变分推理方法
（VARIATIONAL INFERENCE）

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

- 变分推理的要素
  - 概率密度空间
  - 能量泛函
  - 度量、距离
- Wasserstein 梯度流
  - Langevin 动力学
  - 高斯变分推理
- Fisher-Rao 梯度流
  - 生灭过程
  - 自然梯度下降法
  - 指数分布族
- 仿射不变性

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

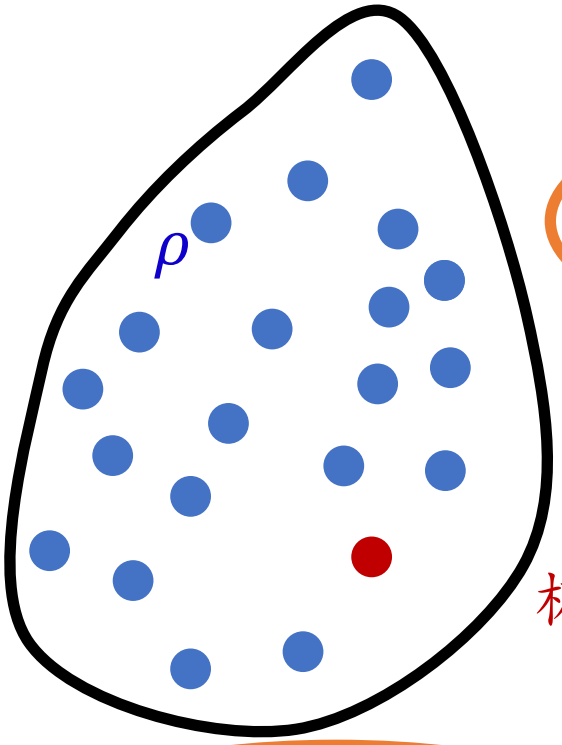
$$\Phi_R(\theta, y) = \frac{1}{2} \| \Sigma_\eta^{-\frac{1}{2}} (y - G(\theta)) \|^2 + \frac{1}{2} \| \Sigma_0^{-\frac{1}{2}} (\theta - r_0) \|^2$$

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

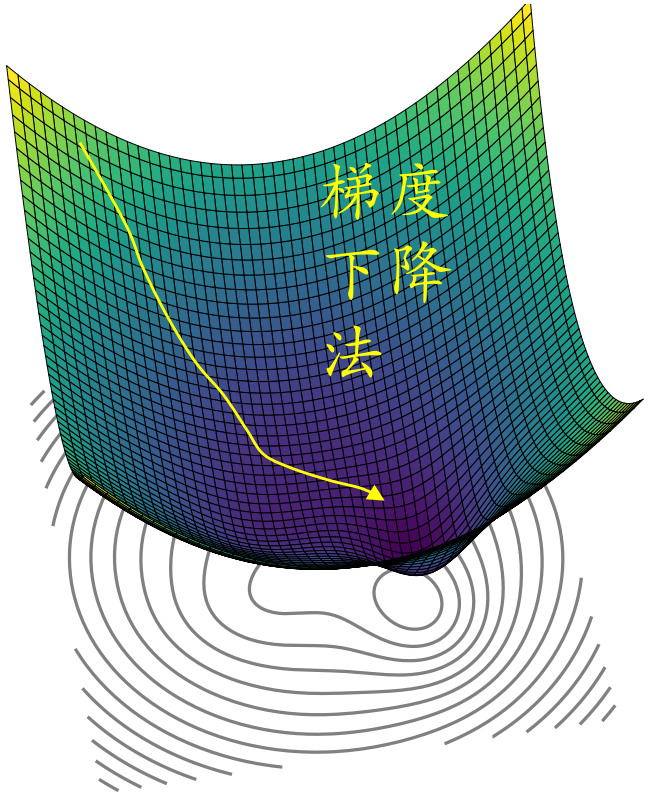
变分推理

- 变分推理

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



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



极小值接近 ρ^*

概率测度空间$\mathcal{P}$

4

能量泛函 (energy functional)

- 概率测度空间上的能量泛函

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

满足：

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

- 能量泛函不一定是距离
- 能量泛函用于区分两个分布、量化两个分布的差异

- KL-散度

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

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

最速梯度下降法

- 最速梯度下降法

$$\text{minimize}_x \ f(x)$$

$$x_{k+1} = x_k - \alpha \nabla f(x_k)$$

$$\nabla f(x_k) = \underset{v}{\text{argmax}} \ \frac{\frac{\delta f}{\delta x}(x_k) \cdot v}{\sqrt{\langle v, v \rangle}}$$

