变分推理方法

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
  - 变分推理的要素(variational inference)
    - 概率密度空间
    - 能量泛函
    - 度量、距离
  - 参数化(parametric)变分推理
  - 非参数化(nonparametric)变分推论

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

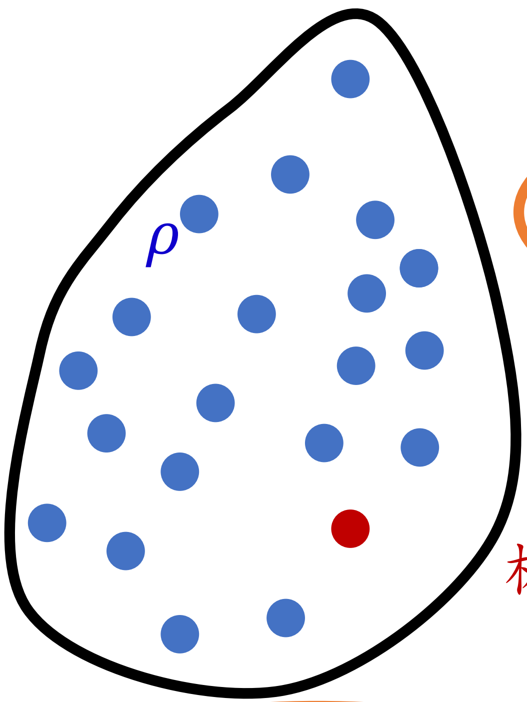
$$\Phi_R(\theta, y) = \frac{1}{2} \| \Sigma_\eta^{-\frac{1}{2}} (y - \mathcal{G}(\theta)) \|^2 + \frac{1}{2} \| \Sigma_0^{-\frac{1}{2}} (\theta - r_0) \|^2$$

- 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 $\{\theta_j\} \sim \rho^*(\theta)$

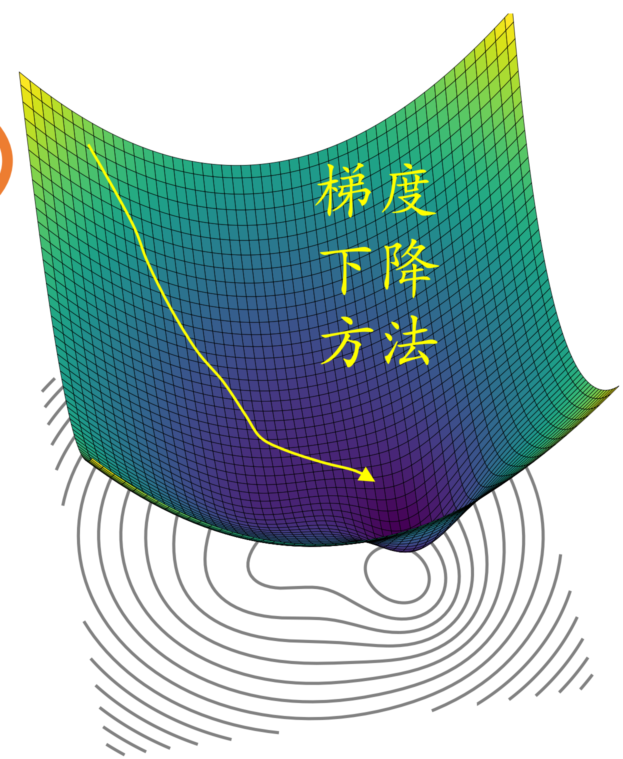
贝叶斯采样、推理

➢ 变分推理

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



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



极小值接近 ρ^*

概率密度空间$\mathcal{P}$

概率密度空间

➢ 参数化方法

高斯密度空间 $\{(m,C),C \succ 0\}$

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

简化的高斯密度空间 $\{(m,\delta),\delta > 0\}$

$$\rho(\theta) \approx \mathcal{N}(\theta;m,\delta^2 I)$$

混合高斯近似 $\{(w_k,\ m_k,\ C_k)_{k=1}^K, C_k \succ 0, w_k \geq 0\}$

$$\rho(\theta) \approx \sum_{k=1}^K w_k \mathcal{N}(\theta;m_k,C_k) \qquad \sum_{k=1}^K w_k = 1$$

……

概率密度空间

- 指数分布族

$$\rho(\theta;a) = h(\theta)e^{T(\theta)\cdot a - A(a)}$$

归一化常数

高斯分布： $\mathcal{N}(\theta;m,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(\theta-m)^2}{2\sigma^2}}$

$$T(\theta) = [\theta;\theta^2] \qquad a = \left[\frac{m}{\sigma^2}; -\frac{1}{\sigma^2}\right]$$

$$h(\theta) = \frac{1}{\sqrt{2\pi}} \qquad A(a) = \frac{m^2}{2\sigma^2} + \log\sigma$$

泊松分布： $\frac{\lambda^\theta e^{-\lambda}}{\theta!}$ $(\theta \in Z^{0+})$

$$T(\theta) = \theta \qquad a = \log\lambda$$

$$h(\theta) = \frac{1}{\theta!} \qquad A(a) = \lambda$$

概率密度空间

- 练习

$$\rho(\theta; a) = h(\theta)\exp\{T(\theta)\cdot a \ - A(a)\}$$

期望：

$$\mathbb{E}_\rho[T(\theta)] = \nabla_a A(a)$$

Fisher信息矩阵(Fisher information matrix)：

$$\begin{aligned}\mathrm{FIM}\big(\rho(\theta; a)\big) &= \mathbb{E}_\rho[\nabla_a \log \rho(\theta; a)^T \nabla_a \log \rho(\theta; a)] \\ &= -\mathbb{E}_\rho[\nabla_a \nabla_a \log \rho(\theta; a)\,] \\ &= \nabla_a \nabla_a A(a)\end{aligned}$$

概率密度空间

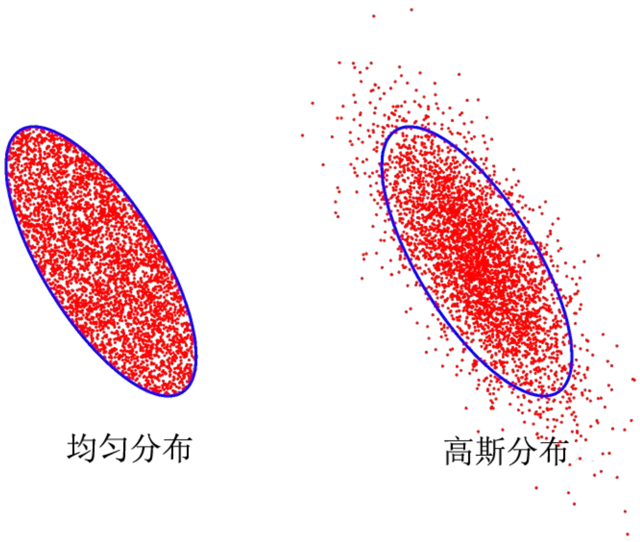
- 非参数化方法

$$\mathcal{P} = \{\rho:\ \rho \in C^{\infty},\ \rho(\theta) > 0\,,\ \int \rho(\theta)d\theta = 1\}$$

粒子近似 $J \gg 1$ ：

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

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



能量泛函(energy functional)

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

- $\mathcal{E}(\rho;\rho^*) \geq 0$
- $\mathcal{E}(\rho;\rho) = 0$

- 理论证明：量化收敛性
- 变分贝叶斯、机器学习：目标函数
- 统计测试：区分两个分布、量化两个分布的差异

- 不一定是距离

距离函数

- 全变差距离(total variation)

$$\mathcal{D}_{\mathrm{TV}}(\rho_A,\rho_B) := \frac{1}{2}\int |\rho_A(\theta)-\rho_B(\theta)|\,d\theta = \frac{1}{2}\|\rho_A(\theta)-\rho_B(\theta)\|_{L_1}$$

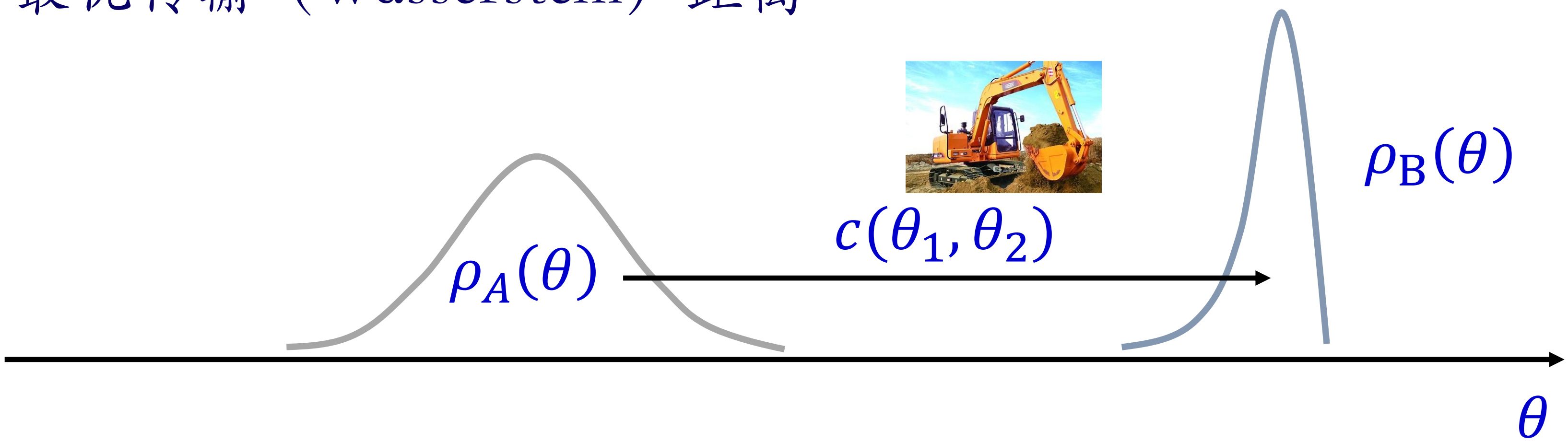
- 海林格(Hillinger)距离

$$\mathcal{D}_{\mathrm{H}}(\rho_A,\rho_B) := \left(\frac{1}{2}\int \left|\sqrt{\rho_A(\theta)}-\sqrt{\rho_B(\theta)}\right|^2 d\theta\right)^{\frac{1}{2}}$$

$$= \frac{1}{\sqrt{2}}\left\|\sqrt{\rho_A(\theta)}-\sqrt{\rho_B(\theta)}\right\|_{L_2}$$

