非线性贝叶斯反问题

黄政宇
北京大学北京国际数学研究中心
北京大学国际机器学习研究中心

本堂课大纲

- 课程内容简介
  - 贝叶斯反问题
  - 贝叶斯采样、推理方法简介
  - 贝叶斯后验分布的性质

贝叶斯反问题

- 贝叶斯反问题

$$y = \mathcal{G}(\theta) + \eta \qquad \eta \sim \rho_\eta \qquad \theta \sim \rho_{\text{prior}}$$

- 假设

高斯先验分布: $\rho_{\text{prior}}(\theta) = \mathcal{N}(\theta; r_0, \Sigma_0)$

高斯噪声: $\rho_\eta = \mathcal{N}(x; 0, \Sigma_\eta)$

- 后验分布

$$\rho_{\text{post}}(\theta; y) \propto \rho(y|\theta)\rho_{\text{prior}}(\theta) \propto e^{-\Phi_R(\theta, y)}$$

$$\rho_{\text{post}}(\theta; y) = \frac{1}{Z} e^{-\Phi_R(\theta, y)}$$

$$\Phi_R(\theta, y) = \frac{1}{2} \| \Sigma_\eta^{-\frac{1}{2}} (y - \mathcal{G}(\theta)) \|^2 + \frac{1}{2} \| \Sigma_0^{-\frac{1}{2}} (\theta - r_0) \|^2$$

贝叶斯反问题

- 优化方法

最大似然估计：

$$\operatorname{argmin}\frac{1}{2}\| \Sigma_{\eta}^{-\frac{1}{2}}(y - \mathcal{G}(\theta))\|^2$$

最大后验估计：

$$\operatorname{argmin}\frac{1}{2}\| \Sigma_{\eta}^{-\frac{1}{2}}(y - \mathcal{G}(\theta))\|^2 + \frac{1}{2}\| \Sigma_{0}^{-\frac{1}{2}}(\theta - r_0)\|^2$$

非线性最小二乘问题：高斯-牛顿方法，Levenberg Marquardt 方法等

- 贝叶斯方法

采样：$\rho_{\text{post}}(\theta; y) = \frac{1}{Z} e^{-\Phi_R(\theta, y)}$

贝叶斯采样、推理

- 有未知归一化常数的目标分布

$$\rho^*(\theta) = \frac{1}{Z} e^{-\Phi_R(\theta)}$$

未知　　已知

- 计算目标分布的期望、协方差等
- 计算目标函数的期望 $\mathbb{E}[f] = \int f(\theta)\rho^*(\theta)d\theta$
- 生成服从目标分布的样本 $\{\theta_j\} \sim \rho^*(\theta)$

贝叶斯采样、推理

- 参数化 (parametric) 方法

  高斯近似：

  $$\rho^*(\theta) \approx \mathcal{N}(\theta; m, C)$$

  混合高斯近似：

  $$\rho^*(\theta) \approx \sum_{j=1}^{J} w_j \mathcal{N}(\theta; m_j, C_j) \qquad \sum_{j=1}^{J} w_j = 1$$

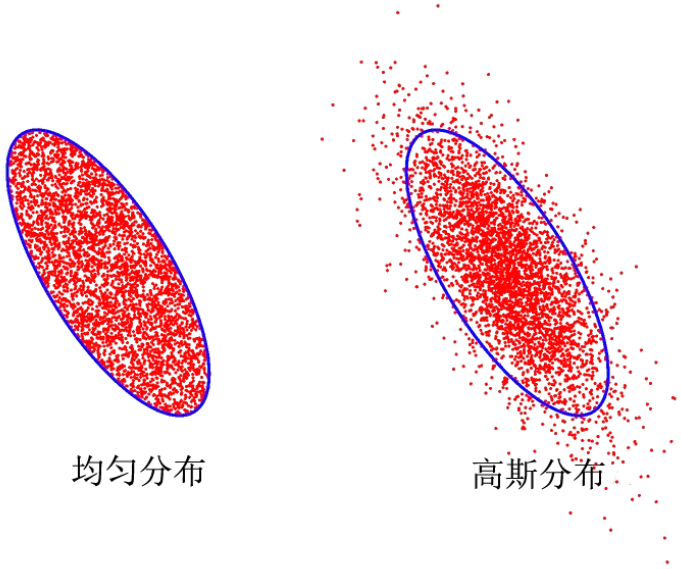
  ……

- 非参数化(nonparametric)方法

  粒子近似 $J \gg 1$：

  $$\rho^*(\theta) \approx \{\theta_j\}_{j=1}^{J}$$

  $$\rho^*(\theta) \approx \frac{1}{J}\sum_{j=1}^{J} \delta(\theta - \theta_j)$$



贝叶斯采样、推理

- 输运的方法（直接近似的方法）

暴力网格搜索：假设$N_\theta = 2$,

$$\theta_{i,j} = [-L + \frac{2(i-1)}{N-1}L, -L + \frac{2(j-1)}{N-1}L] \qquad Z = \sum \rho^*(\theta_{i,j})$$