函数的变化量

$$\frac{\delta f}{\delta x}(x_k) \cdot v = \lim_{\epsilon \to 0} \frac{f(x_k + \epsilon v) - f(x_k)}{\epsilon}$$

长度

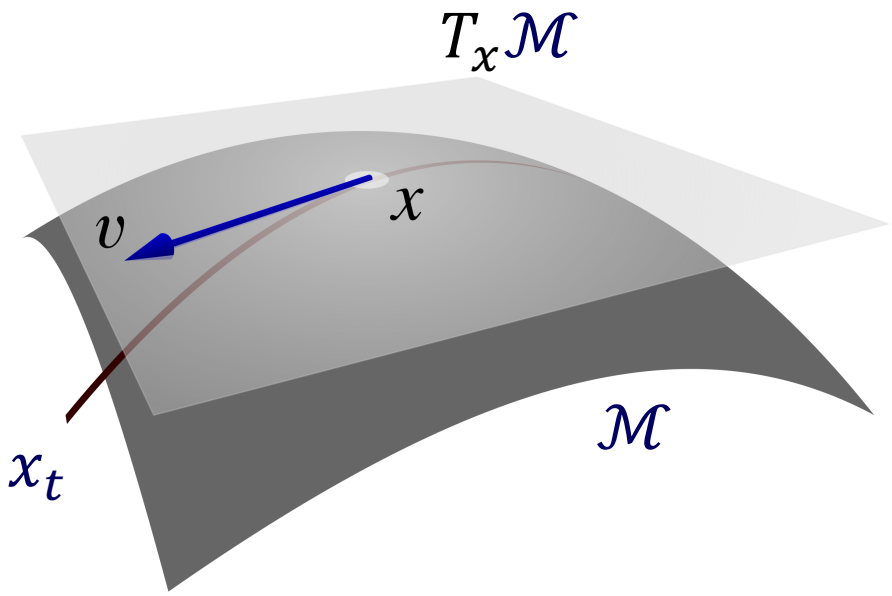
最速梯度下降法

- 流形上的最速梯度下降法

$$\text{minimize}_{x\in\mathcal{M}}\ f(x)$$

在流形 $\mathcal{M}$（比如三维空间的曲面、概率测度空间等）上，

- 如何定义长度
- 如何定义导数



度量

- 流形上的度量
  - 在点x的切空间：$T_x\mathcal{M}$
  - 方向向量：$v \in T_x\mathcal{M}$
  - 曲线：$x_t: [0,1] \to \mathcal{M}$

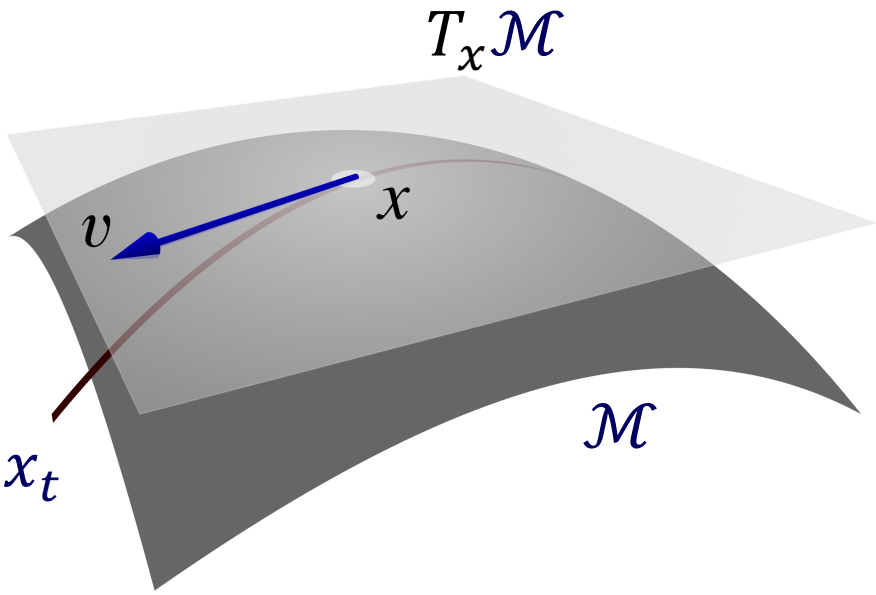
$$\dot{x}_t = v_t$$

  - 曲线长度：$\int_0^1 \sqrt{g_x(v_t, v_t)}dt$
  - 度量：$g_x: T_x\mathcal{M} \times T_x\mathcal{M} \to R_{\geq 0}$
  - 度量矩阵：$M(x): T_x\mathcal{M} \to T_x^*\mathcal{M}$