距离函数

- 最优传输（Wasserstein）距离

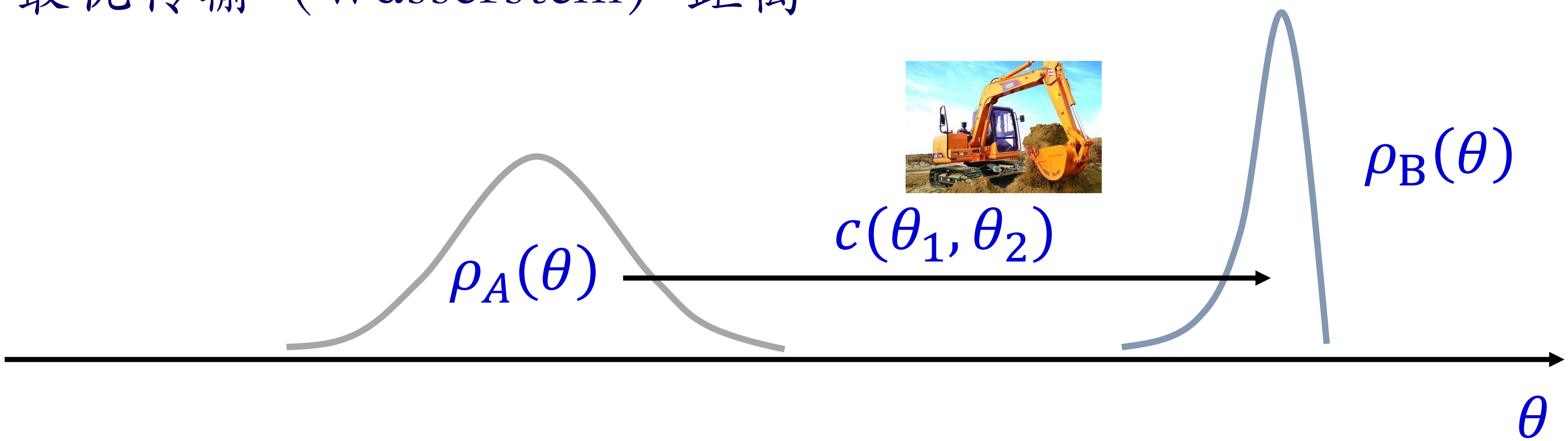


Monge 问题

$$\min_T \iint c(\theta, T(\theta))\rho_A(\theta)d\theta$$

$$T\#\rho_A = \rho_B$$

➢ 最优传输（Wasserstein）距离



$$\min_{\gamma} \iint \gamma(\theta_1,\theta_2)c(\theta_1,\theta_2)d\theta_1 d\theta_2$$

Kantorovich 问题

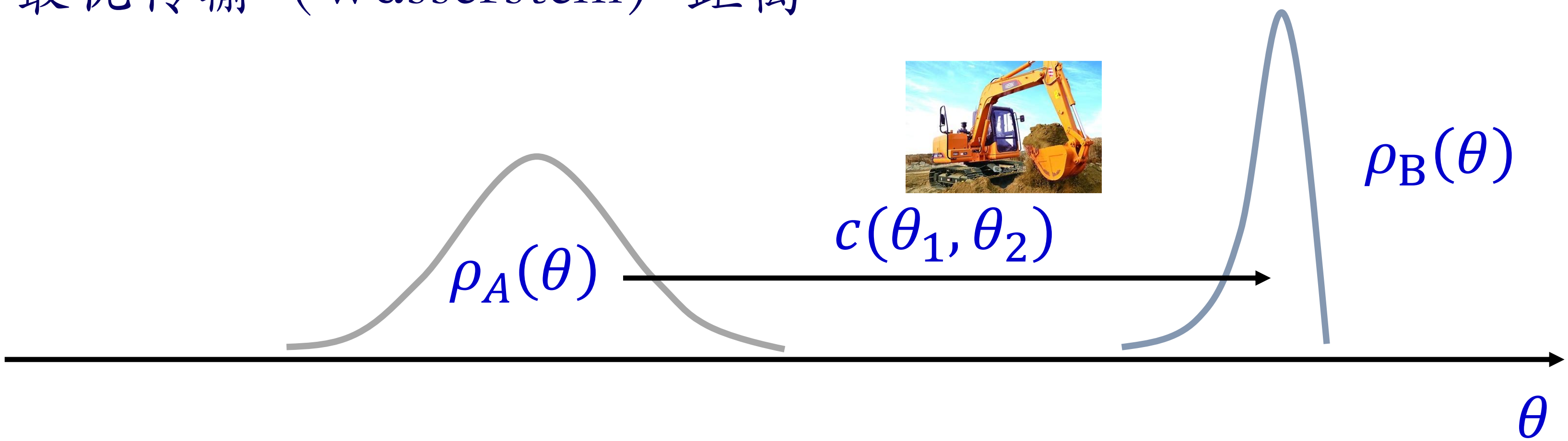
$$\int \gamma(\theta_1,\theta_2)\, d\theta_2 = \rho_A(\theta)$$

$$\int \gamma(\theta_1,\theta_2)\, d\theta_1 = \rho_B(\theta)$$

$$\gamma(\theta_1,\theta_2) \geq 0$$

距离函数

- 最优传输（Wasserstein）距离



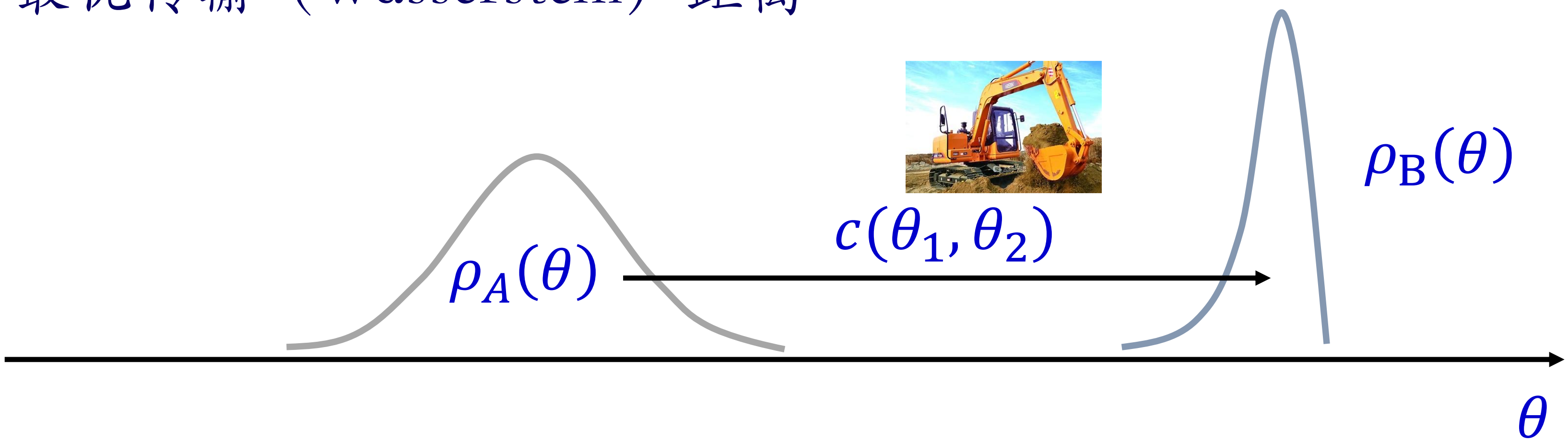
Kantorovich 对偶问题

$$\sup_{f,g} \int \rho_A(\theta) f(\theta) + \rho_B(\theta) g(\theta) d\theta$$

$$f(\theta_1) + g(\theta_2) \leq c(\theta_1, \theta_2)$$

距离函数

- 最优传输（Wasserstein）距离



Brenier 问题 (Wasserstein-2距离 $c(\theta_1, \theta_2) = \| \theta_1 - \theta_2 \|_2^2$)

$$\min_{v} \int_0^1 \int \| v(t,\theta) \|_2^2 \, \rho(t,\theta) d\theta dt$$

$$\frac{\partial \rho(t,\theta)}{\partial t} + \nabla \cdot \big(\rho(t,\theta) v(t,\theta)\big) = 0$$

$$\rho(0,\theta) = \rho_A(\theta) \qquad \rho(1,\theta) = \rho_B(\theta)$$

- f-散度

$$D_f[\rho \parallel \rho^*] = \int \rho^* f\left(\frac{\rho}{\rho^*}\right) d\theta$$