输运：$\{\theta_i\} \sim \rho_{\text{prior}}$ $\qquad \{T\theta_i\} \sim \rho^*$

- 重要性采样
- 卡尔曼方法
- 标准化流方法

……

贝叶斯采样、推理

- 马氏链蒙特卡洛 (Markov chain Monte Carlo) 方法

暴力网格搜索：假设$N_\theta = 2$,

$$\theta_{i,j} = [-L + \frac{2(i-1)}{N-1}L, -L + \frac{2(j-1)}{N-1}L] \qquad Z = \sum \rho(\theta_{i,j})$$

随机游走：$\theta_n \to \theta_{n+1}$ 概率分布 $\psi_I(\theta_n, \theta_{n+1})$

不变分布是目标分布

$$\rho_{\text{post}}(\theta) = \int \psi_I(\theta', \theta)\rho_{\text{post}}(\theta')d\theta'$$

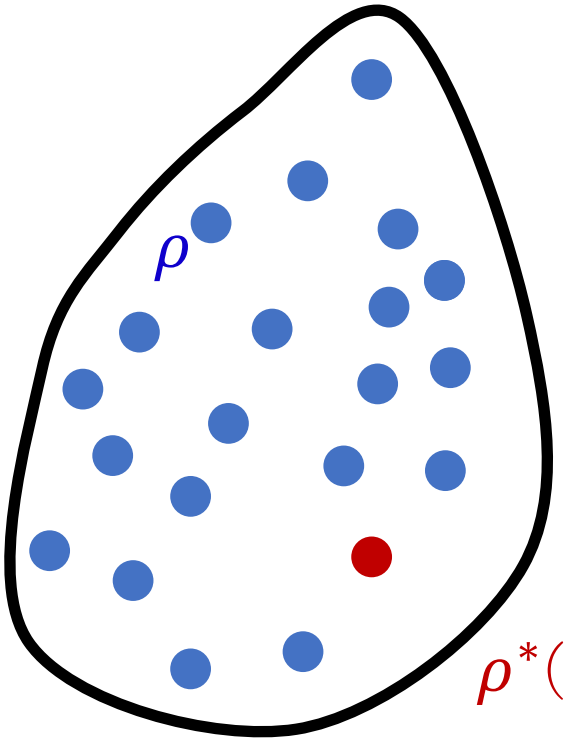
遍历性质

$$\{\theta_n\}_{n \geq N} \sim \rho_{\text{post}}(\theta)$$

贝叶斯采样、推理

- 变分推理 (Variational Inference)

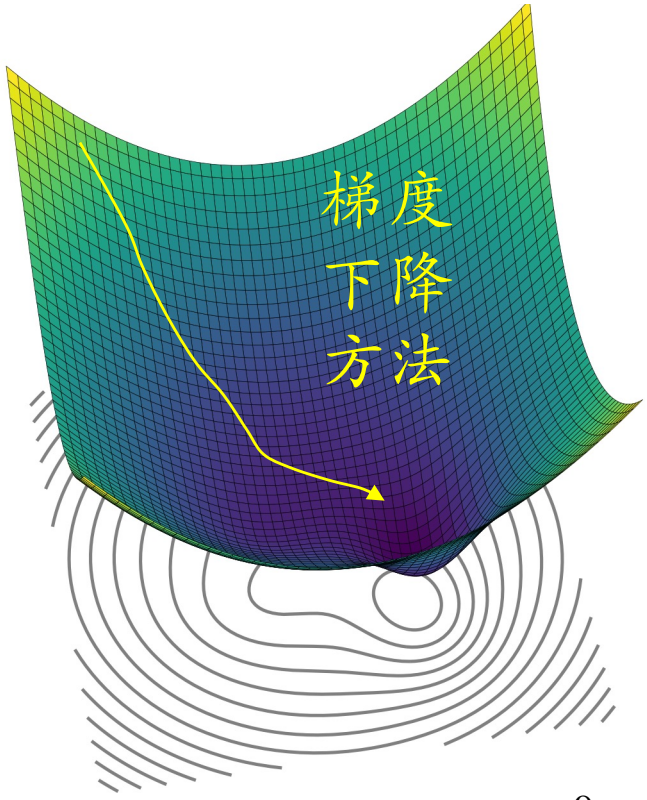
$$\text{minimize}_{\rho}\ \mathcal{E}(\rho;\rho^*)$$