$$g_x(v, v) = \langle M(x)v, v \rangle$$

  - 测地线 (geodesics)：过两点

距离最短的曲线



度量

- 练习：三维欧式空间 R^3

在点x的切空间：

$$T_x R^3 = R^3$$

度量：$g_x(v_t, v_t) = \langle v_t, v_t \rangle, \quad M(x) = I$

曲线：$\gamma: [0,1] \to R^3$

$$\dot{x}_t = v_t$$

曲线长度：$\int \sqrt{v_t \cdot v_t} dt = \int \sqrt{g_x(v_t, v_t)} dt$

度量

- 练习：欧式空间R^3中的球面S^2

球坐标(φ,θ)表示：

$$(\varphi,\theta)\to(\sin\varphi\cos\theta\,,\sin\varphi\sin\theta\,,\cos\varphi)$$

在点x的切空间能参数化为：

$$T_xS^2=R^2$$

曲线：$\gamma\colon[0,1]\to S^2$

$$\dot{x}_t=v_t\ ,\ v_t=(\dot{\varphi}_t,\dot{\theta}_t)$$

度量：$g_x(v_t,v_t)=\dot{\varphi}_t^2+\dot{\theta}_t^2\sin^2\varphi=\langle M_xv_t,v_t\rangle$

$$M(x)=\begin{bmatrix}1 & \\ & \sin^2\varphi\end{bmatrix}$$

度量

- 练习：庞加莱圆盘D

$$D = \{x \in R^2, \|x\|_2 \leq 1\}$$

在点x的切空间能参数化为：

$$T_x D = R^2$$

曲线 $\gamma: [0,1] \to D$:　　$\dot{x}_t = v_t$

度量：$g_x(v_t, v_t) = \langle M(x)v_t, v_t \rangle$

$$M(x) = \begin{bmatrix} \frac{4}{(1-x_1^2-x_2^2)^2} & \\ & \frac{4}{(1-x_1^2-x_2^2)^2} \end{bmatrix}$$

度量

- 练习：庞加莱圆盘D

极坐标(r,θ)表示：$(r,\theta)\to(r\cos\theta\,,r\sin\theta)$

在点x的切空间能参数化为：

$$T_xD=R^2$$

曲线$\gamma:[0,1]\to D$:　　$\dot{x}_t=v_t,\ \ v_t=(\dot{r}_t,\dot{\theta}_t)$

度量：$g_x(v_t,v_t)=\langle M(x)v_t,v_t\rangle$

$$M(x)=\begin{bmatrix}\frac{4}{(1-r^2)^2} & \\ & \frac{4r^2}{(1-r^2)^2}\end{bmatrix}$$