其中f是凸函数，$f(1) = 0$。

琴生不等式：

$$\mathbb{E}_{\rho^*}\left[f\left(\psi(\theta)\right)\right] \geq f(\mathbb{E}_{\rho^*}[\psi(\theta)])$$

$$\sum_j \rho^*(\theta_j) d\theta f\left(\psi(\theta_j)\right) \geq f(\sum_j \rho^*(\theta_j) d\theta \psi(\theta_j))$$

因此

$$D_f[\rho \parallel \rho^*] \geq 0$$

f-散度(f- divergence)

- KL-散度

$$f = x\log x \qquad \mathrm{KL}[\rho \parallel \rho^*] = \int \rho \log\left(\frac{\rho}{\rho^*}\right) d\theta$$

$$\mathrm{KL}(\rho \parallel Z\rho^*) = \mathrm{KL}(\rho \parallel \rho^*) - \log(Z)$$

- 反向 KL-散度

$$f = -\log x \qquad \mathrm{KL}[\rho^* \parallel \rho] = \int \rho^* \log\left(\frac{\rho^*}{\rho}\right) d\theta$$

- χ^2-距离

$$f = (x-1)^2 \qquad \chi^2[\rho \parallel \rho^*] = \int \frac{\rho^2}{\rho^*} d\theta - 1$$

能量泛函(energy functional)

- 最大均值差异(maximum mean discrepancy)

给点任意函数类F

$$\mathrm{MMD}[\rho,\rho^*]=\sup_{f\in F}\left(\mathbb{E}_\rho[f(\theta)]-\mathbb{E}_{\rho^*}[f(\theta)]\right)$$

离散情况：

$$\mathrm{MMD}[\mathrm{X},X^*]=\sup_{f\in F}\left(\frac{1}{m}\sum_i f(x_i)-\frac{1}{n}\sum_i f(x_i^*)\right)$$

$F=\{f:|f|_\infty\leq 1\}$，$F=\mathrm{span}\{x,x^2\}\cdots\cdots$

度量(metric)

- 欧式空间的度量

欧式空间：R^N

线性切空间：R^N

度量： $g_x : T_xR^N \times T_xR^N \to R$

$$g_x(\sigma_1, \sigma_2):= \sqrt{\langle M(x)\sigma_1 \cdot \sigma_2 \rangle}$$

度量张量： $M(x): T_xR^N \to T_x^*R^N$

曲线距离: $\dot{x}_t = \sigma_t$

$$D(\rho_0, \rho_1) = \int_0^1 g_\rho(\sigma_t, \sigma_t) dt$$

度量(metric)

- 概率空间的度量

概率密度空间：$\mathcal{P}=\left\{\rho\in C^{\infty},\int\rho d\theta=1\right\}$

线性切空间：$T_\rho\mathcal{P}\subseteq\left\{\sigma\in C^{\infty},\int\sigma d\theta=0\right\}$

$$\sigma=\rho'-\rho$$

度量： $g_\rho:T_\rho\mathcal{P}\times T_\rho\mathcal{P}\rightarrow R$

$$g_\rho(\sigma_1,\sigma_2):=\langle M(\rho)\sigma_1\cdot\sigma_2\rangle=\int M(\rho)\sigma_1\cdot\sigma_2 d\theta$$

度量张量： $M(\rho):T_\rho\mathcal{P}\rightarrow T_\rho^*\mathcal{P}$

曲线距离: $\dot{\rho}_t=\sigma_t$

$D(\rho_0,\rho_1)^2=\int_0^1 g_\rho(\sigma_t,\sigma_t)dt$

度量(metric)

- Wasserstein 2度量

$$D(\rho_0,\rho_1)^2=\int_0^1\int \| v_t \|_2^2\ \rho_t d\theta dt$$

$$\frac{\partial\rho_t}{\partial t}=\sigma_t$$

其中v_t 满足：

$$\sigma_t=-\nabla\cdot(\rho_t v_t) \qquad v_t=\mathrm{argmin}_v\int \| v \|_2^2\ \rho_t d\theta$$

使用拉格朗日乘子法，我们可以进一步得到：

$$v_t=\nabla\psi_t \quad \sigma_t=-\nabla\cdot(\rho_t\nabla\psi_t)$$

度量(metric)

➤ Wasserstein 2度量

$$D(\rho_0,\rho_1)^2=\int_0^1\int \|\nabla\psi_t\|_2^2\,\rho_t d\theta dt$$

$$\frac{\partial\rho_t}{\partial t}+\nabla\cdot(\rho_t\nabla\psi_t)=0$$

$$D(\rho_0,\rho_1)^2=\int_0^1 g_\rho(\sigma_t,\sigma_t)dt$$

对比两边，我们有$\sigma_t=-\nabla\cdot(\rho_t\nabla\psi_t)$，我们有

$$g_{\rho_t}^W(\nabla\cdot(\rho_t\nabla\psi_t),\nabla\cdot(\rho_t\nabla\psi_t))=\int \|\nabla\psi_t\|_2^2\,\rho_t d\theta$$

使用定义 $g_\rho(\sigma_1,\sigma_2)=\int M(\rho)\sigma_1\cdot\sigma_2 d\theta$，我们得到

$$M^W(\rho)\sigma=\psi \qquad -\nabla\cdot(\rho\nabla\psi)=\sigma \qquad g_\rho^W(\sigma,\sigma)=\int\psi\sigma d\theta$$

度量(metric)

- Fisher-Rao 度量

$$D(\rho_0,\rho_1)^2 = \int_0^1 \int \frac{\sigma_t^2}{\rho_t}\, d\theta dt$$

$$\frac{\partial \rho_t}{\partial t} = \sigma_t$$

$$D(\rho_0,\rho_1)^2 = \int_0^1 g_\rho(\sigma_t,\sigma_t) dt$$

对比两边，我们有

$$g_{\rho_t}^{FR}(\sigma_t,\sigma_t) = \int \frac{\sigma_t^2}{\rho_t}\, d\theta$$

使用定义 $g_\rho(\sigma_1,\sigma_2) = \int M(\rho)\sigma_1 \cdot \sigma_2 d\theta$，我们得到

$$M^{FR}(\rho)\sigma = \psi = \frac{\sigma}{\rho} \qquad g_\rho^{FR}(\sigma,\sigma) = \int \frac{\sigma^2}{\rho} d\theta$$

度量(metric)

- 参数密度空间的度量

参数密度空间：$\mathcal{P} = \{a \in R^{N_a}\}$

线性切空间：$T_{\rho_a}\mathcal{P} = \{\sigma \in R^{N_a}\}$

$$\lim_{\epsilon \to 0} \frac{\rho_{a+\epsilon\sigma} - \rho_a}{\epsilon} = \nabla_a \rho_a \cdot \sigma$$

$$\mathfrak{g}_a(\sigma_1, \sigma_2) = g_{\rho_a}(\nabla_a \rho_a \cdot \sigma_1, \nabla_a \rho_a \cdot \sigma_2) = \sigma_1^T \mathfrak{M}(a) \sigma_2$$

