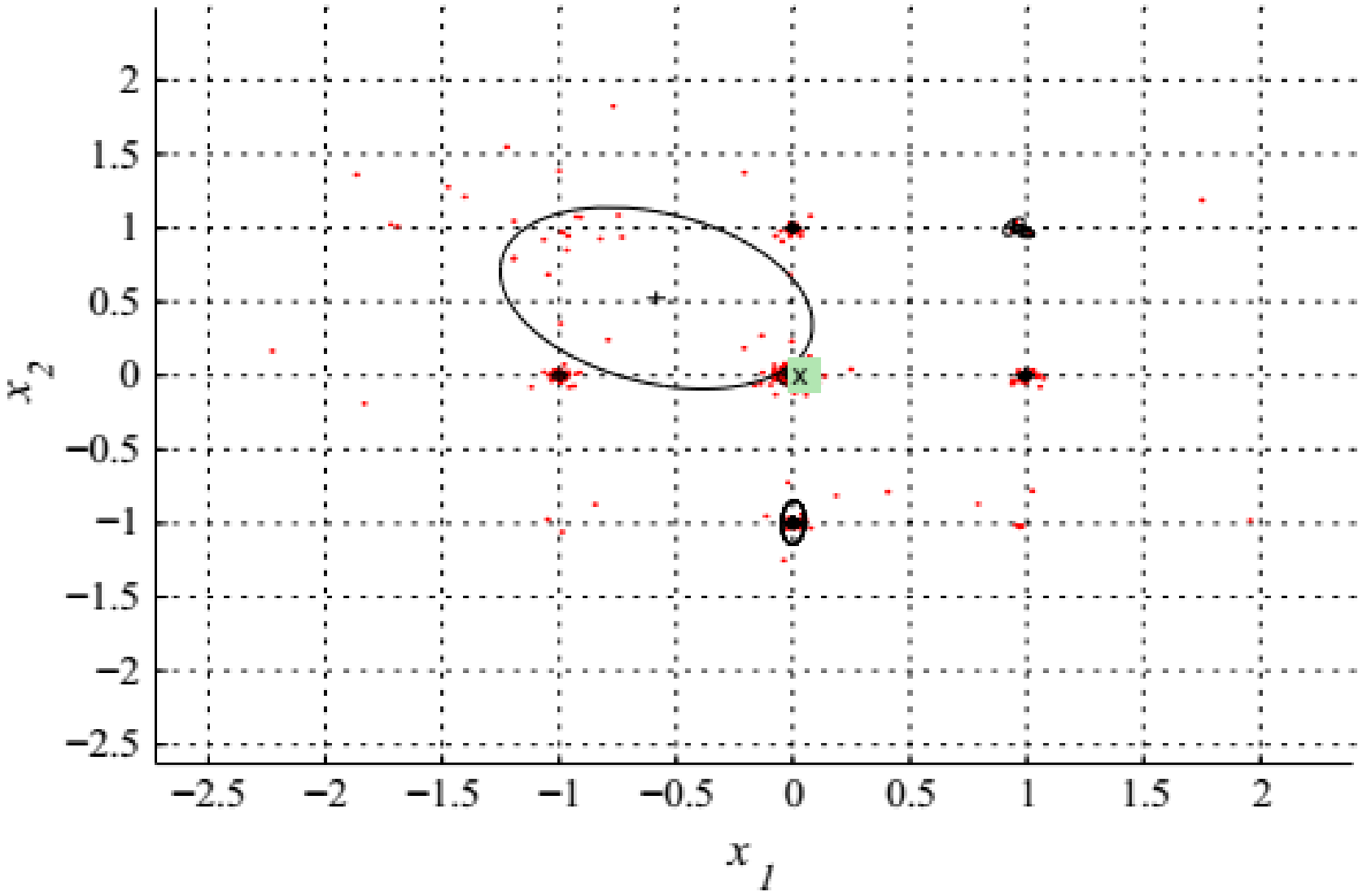
搜索全局最优点：初始化



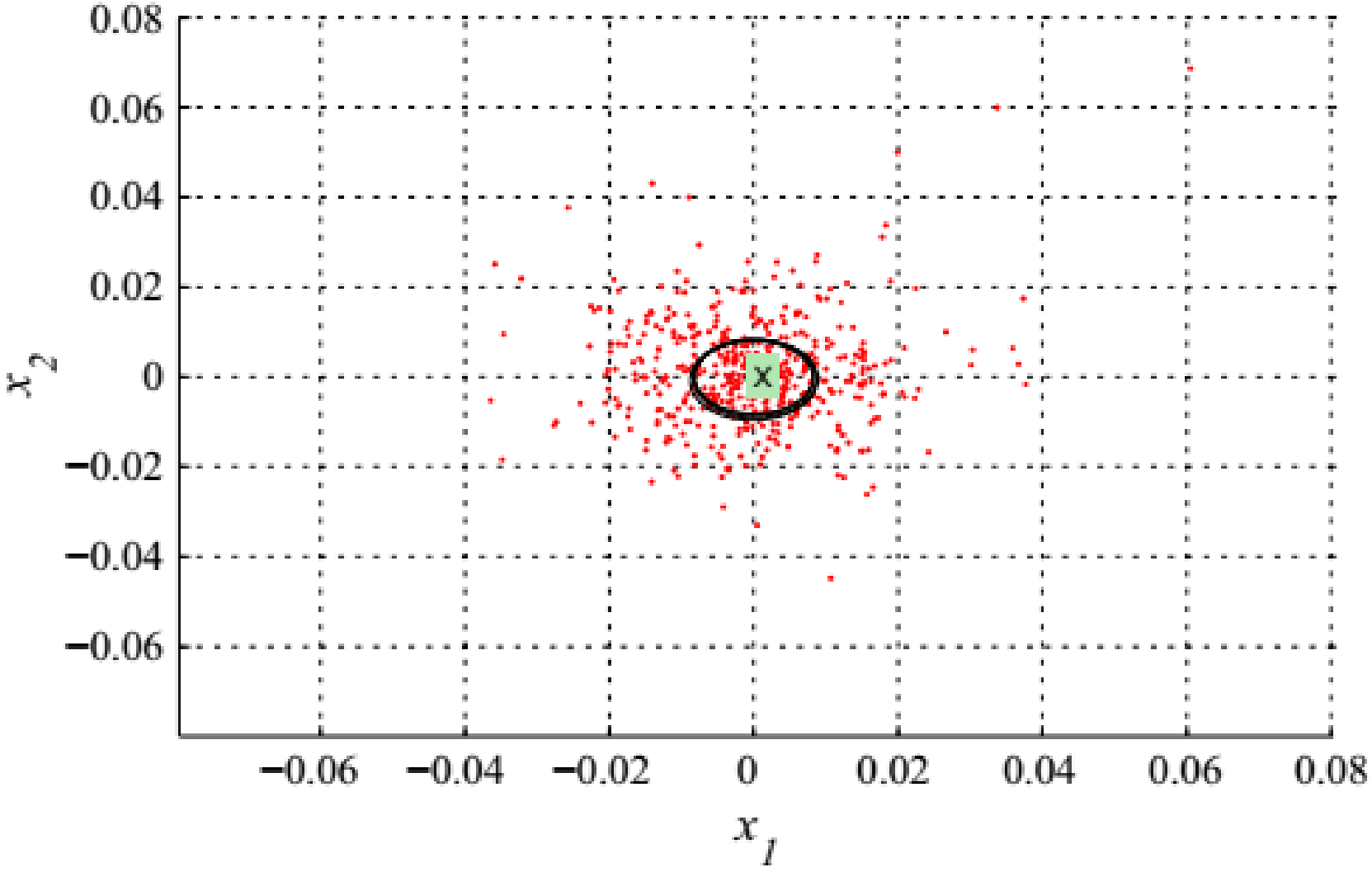
搜索全局最优点：第1次算法迭代结束



搜索全局最优点：第4次算法迭代



搜索全局最优点：最后一次迭代



多基准测试函数、多算法性能比对

Table 1 Convergence results yielded from 100 independent runs of each algorithm on each test problem