度量矩阵的形式会随内积定义而改变！

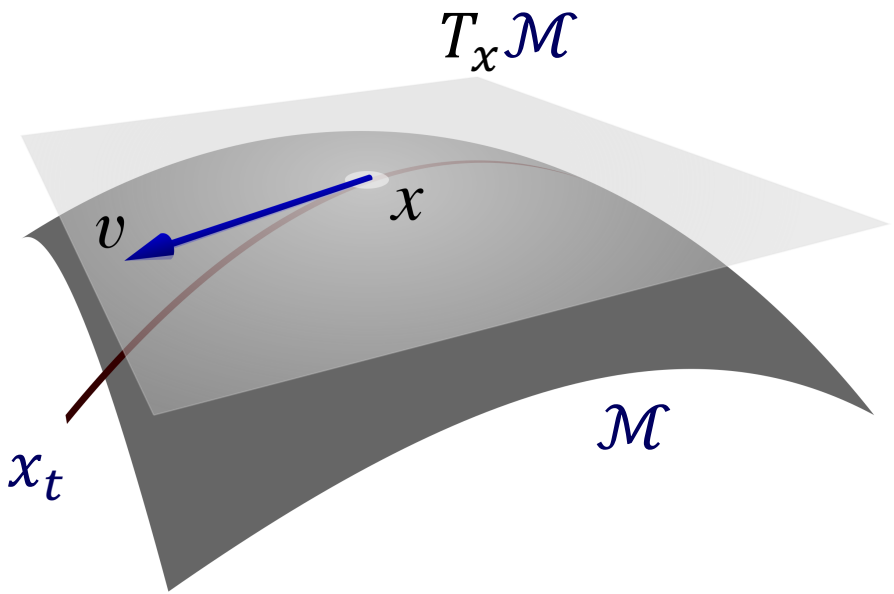
最速梯度下降法

- 流形上的最速梯度下降法

$$\text{minimize}_{x\in\mathcal{M}}\ f(x)$$

在流形 $\mathcal{M}$（比如三维空间的曲面、概率测度空间等）上，

- 如何定义长度
- 如何定义导数



度量

- 流形上的度量

流形上的函数或泛函 (functional)

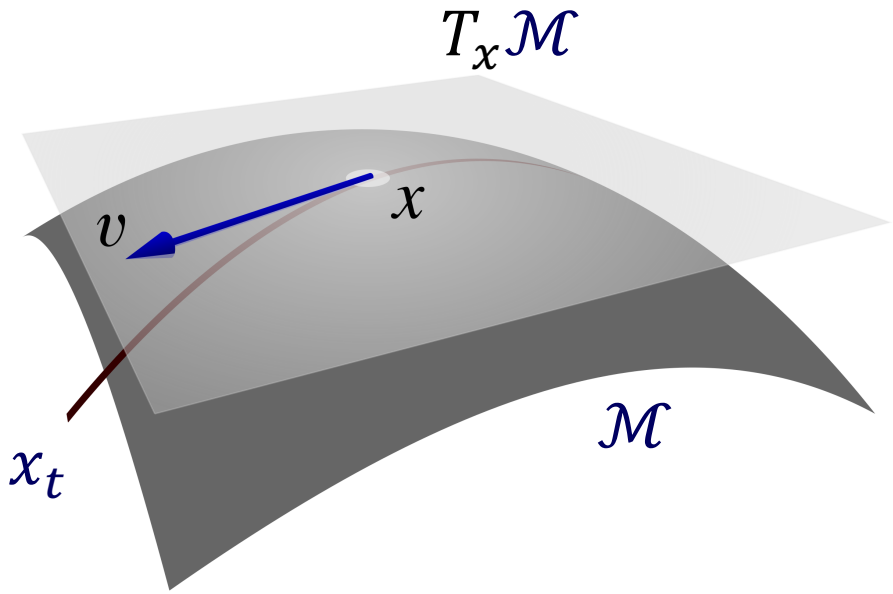
$$f : \mathcal{M} \to R$$

第一变分 (first variation) $v \in T_x\mathcal{P}$

$$\frac{\delta f}{\delta x} v = \lim_{\epsilon \to 0} \frac{f(x + \epsilon v) - f(x)}{\epsilon}$$

Gâteaux 导数：

$$\frac{\delta f}{\delta x} \in T_x^*\mathcal{P}$$



最速梯度下降法

- 流形上的最速梯度下降法

$$\text{minimize}_{x\in\mathcal{M}}\ f(x)$$

$$x_{k+1} = x_k - \alpha\,\nabla f(x_k)$$

$$\nabla f(x_k) = \underset{v}{\text{argmin}}\ \frac{\frac{\delta f}{\delta x}(x_k)\cdot v}{\sqrt{\langle M(x_k)v, v\rangle}}$$