ρ^*(极小值)

概率密度空间$\mathcal{P}$

度量 $M(\rho)$
距离 $\mathcal{D}(\rho_A,\rho_B)$



9

贝叶斯采样、推理

- 方法总结

	效率	精度	概率密度函数空间
基于输运的方法	高	低	参数化空间（比如高斯族、流模型）
马氏链蒙特卡洛方法	低	高	全空间
变分推理方法	中	中	参数化空间（比如高斯族）

备注：这只是通常情况，基于输运的方法或变分推理方法也可以通过更强的参数化来提高近似精度，但往往会牺牲计算效率。

贝叶斯后验分布

- Rosenbrock 函数 $(N_\theta = 2)$

$$y = \mathcal{G}(\theta) + \eta$$

$$\mathcal{G}(\theta) = \begin{bmatrix} \theta_2 - c_1\theta_1^2 \\ \theta_1 \end{bmatrix} \qquad y = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

高斯先验分布: $\rho_{\text{prior}}(\theta) = \mathcal{N}(\theta; 0, \begin{bmatrix} 10^2 & \\ & 10^2 \end{bmatrix})$

高斯噪声: $\rho_\eta = \mathcal{N}(x; 0, c_2 \begin{bmatrix} \frac{1}{10^2} & \\ & 1 \end{bmatrix})$

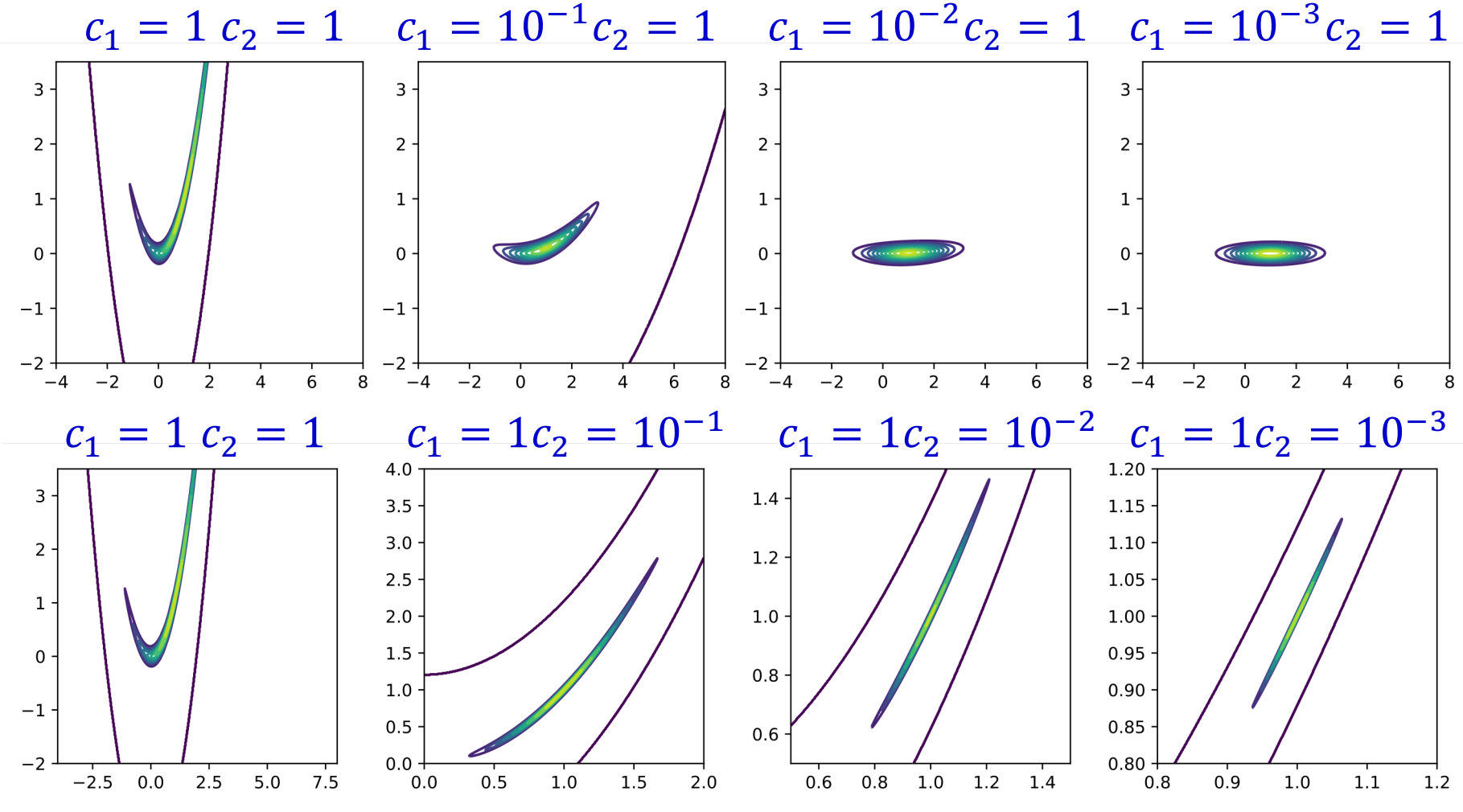
考虑:

$c_1 = 1, 10^{-1}, 10^{-2}, 10^{-3}$ $\qquad c_2 = 1$

$c_2 = 1, 10^{-1}, 10^{-2}, 10^{-3}$ $\qquad c_1 = 1$

贝叶斯后验分布

- Rosenbrock 函数 $(N_\theta = 2)$



- 在一定条件下，后验分布会趋于高斯
- 强非线性、多尺度问题中，预条件至关重要

贝叶斯后验分布

- Bernstein Von Mises 理论

对于反问题

$$y = \mathcal{G}(\theta) + \eta$$

我们假设噪声满足$\eta \sim \mathcal{N}\left(x; 0, \gamma^2 \Sigma_\eta\right)$，先验分布满足$\theta \sim \rho_{\text{prior}}$。

在一定条件下，当$\gamma \to 0$

$$\rho_{\text{post}}(\theta) \to \mathcal{N}\left(\theta; \theta^\dagger, \gamma^2 \left(\nabla_\theta \mathcal{G}\left(\theta^\dagger\right)^T \Sigma_\eta^{-1} \nabla_\theta \mathcal{G}\left(\theta^\dagger\right)\right)^{-1}\right)$$

其中$\theta^\dagger$是唯一最大似然估计，即$(y - \mathcal{G}(\theta))\Sigma_\eta^{-1}(y - \mathcal{G}(\theta))$的唯一极小值。

贝叶斯后验分布

- Bernstein Von Mises 理论

对于数据

$$y_i = \mathcal{G}(\theta) + \eta, \qquad i = 1,2\cdots,N$$

我们假设噪声满足$\eta \sim \mathcal{N}(x; 0, \Sigma_\eta)$，先验分布满足$\theta \sim \rho_{\text{prior}}$。

在一定条件下，当$N \to \infty$

$$\rho_{\text{post}}(\theta) \to \mathcal{N}\left(\theta; \theta^\dagger, N^{-1}\left(\nabla_\theta \mathcal{G}(\theta^*)^T \Sigma_\eta^{-1} \nabla_\theta \mathcal{G}(\theta^*)\right)^{-1}\right)$$

其中$\theta^\dagger$是唯一最大似然估计，θ^*是真实值。

贝叶斯后验分布

- Bernstein Von Mises 理论

对于数据

$$y_i \sim P_\theta(\cdot), \qquad i = 1,2\cdots,N$$

我们假设先验分布满足$\theta \sim \rho_{\text{prior}}$。

在一定条件下，当$N \to \infty$

$$\rho_{\text{post}}(\theta) \to \mathcal{N}(\theta; \theta^\dagger, N^{-1} I(\theta^*)^{-1})$$

其中 $\theta^\dagger$是唯一最大似然估计，θ^*是真实值，$I(\theta^*)$是y分布的Fisher信息矩阵（可逆）

$$I(\theta^*) = -\mathbb{E}_{P_{\theta^*}}[\nabla_\theta^2 \log P_{\theta^*}]$$

贝叶斯后验分布

- 多峰函数 $(N_\theta = 2)$

$$y = \mathcal{G}(\theta) + \eta$$

$$\mathcal{G}(\theta) = (\theta_1 - \theta_2)^2 \qquad y = 4$$

高斯先验分布: $\rho_{\text{prior}}(\theta) = \mathcal{N}(\theta; \begin{bmatrix} c \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \\ & 1 \end{bmatrix})$