Test Problems		Goal: $f(x^*)$	PSO	PFO	SMC-SA	PE-SMC
TF1	2D	30	**30** ± 0	29.4447 ± 0.53	$29.9943 \pm 6 \times 10^{-3}$	*29.9989* $\pm 6.14 \times 10^{-4}$
TF2	2D	1.56261	**1.5626** $\pm 2.23 \times 10^{-15}$	$1.5625 \pm 6.14 \times 10^{-5}$	$1.5626 \pm 7.80 \times 10^{-6}$	*1.5626* $\pm 5.01 \times 10^{-7}$
TF3	2D	1	*0.9974* $\pm 1.26 \times 10^{-2}$	$0.9836 \pm 1.60 \times 10^{-2}$	**0.9978** $\pm 8.60 \times 10^{-3}$	$0.9960 \pm 5.50 \times 10^{-3}$
TF4	2D	2459.6407	2410.6 ± 75.08	2429.9 ± 26.6326	**2436.9** ± 24.1397	*2431.4* ± 30.2447
TF5	2D	1000	**999.9979** $\pm 3.30 \times 10^{-3}$	999.8415 ± 0.16	$999.9058 \pm 8.16 \times 10^{-2}$	*999.9895* $\pm 8.3 \times 10^{-3}$
TF6	2D	19.2085	**19.2085** $\pm 4.04 \times 10^{-14}$	$19.1981 \pm 1.33 \times 10^{-2}$	$19.2084 \pm 1.42 \times 10^{-4}$	*19.2085* $\pm 6.11 \times 10^{-6}$
TF7	2D	100	**100** ± 0	$99.9996 \pm 5.64 \times 10^{-4}$	*100* $\pm 4.95 \times 10^{-5}$	**100** ± 0
TF8	2D	450	**450** ± 0	$449.9759 \pm 2.92 \times 10^{-2}$	*449.9990* $\pm 1.20 \times 10^{-3}$	**450** ± 0
TF9	2D	200	**200** ± 0	$199.9651 \pm 3.60 \times 10^{-2}$	$199.9977 \pm 3.00 \times 10^{-3}$	*199.9999* $\pm 1.12 \times 10^{-6}$
	5D	200	**200** ± 0	197.0448 ± 0.62	$199.9854 \pm 3.67 \times 10^{-2}$	*199.9997* $\pm 4.38 \times 10^{-4}$
	10D	200	*199.8007* ± 0.4690	193.9105 ± 34.27	$199.6702 \pm 4.10 \times 10^{-2}$	**199.9487** $\pm 6.48 \times 10^{-3}$
	20D	200	*199.7810* ± 0.5015	188.6582 ± 44.83	$196.3677 \pm 5.33 \times 10^{-1}$	**199.9228** $\pm 1.13 \times 10^{-2}$
TF10	2D	1	**1** ± 0	*1* $\pm 3.25 \times 10^{-5}$	$1 \pm 6.55 \times 10^{-5}$	$1 \pm 3.77 \times 10^{-5}$
TF11	2D	1800	1758.5 ± 59.22	1777.6 ± 32.09	*1800* $\pm 3.19 \times 10^{-4}$	**1800** $\pm 1.35 \times 10^{-4}$
TF12	2D	486.7309	**486.7309** $\pm 5.98 \times 10^{-13}$	485.599 ± 1.33	*486.7309* $\pm 9.91 \times 10^{-6}$	$486.7298 \pm 9.50 \times 10^{-3}$
TF13	2D	120	**120** ± 0	$119.9994 \pm 5.16 \times 10^{-4}$	$119.9999 \pm 1.63 \times 10^{-4}$	*120* $\pm 9.66 \times 10^{-6}$
TF14	2D	1.8×10^5	*1.8* $\times 10^5 \pm 2.32 \times 10^{-11}$	$1.8 \times 10^5 \pm 1.37 \times 10^{-2}$	$1.8 \times 10^5 \pm 8.4 \times 10^{-3}$	**1.8** $\times 10^5 \pm 0$
TF15	2D	≈509[31]	508.9427 ± 0.50	508.9830 ± 0.11	*509.0020* $\pm 1.93 \times 10^{-10}$	**509.0020** $\pm 1.50 \times 10^{-11}$
TF16	2D	1	0.7900 ± 0.4094	*0.9974* $\pm 5.30 \times 10^{-3}$	0.5000 ± 0.50	**1** $\pm 1.21 \times 10^{-7}$
TF17	2D	1.8013	**1.8013** $\pm 2.89 \times 10^{-15}$	$1.7982 \pm 3.10 \times 10^{-3}$	$1.8008 \pm 7.04 \times 10^{-4}$	*1.8013* $\pm 1.67 \times 10^{-8}$
	5D	4.687658	*4.6624* $\pm 4.01 \times 10^{-2}$	4.2103 ± 0.1648	4.6267 ± 0.1616	**4.6875** $\pm 2.81 \times 10^{-4}$
	10D	9.66015	*9.4562* ± 0.16	5.9357 ± 0.2852	9.3155 ± 0.8074	**9.6596** *pm* 1.77×10^{-2}