$$= M(x_k)^{-1}\frac{\delta f}{\delta x}(x_k)$$

函数的变化量

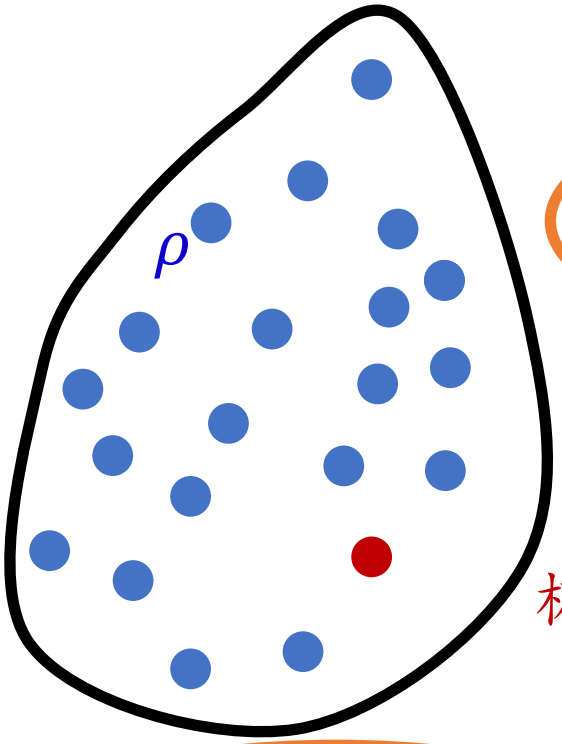
$$\frac{\delta f}{\delta x}(x_k)\cdot v = \lim_{\epsilon\to 0}\frac{f(x_k+\epsilon v) - f(x_k)}{\epsilon}$$

长度

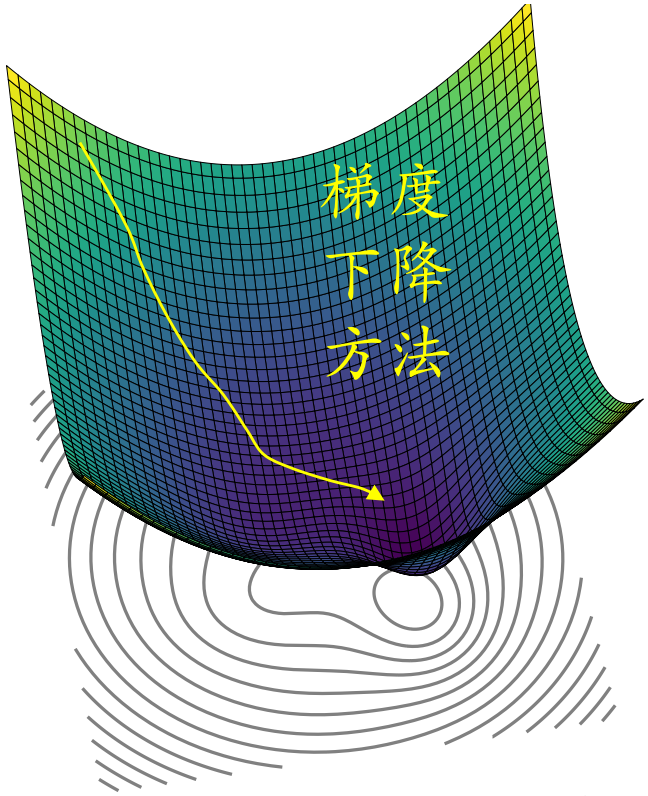
变分推理

- 变分推理

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



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



极小值接近 ρ^*

概率测度空间$\mathcal{P}$

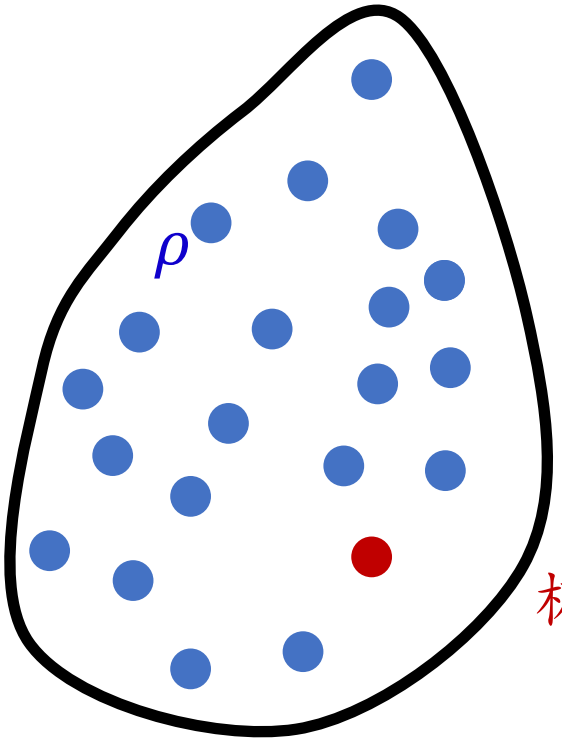
16

概率测度空间

- 概率测度空间

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

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



极小值 ρ^*

概率测度空间$\mathcal{P}$

能量泛函 (energy functional)