练习：Fisher Rao 度量对应的参数密度空间的度量张量

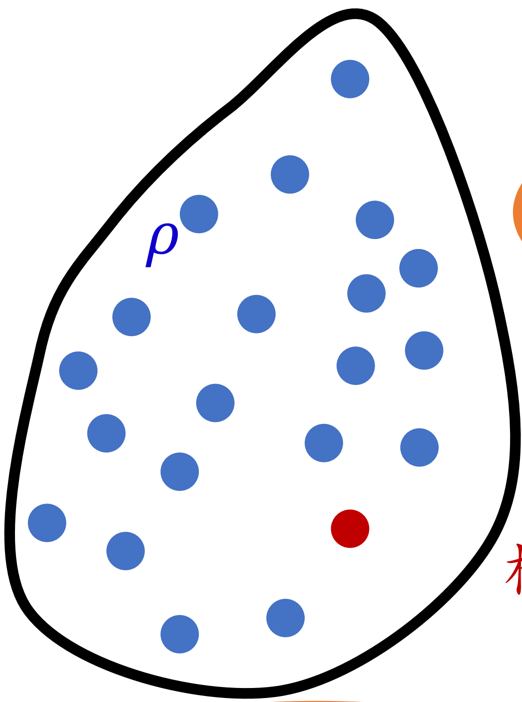
$\mathfrak{M}(a)$是什么？

- 变分推理

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

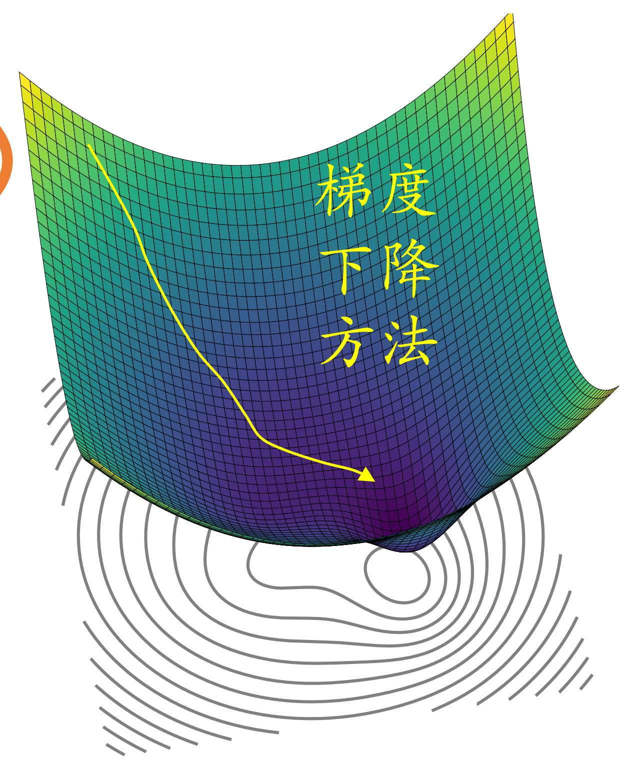
度量 $M(\rho)$

距离 $\mathcal{D}(\rho_A,\rho_B)$



极小值接近 ρ^*

概率密度空间 $\mathcal{P}$



24

变分推理

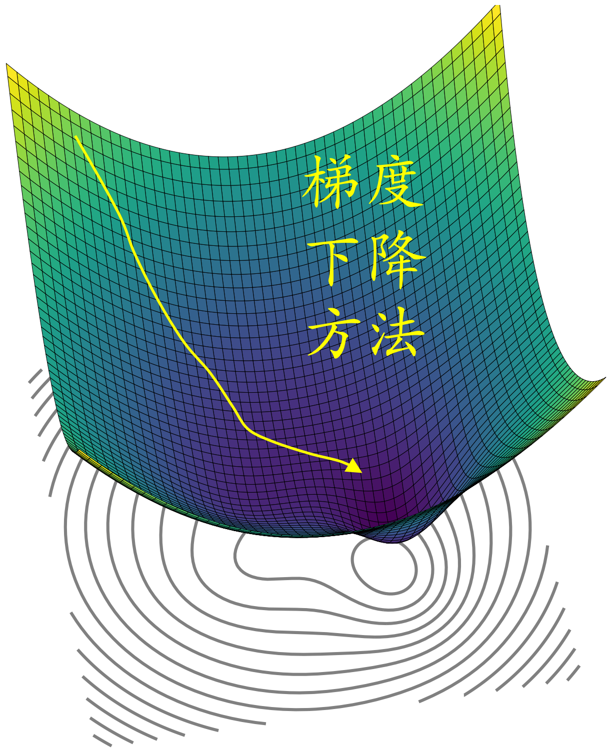
- 梯度流

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

$$\frac{\partial \rho_t}{\partial t} = \sigma_t$$

$$\sigma = \operatorname{argmin}_\sigma \lim_{\epsilon\to 0} \frac{\mathcal{E}(\rho+\epsilon\sigma;\rho^*) - \mathcal{E}(\rho;\rho^*)}{\epsilon\sqrt{g_\rho(\sigma,\sigma)}}$$

$$= \operatorname{argmin}_\sigma \frac{\left\langle \frac{\delta \mathcal{E}(\rho;\rho^*)}{\delta \rho}, \sigma \right\rangle}{\sqrt{\langle M(\rho)\sigma,\sigma\rangle}}$$

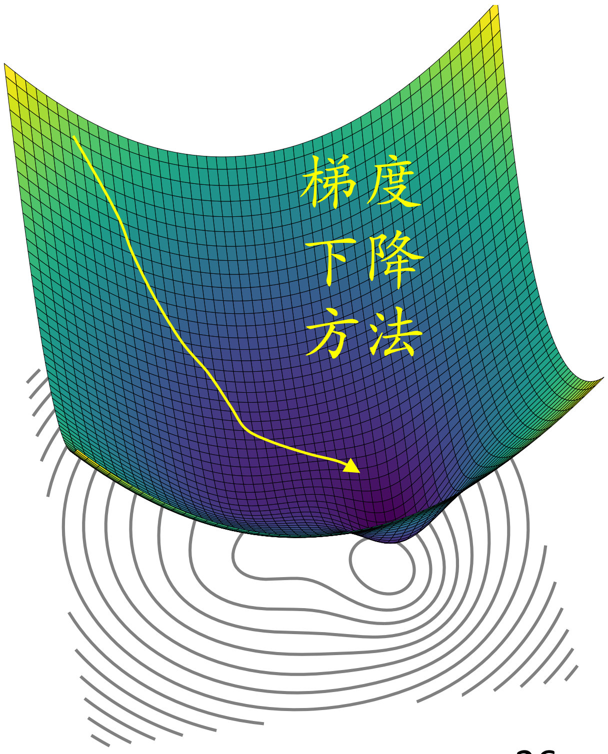


➢ 梯度流

$$\text{minimize}_a\, \mathcal{E}(\rho_a; \rho^*)$$

$$\frac{\partial a_t}{\partial t} = \sigma_t$$

$$\begin{aligned}\sigma &= \operatorname{argmin}_\sigma \lim_{\epsilon\to 0} \frac{\mathcal{E}(\rho_{a+\epsilon\sigma}; \rho^*) - \mathcal{E}(\rho_a; \rho^*)}{\epsilon\sqrt{\mathfrak{g}_a(\sigma,\sigma)}} \\ &= \operatorname{argmin}_\sigma \frac{\nabla_a \mathcal{E}(\rho_a; \rho^*)\sigma}{\sqrt{\sigma^T \mathfrak{M}(a)\sigma}} \\ &= -\mathfrak{M}(a)^{-1}\, \nabla_a \mathcal{E}(\rho_a; \rho^*)\end{aligned}$$



26

参数化变分推理

- 高斯变分推理

$$\min_{\rho_a} \mathrm{KL}[\rho_a \parallel \rho^*]$$

其中 $\rho_a = \mathcal{N}(\theta; m, C)$， $a = [m, C]$

$$\begin{aligned}\nabla_a \mathrm{KL}[\rho_a \parallel \rho^*] &= \int \nabla_a \rho_a \log\left(\frac{\rho_a}{\rho^*}\right) + \nabla_a \rho_a \, d\theta \\ &= \int \nabla_a \rho_a \left(\log(\rho_a) + \Phi_R\right) d\theta\end{aligned}$$

参数化变分推理

高斯变分推理

导数满足

$$\nabla_m \mathrm{KL}[\rho_a \parallel \rho^*] = \mathbb{E}_{\rho_a}[\nabla_\theta \Phi_R]$$

$$\nabla_C \mathrm{KL}[\rho_a \parallel \rho^*] = -\frac{1}{2} C^{-1} + \frac{1}{2} \mathbb{E}_{\rho_a}[\nabla_\theta \nabla_\theta \Phi_R]$$

梯度下降方法

$$\frac{dm_t}{dt} = -\mathbb{E}_{\rho_{a_t}}[\nabla_\theta \Phi_R]$$

$$\frac{dC_t}{dt} = \frac{1}{2} C_t^{-1} - \frac{1}{2} \mathbb{E}_{\rho_{a_t}}[\nabla_\theta \nabla_\theta \Phi_R]$$

参数化变分推理

- 自然梯度下降法(natural gradient descent)

$$\frac{\partial a_t}{\partial t} = -\mathfrak{M}(a_t)^{-1}\,\nabla_a \mathrm{KL}\left(\rho_{a_t} \parallel \rho^*\right)$$

$$\mathfrak{M}(a) = \mathrm{FIM}(\rho_a)$$

海瑟矩阵近似：

$$\mathrm{KL}[\rho_{a+da} \parallel \rho_a] \approx da^T \mathrm{FIM}(\rho_a) da$$

参数化变分推理

自然梯度下降法收敛性

Fisher-Rao度量下，我们有

$$\frac{dm_t}{dt} = -C_t \mathbb{E}_{\rho_{a_t}}[\nabla_\theta \Phi_R]$$

$$\frac{dC_t}{dt} = C_t - C_t \mathbb{E}_{\rho_{a_t}}[\nabla_\theta \nabla_\theta \Phi_R] C_t$$

当目标分布是高斯，$\Phi_R = -\frac{1}{2}(\theta - m^*)^T {C^*}^{-1}(\theta - m^*)$，自然梯度下降方法指数收敛

$$C_t^{-1} = {C^*}^{-1} + e^{-t}(C_0^{-1} - {C^*}^{-1})$$

$$m_t = m^* + e^{-t} C_t C_0^{-1}(m_0 - m^*)$$

参数化变分推理

- 练习

梯度下降方法：

$$\frac{dm_t}{dt} = -\mathbb{E}_{\rho_{a_t}}[\nabla_\theta \Phi_R] \qquad \frac{dC_t}{dt} = \frac{1}{2}C_t^{-1} - \frac{1}{2}\mathbb{E}_{\rho_{a_t}}[\nabla_\theta\nabla_\theta \Phi_R]$$