This table shows the averages and the standard deviations corresponding with the found optimum over these runs. The best and the 2nd best results found by the different algorithms are indicated with boldface and italics font, respectively

详细内容请参见:

- *Liu, B.*, Posterior Exploration based Sequential Monte Carlo for Global Optimization, Journal of Global Optimization, vol.69, no.4, pp. 847-868, 2017.
.
- *Liu, B.*, Cheng, S. and Shi, Y., Particle Filter Optimization: A Brief Introduction, Advances in Swarm Intelligence, pp.95-104, 2016.
.

小结

- **不确定性(Uncertainty) & 概率(Probability)**
  - 物理概率 & 主观概率
  - Principal Principle
- **贝叶斯推理**
  - 一个小例子
  - 模型参数不确定性&模型不确定性
- **贝叶斯采样之序列蒙特卡洛**
- **“拥抱”不确定性**
  - 利用贝叶斯采样搜索全局最优点

很重要但没有讲到的

- 先验怎么选？
  - 主观 or 客观？
  - 经验 or 数据？
  - 共轭 or 非共轭？
  - 。。。
- 模型怎么建？
  - 参数 or 非参数？
  - 线性 or 非线性？
  - 高斯 or 非高斯？
  - 。。。
- 样本怎么采？
  - MCMC: Gibbs、Slice sampling
  - Rejection sampling
  - Hamiltonian Monte Carlo
  - Quasi Monte Carlo
  - 。。。