- KL-散度

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

第一变分 (first variation) $\sigma \in T_\rho \mathcal{P}$

$$\frac{\delta \mathcal{E}}{\delta \rho}\sigma = \lim_{\epsilon \to 0} \frac{\mathcal{E}(\rho + \epsilon\sigma) - \mathcal{E}(\rho)}{\epsilon}$$

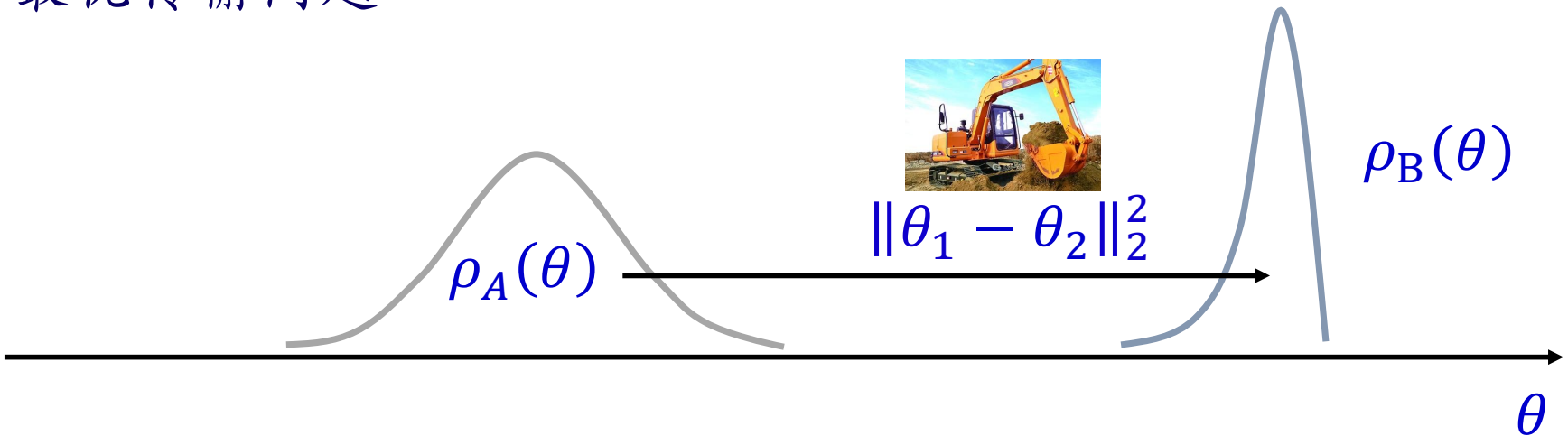
$$= \int \sigma \log\left(\frac{\rho}{\rho^*}\right) d\theta$$

Gâteaux 导数：

$$\frac{\delta \mathcal{E}}{\delta \rho} = \log \rho - \log \rho^* - \mathbb{E}_\rho[\log \rho - \log \rho^*]$$

Wasseratein-2度量

- 最优传输问题



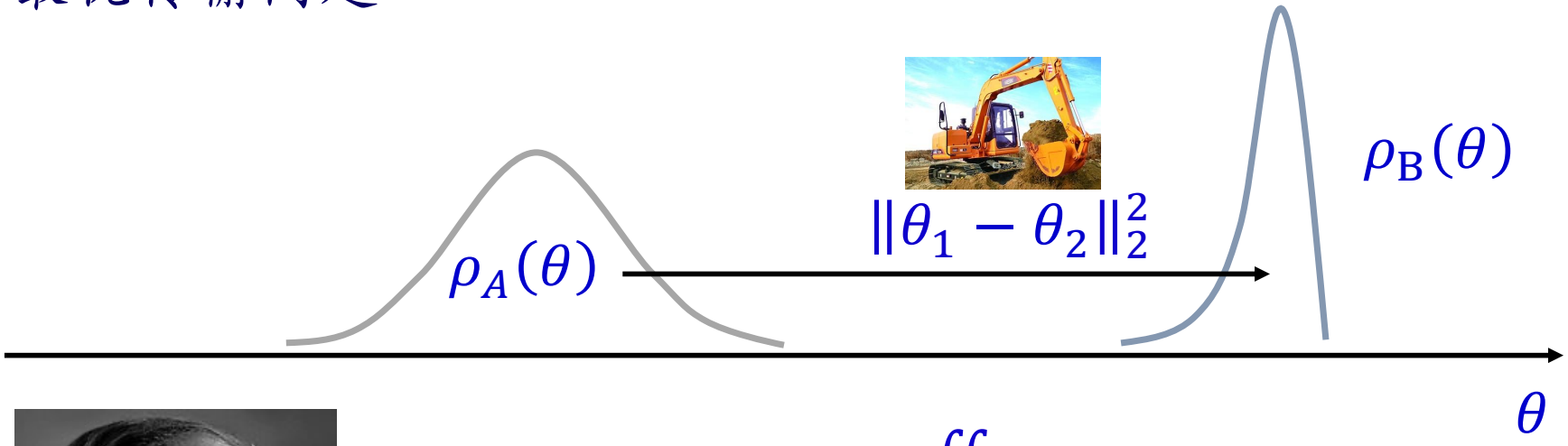
$$W_2^2(\rho_A, \rho_B) = \inf_T \iint \|\theta - T(\theta)\|_2^2 \rho_A(\theta) d\theta$$