自然梯度下降法：

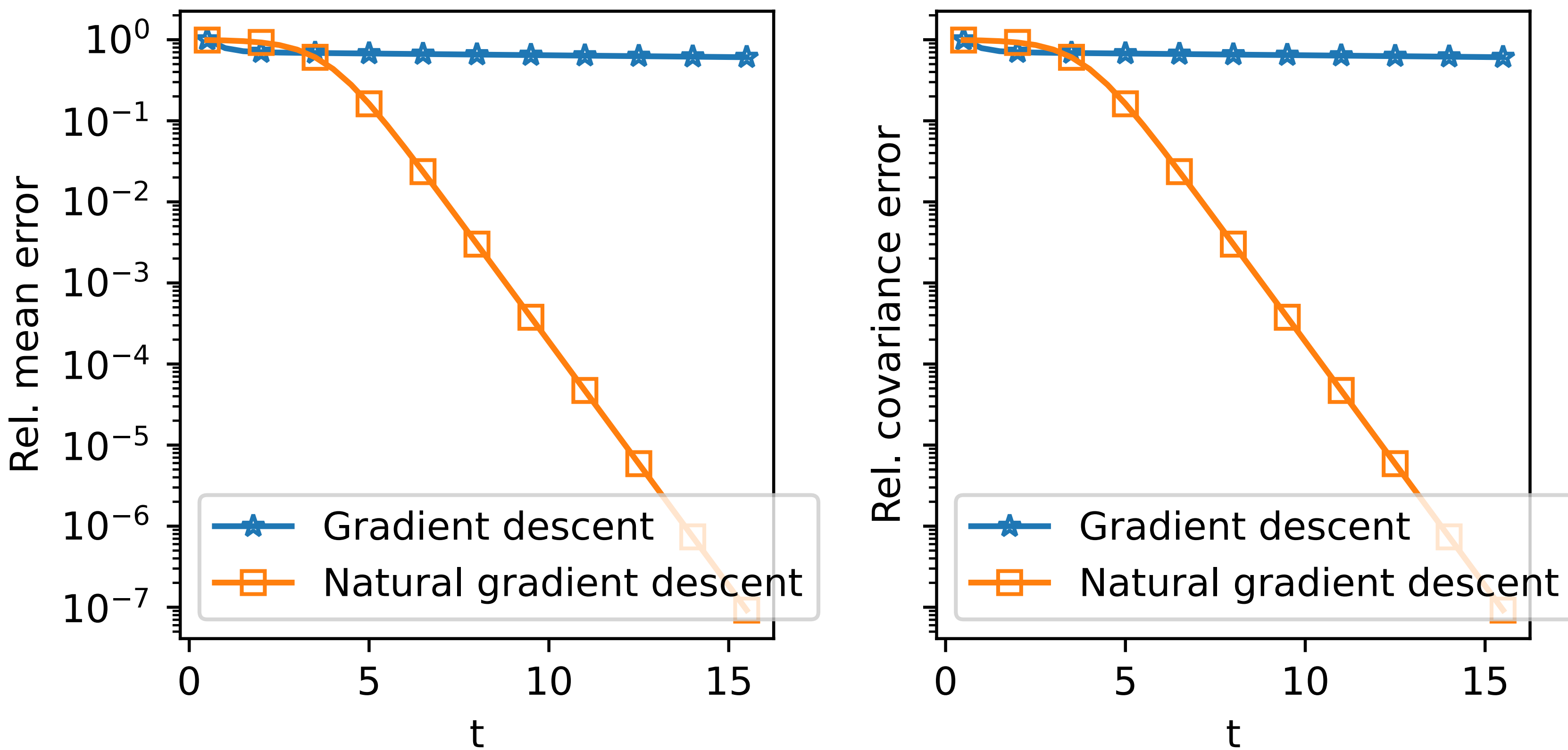
$$\frac{dm_t}{dt} = -C_t\mathbb{E}_{\rho_{a_t}}[\nabla_\theta \Phi_R] \qquad \frac{dC_t}{dt} = C_t - C_t\mathbb{E}_{\rho_{a_t}}[\nabla_\theta\nabla_\theta \Phi_R]C_t$$

$$\frac{dC_t^{-1}}{dt} = -C_t^{-1} + \mathbb{E}_{\rho_{a_t}}[\nabla_\theta\nabla_\theta \Phi_R]$$

其中目标分布是高斯，$\Phi_R = \frac{1}{2}(\theta - m^*)^T {C^*}^{-1}(\theta - m^*)$，$m^* = [1;1]$，$C^* = \mathrm{diag}\{100, 1\}$。初始值选取$m_0 = [0;0]$，$C_0 = \mathrm{diag}\{1,1\}$。

- 练习

目标分布是高斯，$\Phi_R = \frac{1}{2}(\theta - m^*)^T {C^*}^{-1}(\theta - m^*)$，$m^* = [1;1]$，$C^* = \mathrm{diag}\{100, 1\}$。初始值选取$m_0 = [0;0]$，$C_0 = \mathrm{diag}\{1,1\}$。



参数化变分推理

- 高斯变分推理

$$\min_{\rho_a} \mathrm{KL}[\rho_a \parallel \rho^*]$$

其中 $\rho_a = \mathcal{N}(\theta; m, C)$， $a = [m, C]$

极小值点满足

$$\nabla_m \mathrm{KL}[\rho_a \parallel \rho^*] = \mathbb{E}_{\rho_a}[\nabla_\theta \Phi_R] = 0$$

$$\nabla_C \mathrm{KL}[\rho_a \parallel \rho^*] = -\frac{1}{2} C^{-1} + \frac{1}{2} \mathbb{E}_{\rho_a}[\nabla_\theta \nabla_\theta \Phi_R] = 0$$

当Φ_R是强凸函数时，极小值点唯一。

参数化变分推理

- 凸函数

$$\nabla_x \nabla_x f \geq \alpha I \quad (\alpha > 0)$$

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\alpha}{2} \|y - x\|_2^2$$

- 对数强凹(log-concave)密度函数

$$\rho^*(\theta) = \frac{1}{Z} e^{-\Phi_R(\theta)} \qquad \nabla_\theta \nabla_\theta \Phi_R \geq \alpha I \quad (\alpha > 0)$$

KL散度在Wasserstein 度量下是强凸的(常速测底线ρ_0，ρ_1，σ_0)：

$$\mathrm{KL}[\rho_1 \parallel \rho^*] \geq \mathrm{KL}[\rho_0 \parallel \rho^*] + g_{\rho_0}(\nabla \mathrm{KL}[\rho_0 \parallel \rho^*], \sigma_0) + \frac{\alpha}{2} D(\rho_0, \rho_1)^2$$

参数化变分推理

- 高斯变分推理

$$\min_{\rho_a} \mathrm{KL}[\rho^* \parallel \rho_a]$$

其中 $\rho_a = \mathcal{N}(\theta; m, C)$， $a = [m, C]$

极小值点唯一，满足 $m = \mathbb{E}_{\rho^*}[\theta]$， $C = \mathrm{Cov}_{\rho^*}[\theta]$

$$\nabla_m \mathrm{KL}[\rho^* \parallel \rho_a] = -C^{-1}(\mathbb{E}_{\rho^*}[\theta] - m)$$

$$\nabla_C \mathrm{KL}[\rho^* \parallel \rho_a]$$

$$= \frac{1}{2}C^{-1} - \frac{1}{2}C^{-1}\left[\mathrm{Cov}_{\rho^*}[\theta] + \left(\mathbb{E}_{\rho^*}[\theta] - m\right)\left(\mathbb{E}_{\rho^*}[\theta] - m\right)^T\right]C^{-1}$$

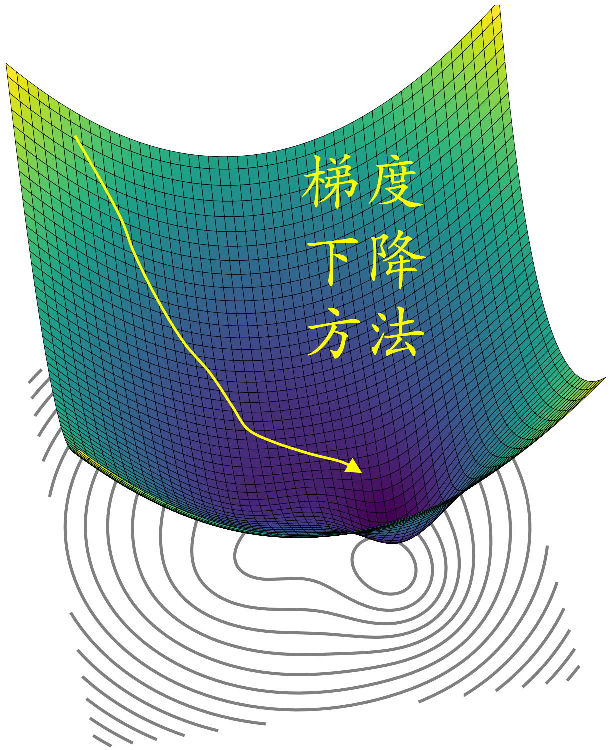
➢ 梯度流

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

$$\frac{\partial \rho_t}{\partial t} = \sigma_t$$

$$\sigma = \mathrm{argmin}_{\sigma} \lim_{\epsilon \to 0} \frac{\mathcal{E}(\rho+\epsilon\sigma;\rho^*) - \mathcal{E}(\rho;\rho^*)}{\epsilon\sqrt{g_{\rho}(\sigma,\sigma)}}$$

$$= \mathrm{argmin}_{\sigma} \frac{\left\langle \frac{\delta\mathcal{E}(\rho;\rho^*)}{\delta\rho}, \sigma \right\rangle}{\sqrt{\langle M(\rho)\sigma,\sigma\rangle}}$$