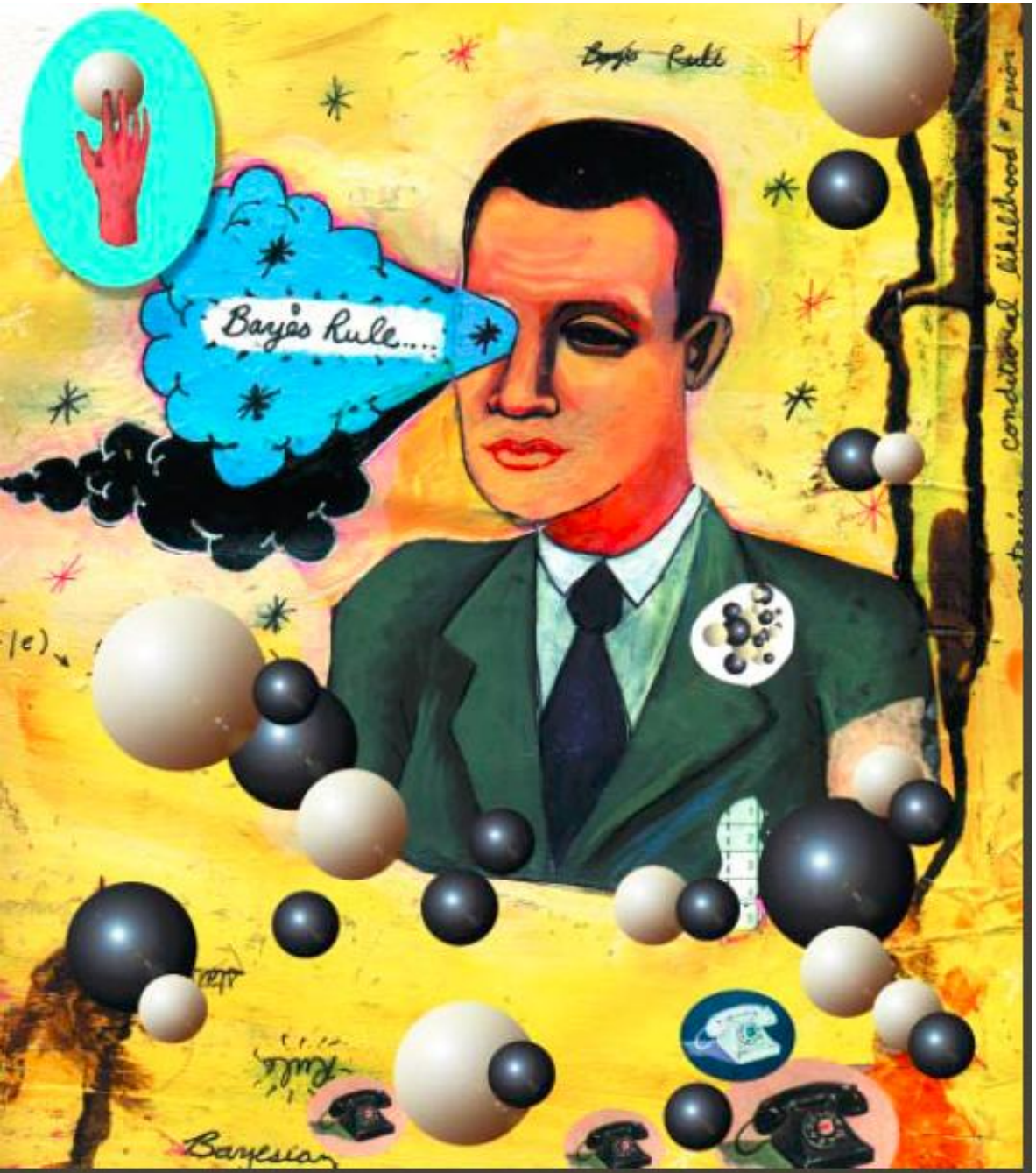


Illustration by P. J. Fidler

贝叶斯 & 人类认知

Special Issue: Probabilistic models of cognition

Bayesian decision theory in sensorimotor control

Konrad P. Körding[1] and Daniel M. Wolpert[2]

[1]Brain and Cognitive Sciences, Massachusetts Institute of Technology, Building NE46-4053, Cambridge, Massachusetts, 02139, USA
[2]Department of Engineering, University of Cambridge, Trumpington Street, Cambridge, CB2 1PZ, UK

rewards of different action outcomes. We review recent studies that have investigated the mechanisms used by the nervous system to solve such estimation and decision problems, which show that human behaviour is close to that predicted by Bayesian Decision Theory. This theory defines optimal behaviour in a world characterized by uncertainty, and provides a coherent way of describing sensorimotor processes.

Bayes in the Brain—On Bayesian Modelling in Neuroscience

Matteo Colombo and Peggy Seriès

Matteo Colombo
Department of Philosophy
University of Edinburgh
Dugald Stewart Building, 3 Charles Street
George Square, EH8 9AD, Edinburgh
UK

Peggy Seriès
Institute for Adaptive and Neural Computation
School of Informatics, University of Edinburgh
10 Crichton Street, EH8 9AB, Edinburgh
UK

Neurons as Monte Carlo Samplers: Bayesian Inference and Learning in Spiking Networks

Yanping Huang
University of Washington

Rajesh P.N. Rao
University of Washington

from spike trains generated by sensory neurons. We show how such a neuronal network with synaptic plasticity can implement a form of Bayesian inference similar to Monte Carlo methods such as particle filtering. Each spike in the population of inference neurons represents a sample of a particular hidden world state. The spiking activity across the neural population approximates the posterior distribution of hidden state. The model provides a functional explanation for the Poisson-

“According to a growing trend in theoretical neuroscience, the human perceptual system is akin to a Bayesian machine.”

Is Perception Bayesian Inference?

tions of the mechanisms of sensory systems. The most we can acknowledge from existing evidence is that viewing ‘perception as Bayesian inference’ is useful for generating *predictions* about people’s *performance* in perceptual tasks. We explain that the goal of Bayesian models in psychophysics experiments is *not* to describe sensory mechanisms. Bayesian models are used as tools for predicting, systematizing and classifying statements about people’s observable performance. Hence, claims about perception as Bayesian infer-

J Neurophysiol 102: 1–6, 2009.
First published April 29, 2009; doi:10.1152/jn.00239.2009.