高斯噪声: $\rho_\eta = \mathcal{N}(x; 0, 1)$

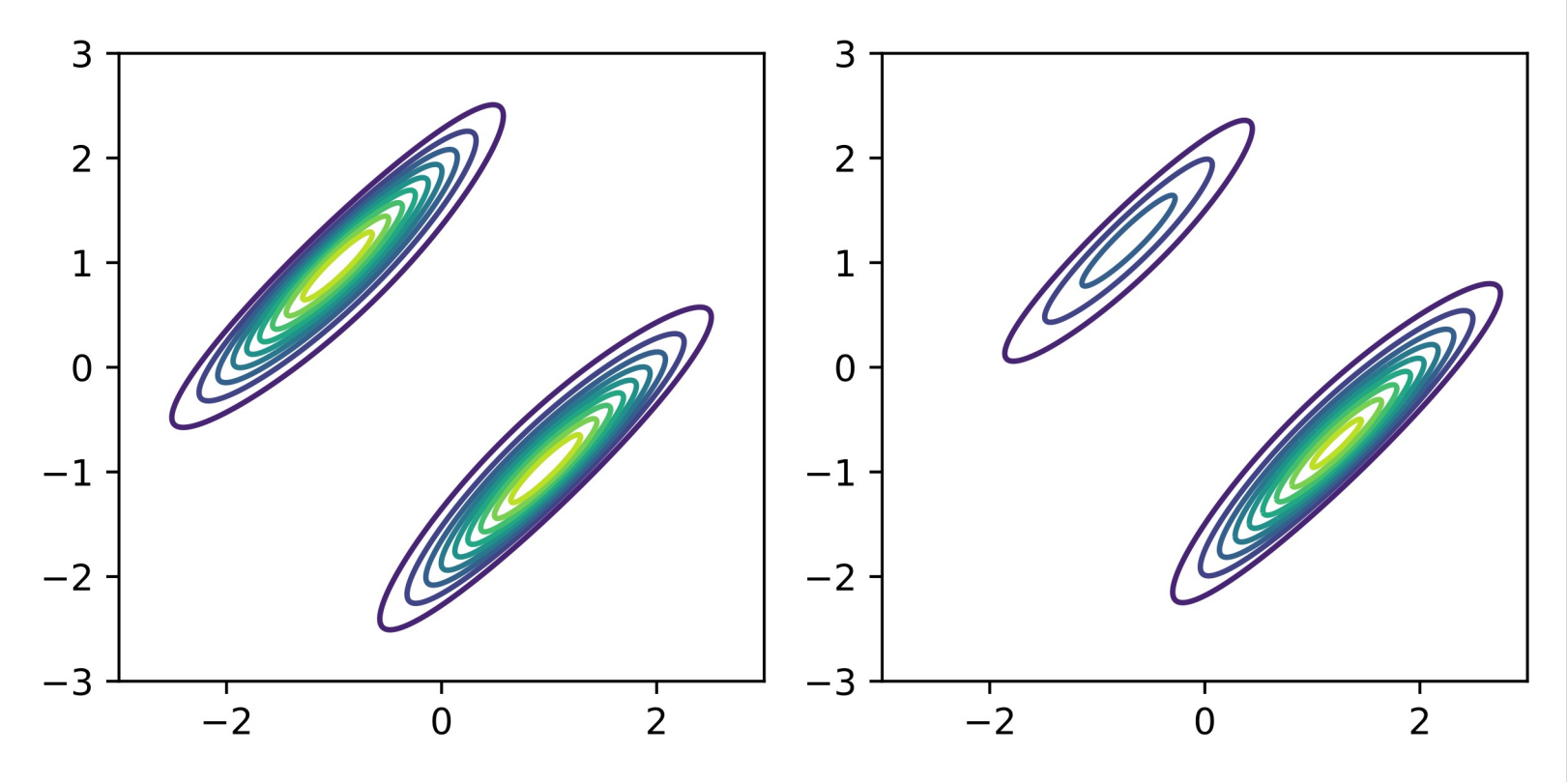
考虑:

$c = 0,\ 0.5$

贝叶斯后验分布

- 多峰函数 $(N_\theta = 2)$

$c = 0$ $\quad$ $c = 0.5$



- 马氏链蒙特卡洛容易陷入单峰，跨峰慢
- 基于高斯的变分推理或其它方法容易把两峰“抹平”或“漏峰”

扩展阅读

- Bernstein Von Mises 理论

  Wikipedia:
  https://en.wikipedia.org/wiki/Bernstein%E2%80%93von_Mises_theorem.

  Bickel, Peter J., and Joseph A. Yahav. "Some contributions to the asymptotic theory of Bayes solutions." Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 11.4 (1969): 257-276.