36

非参数化变分推理

- 第一变分（Frechet 导数）

$\frac{\delta \mathrm{KL}}{\delta \rho} \in T_\rho^* \mathcal{P}$ 对切向量$\sigma(\theta) \in T_\rho \mathcal{P}$的作用满足

$$\int \frac{\delta \mathrm{KL}}{\delta \rho} \sigma d\theta = \lim_{\epsilon \to 0} \frac{\mathrm{KL}[\rho + \epsilon\sigma \parallel \rho^*] - \mathrm{KL}[\rho \parallel \rho^*]}{\epsilon}$$

因此

$$\frac{\delta \mathrm{KL}[\rho \parallel \rho^*]}{\delta \rho} = \log \rho - \log \rho^* + \mathrm{const.}$$

- 梯度流

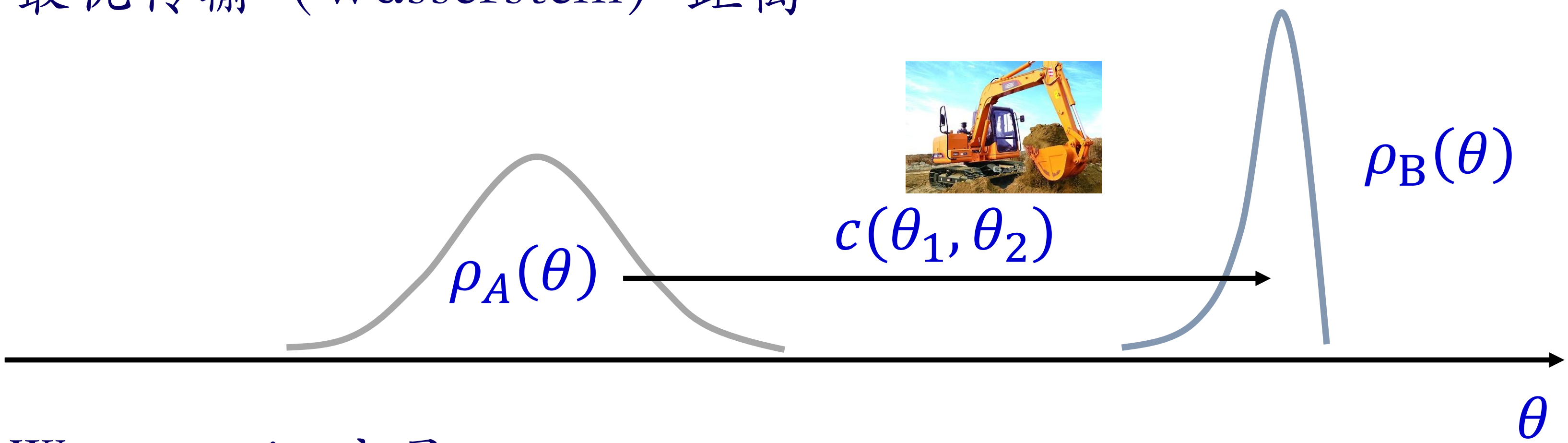
$$\frac{\partial \rho_t}{\partial t} = -M(\rho_t)^{-1} \frac{\delta \mathrm{KL}}{\delta \rho_t}$$

我们有

$$\frac{\partial \mathrm{KL}(\rho_t)}{\partial t} = -\int \frac{\delta \mathrm{KL}}{\delta \rho_t} M(\rho_t)^{-1} \frac{\delta \mathrm{KL}}{\delta \rho_t} d\theta \le 0$$

非参数化变分推理

- 最优传输（Wasserstein）距离



- Wasserstein 度量

$$M(\rho)^{-1}\psi = -\nabla_\theta \cdot (\rho \nabla_\theta \psi)$$

- Wasserstein 梯度流

$$\frac{\partial \rho_t}{\partial t} = -M(\rho_t)^{-1} \frac{\delta \mathrm{KL}[\rho_t \parallel \rho^*]}{\delta \rho_t}$$
$$= -\nabla_\theta \cdot [\rho_t (\nabla_\theta \log \rho^* - \nabla_\theta \log \rho_t)]$$

非参数化变分推理

- Wasserstein 梯度流

$$\frac{\partial \rho_t}{\partial t} = -\nabla_\theta \cdot [\rho_t(\nabla_\theta \log \rho^* - \nabla_\theta \log \rho_t)]$$

$$\frac{\partial \rho_t}{\partial t} = \nabla_\theta \cdot [\rho_t \nabla_\theta \Phi_R + \nabla_\theta \rho_t]$$

粒子系统 $\theta_t \sim \rho_t$

- Langevin 动力系统

$$d\theta_t = -\nabla_\theta \Phi_R + \sqrt{2} dW_t$$

$$\theta_{n+1} = \theta_n - \epsilon \nabla_\theta \Phi_R + \sqrt{2} \mathcal{N}(0, \epsilon)$$

非参数化变分推理

Langevin 动力系统

假设$\theta_0 \sim \rho_0$，对于Langevin 动力系统

$$d\theta_t = -\nabla_\theta \Phi_R + \sqrt{2} dW_t$$

那么$\theta_t \sim \rho_t$， ρ_t满足

$$\frac{\partial \rho_t}{\partial t} = \nabla_\theta \cdot [\rho_t \nabla_\theta \Phi_R + \nabla_\theta \rho_t]$$

Metropolis-Hastings 算法

- Metropolis-Hastings 算法

提议核： $q(\cdot,\cdot): R^{N_\theta \times N_\theta} \rightarrow R^+$

修正： $a(\theta,\theta') = \min\left\{\frac{\rho^*(\theta')q(\theta',\theta)}{\rho^*(\theta)q(\theta,\theta')}，1\right\}$

- Metropolis-Adjusted Langevin 算法

$\rho^*(\theta) \propto e^{-\Phi_R(\theta)}$

梯度下降方法： $\theta \rightarrow \theta - \epsilon\nabla_\theta\Phi_R(\theta)$

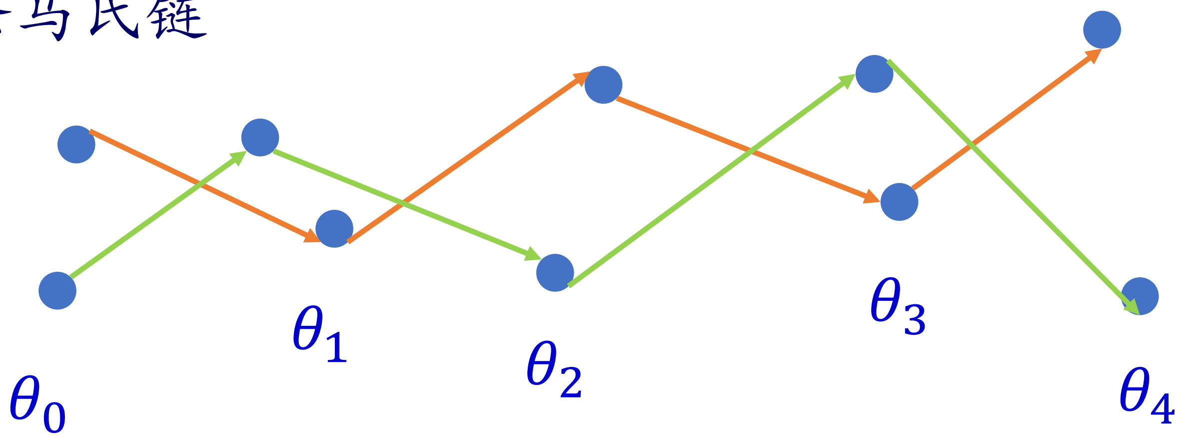
$\Phi_R\big(\theta - \epsilon\nabla_\theta\Phi_R(\theta)\big) < \Phi_R(\theta) \qquad \rho^*\big(\theta - \epsilon\nabla_\theta\Phi_R(\theta)\big) > \rho^*(\theta)$

$$q(\theta,\theta') = \mathcal{N}(\theta';\theta - \epsilon\nabla_\theta\Phi_R(\theta),\delta^2 I)$$

$$a(\theta,\theta') = \min\left\{\frac{\rho^*(\theta')\mathcal{N}(\theta;\ \theta' - \epsilon\nabla_\theta\Phi_R(\theta'),\delta^2 I)}{\rho^*(\theta)\mathcal{N}(\theta';\theta - \epsilon\nabla_\theta\Phi_R(\theta),\delta^2 I)}，1\right\}$$