Neuro Forum

Dynamical Foundations of the Neural Circuit for Bayesian Decision Making

Kenji Morita
RIKEN Brain Science Institute, Wako, Japan

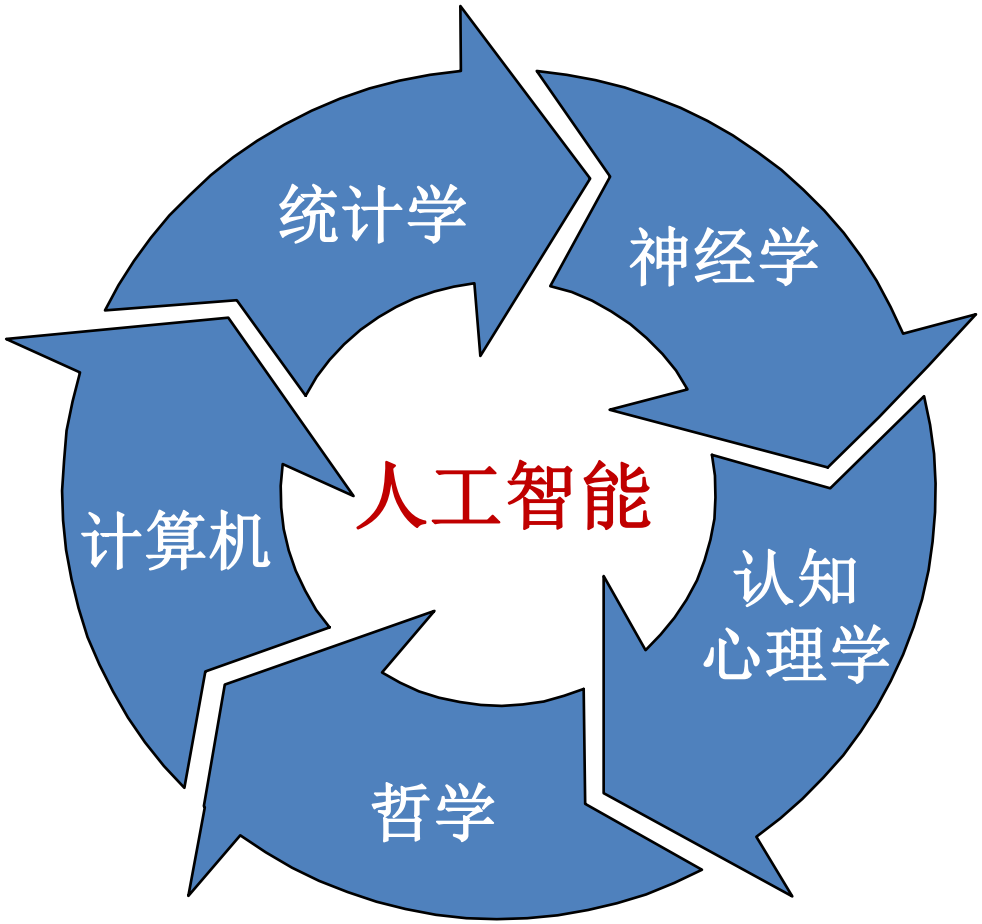
There has been an increasing number of psychological evidences indicating that humans and animals can often make decision behaviors that are nearly optimal in terms of Bayesian inference. Given that animals must have been somehow opti-

sions. The emerging question is how such computations as Bayesian inference can be implemented by biological substrates in the brain neural circuits. In fact, this is one of the

总结

- 贝叶斯是一种世界观、方法论（甚至哲学）
  - 以Degree of belief 描述/定义“概率”、量测“不确定性”
  - 刻划了Belief Propagation的过程
  - 认为研究者（人）与被研究对象（物）是一个有机整体，而非互相独立
- 贝叶斯是一种信息融合/处理机制
- 采样算法令贝叶斯方法可算、可用
- 贝叶斯是研究人脑认知的重要工具

通往真正AI之路?



-

个人观点，欢迎批评指正，谢谢！