马氏链蒙特卡洛方法

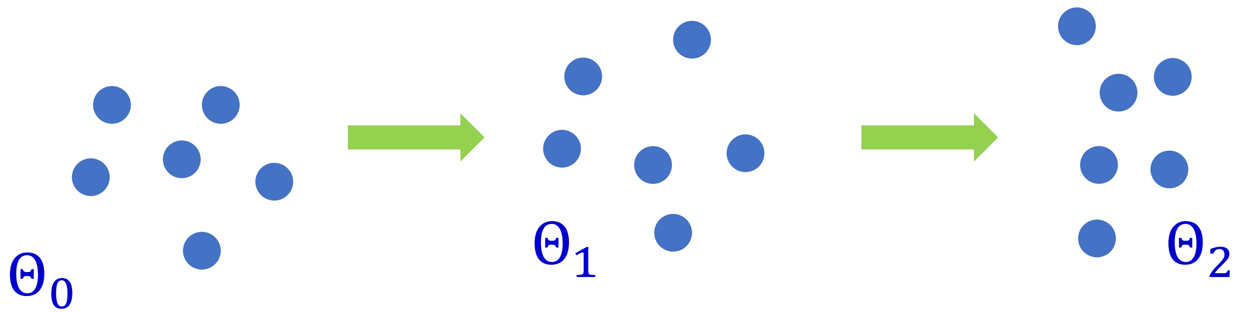
- 两条马氏链



- 交互粒子系统

$$\Theta_n = [\theta_n^1; \theta_n^2; \cdots \theta_n^J] \in R^{N_\theta} \otimes R^{N_\theta} \cdots \otimes R^{N_\theta}$$

$$\mathrm{P}^* = \rho^* \otimes \rho^* \cdots \otimes \ \rho^*$$



非参数化变分推理

- Kalman-Wasserstein 梯度流

$$\frac{\partial \rho_t}{\partial t} = \nabla_\theta \cdot [\rho_t C_t (\nabla_\theta \log \rho^* - \nabla_\theta \log \rho_t)]$$

$$\frac{\partial \rho_t}{\partial t} = \nabla_\theta \cdot [\rho_t C_t (\nabla_\theta \Phi_R + \nabla_\theta \log \rho_t)]$$

- 预处理 Langevin 动力系统

$$d\theta_t = -C_t \nabla_\theta \Phi_R + \sqrt{2C_t}\, dW_t$$

$$\theta_{n+1}^j = \theta_n^j - \epsilon C_n \nabla_\theta \Phi_R(\theta_n^j) + \sqrt{2C_n}\, \mathcal{N}(0, \epsilon)$$

非参数化变分推理

仿射不变性

我们的梯度流

$$\frac{\partial \rho_t}{\partial t} = \nabla_\theta \cdot [\rho_t C_t (\nabla_\theta \log \rho^* - \nabla_\theta \log \rho_t)]$$

对于任意可逆仿射变换 $\mathcal{T}: \tilde{\theta} \to A\theta + b$，我们的梯度流满足

$$\frac{\partial \tilde{\rho}_t}{\partial t} = \nabla_{\tilde{\theta}} \cdot [\tilde{\rho}_t \tilde{C}_t (\nabla_{\tilde{\theta}} \log \tilde{\rho}^* - \nabla_{\tilde{\theta}} \log \tilde{\rho}_t)]$$

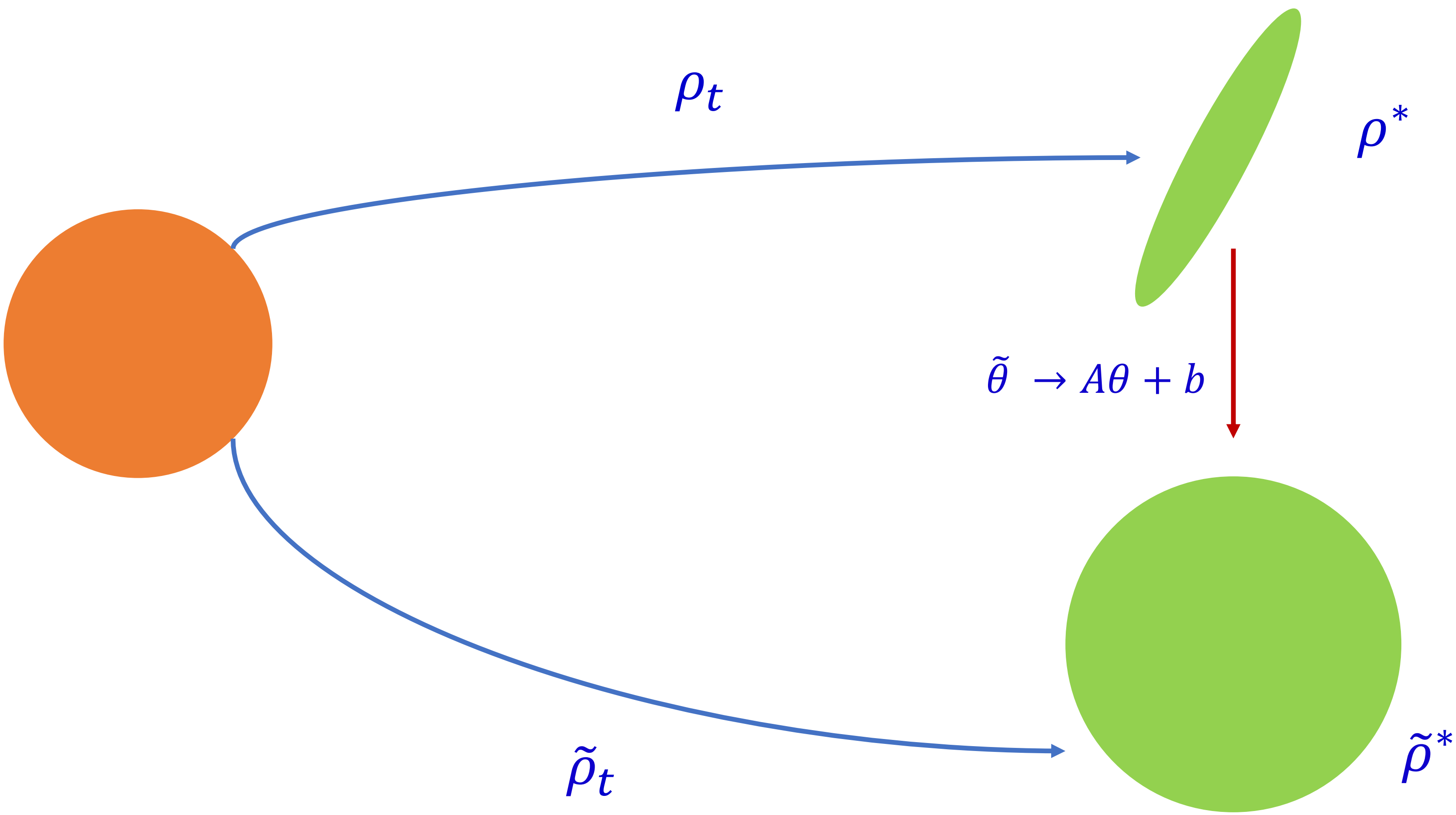
其中

$$\tilde{\rho}\,(\tilde{\theta}) = \rho\left(\mathcal{T}^{-1}(\tilde{\theta})\right) |\nabla_{\tilde{\theta}} \mathcal{T}^{-1}(\tilde{\theta})|$$

$$\tilde{C}_t = A C_t A^T$$

非参数化变分推理

- 仿射不变性

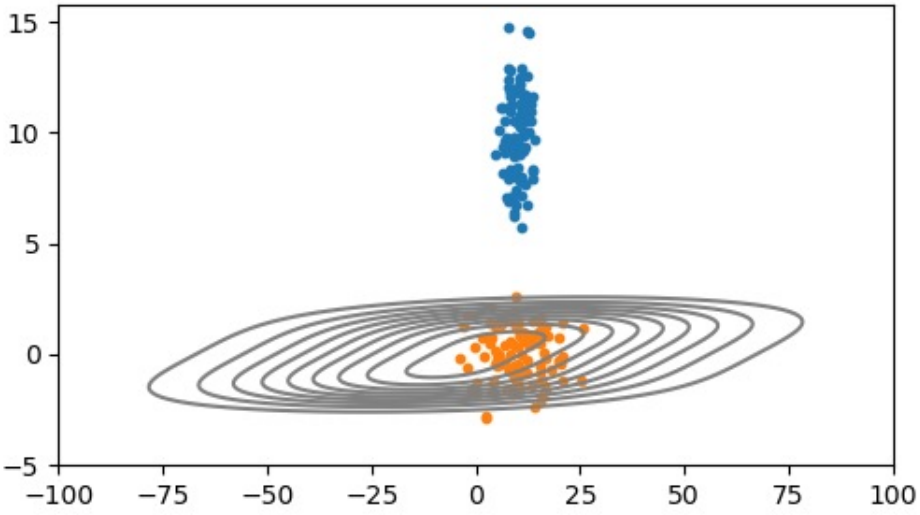


- 练习

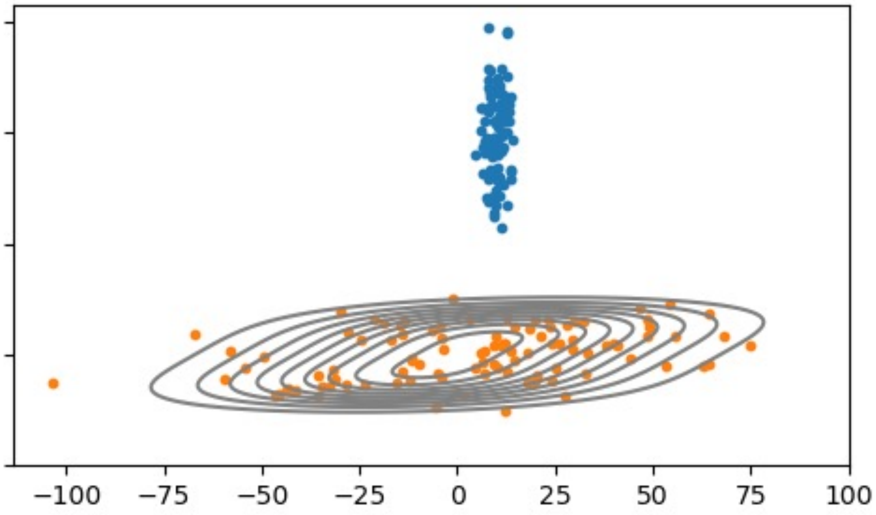
$$\rho^*(\theta) = \frac{1}{Z} e^{-\Phi_R(\theta)}$$

$$\Phi_R(\theta) = \frac{(0.1\theta_1 - \theta_2)^2 + \theta_2^4}{20} \qquad \rho_0 \sim \mathcal{N}([10; 10], 4I)$$

Langevin 动力



预处理 Langevin 动力



非参数化变分推理

- Fisher-Rao 度量

$$M(\rho)\sigma = \psi = \frac{\sigma}{\rho}$$

$$M(\rho)^{-1}\psi = \rho\psi - \mathbb{E}_\rho \psi$$

- Fisher-Rao 梯度流

$$\frac{\partial \rho_t}{\partial t} = -M(\rho_t)^{-1} \frac{\delta \mathrm{KL}[\rho_t \parallel \rho^*]}{\delta \rho_t}$$
$$= \rho_t(\log \rho^* - \log \rho_t) - \rho_t \mathbb{E}[\log \rho^* - \log \rho_t]$$

非参数化变分推理

不变性

我们的梯度流

$$\frac{\partial \rho_t}{\partial t} = \rho_t(\log \rho^* - \log \rho_t) - \rho_t \mathbb{E}[\log \rho^* - \log \rho_t]$$

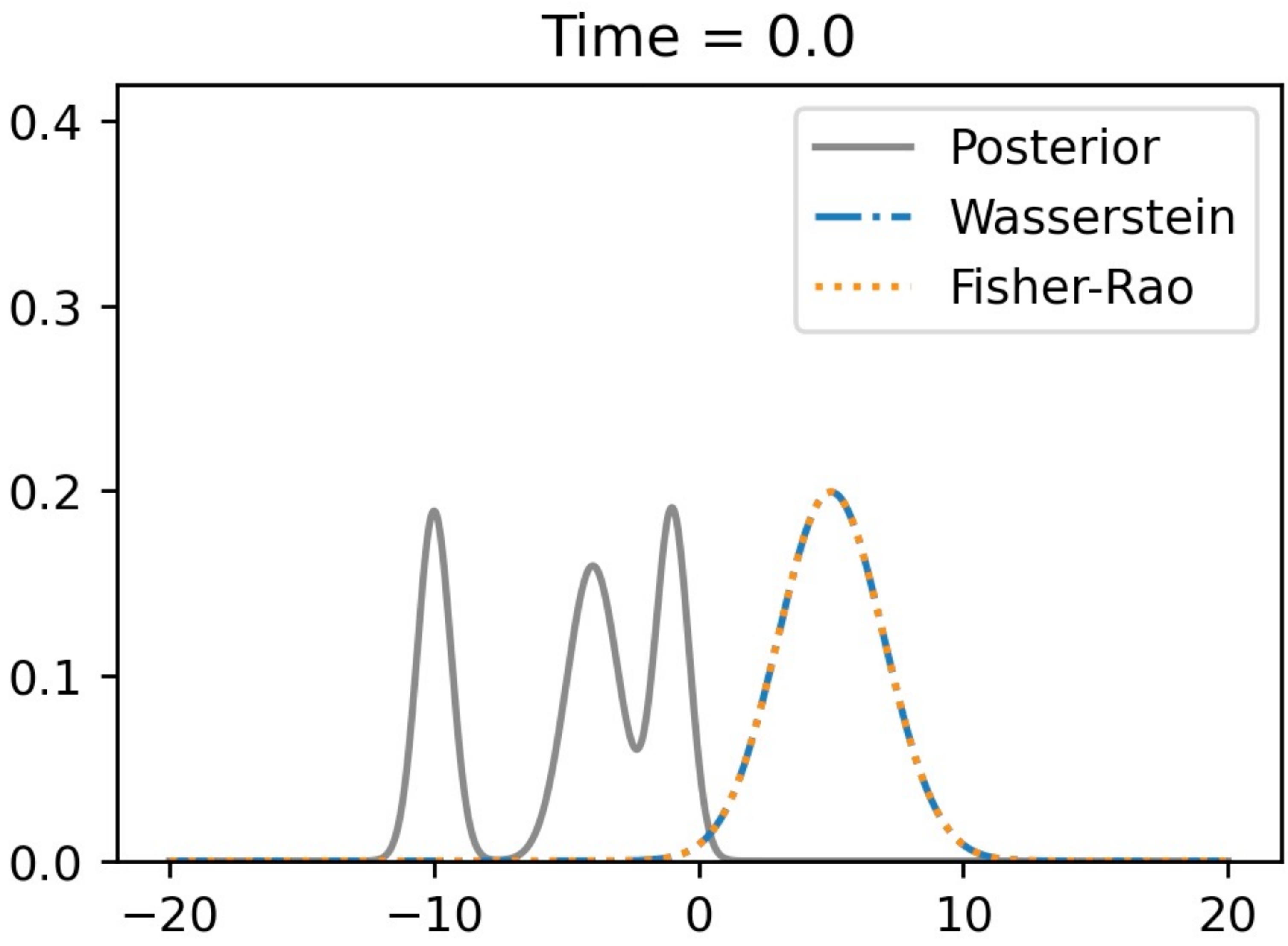
对于任意可逆变换$\mathcal{T}: \theta \to \tilde{\theta}$，我们的梯度流满足

$$\frac{\partial \tilde{\rho}_t(\tilde{\theta})}{\partial t} = \tilde{\rho}_t(\log \tilde{\rho}^* - \log \tilde{\rho}_t) - \tilde{\rho}_t \mathbb{E}_{\tilde{\rho}_t}(\log \tilde{\rho}^* - \log \tilde{\rho}_t)$$

其中

$$\tilde{\rho}\left(\tilde{\theta}\right) = \rho\left(\mathcal{T}^{-1}(\tilde{\theta})\right) |\nabla_{\tilde{\theta}} \mathcal{T}^{-1}(\tilde{\theta})|$$

Gradient Flow for Sampling



扩展阅读

梯度流(gradient flow)

综述: Chen, Yifan, et al. "Sampling via Gradient Flows in the Space of Probability Measures." arXiv preprint arXiv:2310.03597 (2023).

综述: Trillos, N. García, Bamdad Hosseini, and Daniel Sanz-Alonso. "From optimization to sampling through gradient flows." NOTICES OF THE AMERICAN MATHEMATICAL SOCIETY 70.6 (2023).

参数化变分推理

对于对数凹目标函数，自然梯度下降方法的收敛速度。

高斯变分推理：Opper, Manfred, and Cédric Archambeau. "The variational Gaussian approximation revisited." Neural computation 21.3 (2009): 786-792.

高斯Wasserstein梯度下降方法： Lambert, Marc, et al. "Variational inference via Wasserstein gradient flows. Advances in Neural Information Processing Systems 35 (2022): 14434-14447.

综述(统计学角度)：Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. "Variational inference: A review for statisticians." Journal of the American statistical Association 112.518 (2017): 859-877.

扩展阅读

综述(统计学角度) Wainwright, Martin J., and Michael I. Jordan. "Graphical models, exponential families, and variational inference." Foundations and Trends® in Machine Learning 1.1 – 2 (2008): 1-305.

➢ 非参数化梯度流

Stein梯度流：Liu, Qiang, and Dilin Wang. "Stein variational gradient descent: A general purpose bayesian inference algorithm." Advances in neural information processing systems 29 (2016).

Kalman-Wasserstein梯度流：Garbuno-Inigo, Alfredo, et al. "Interacting Langevin diffusions: Gradient structure and ensemble Kalman sampler." SIAM Journal on Applied Dynamical Systems 19.1 (2020): 412-441.

理论：Log-concave sampling (https://chewisinho.github.io).

Wasserstein-Fisher-Rao 梯度流： Lu, Yulong, Jianfeng Lu, and James Nolen. "Accelerating langevin sampling with birth-death." arXiv preprint arXiv:1905.09863 (2019).

Fisher-Rao 梯度流： Maurais, Aimee, and Youssef Marzouk. "Sampling in unit time with kernel Fisher-Rao flow." arXiv preprint arXiv:2401.03892 (2024).

扩展阅读

最优输运算法

Cuturi, Marco. "Sinkhorn distances: Lightspeed computation of optimal transport." Advances in neural information processing systems 26 (2013).

Altschuler, Jason, Jonathan Niles-Weed, and Philippe Rigollet. "Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration." Advances in neural information processing systems 30 (2017).

自然梯度下降法

优化：Amari, Shun-Ichi. "Natural gradient works efficiently in learning." Neural computation 10.2 (1998): 251-276.

综述：Martens, James. "New insights and perspectives on the natural gradient method." Journal of Machine Learning Research 21.146 (2020): 1-76.