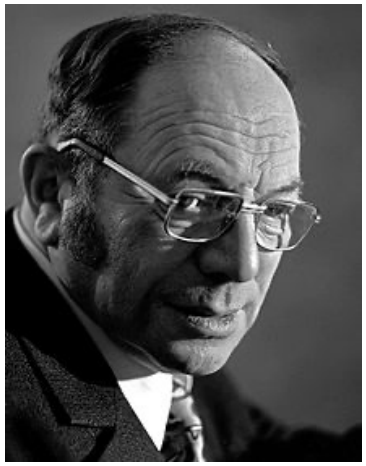
$$T_{\#}\rho_A = \rho_B$$

Monge 问题 (1781)

Wasseratein-2度量

- 最优传输问题





Kantorovich 问题 (1942)

$$W_2^2(\rho_A, \rho_B) = \inf_{\gamma} \iint \gamma(\theta_1, \theta_2) \|\theta_1 - \theta_2\|_2^2 d\theta_1 d\theta_2$$

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

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

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

Wasserstein-2度量

动力学观点 (Benamou, Brenier, 2000)

$$W_2^2(\rho_A, \rho_B) = \inf_{v_t} \int_0^1 \int \| v_t(\theta) \|_2^2 \, \rho_t(\theta) d\theta dt$$

$$\frac{\partial \rho_t}{\partial t} + \nabla_\theta \cdot (\rho_t v_t) = 0$$

$$\rho_0(\theta) = \rho_A(\theta) \qquad \rho_1(\theta) = \rho_B(\theta)$$

ρ_t为连接ρ_A和ρ_B的匀速测地线。

Wasserstein-2度量

黎曼几何表述 (Otto calculus, 2001)

定义Wassserstein-2黎曼度量：

$$g_\rho^{W_2}(\sigma,\sigma)=\inf_v\{\int \parallel v \parallel_2^2 \rho d\theta\ ,\ \sigma=-\nabla_\theta\cdot(\rho v)\}$$

使用拉格朗日乘子法，我们可以进一步得到：

$$v=\nabla_\theta\psi \qquad \sigma=-\nabla_\theta\cdot(\rho\nabla_\theta\psi)$$

度量可以简化为：

$$g_\rho^{W_2}(\sigma,\sigma)=\int \parallel \nabla_\theta\psi \parallel_2^2 \rho d\theta\ ,\ \sigma=-\nabla_\theta\cdot(\rho\nabla_\theta\psi)$$

Wasserstein-2 距离可以表述为：

$$W_2^2(\rho_A,\rho_B)=\inf_{\psi_t}\int_0^1 g_{\rho_t}(\sigma_t,\sigma_t)dt$$

σ_t 和ρ_t（连接ρ_A和ρ_B）满足：

$$\frac{\partial\rho_t}{\partial t}+\nabla\cdot(\rho_t\nabla_\theta\psi_t)=0 \qquad \sigma_t=-\nabla\cdot(\rho_t\nabla_\theta\psi_t)$$

Wasserstein-2度量

黎曼几何表述 (Otto calculus, 2001)

给定Wassserstein-2黎曼度量：

$$g_\rho^{W_2}(\sigma,\sigma)=\int \|\nabla_\theta\psi\|_2^2\,\rho d\theta\,,\qquad \sigma=-\nabla_\theta\cdot(\rho\nabla_\theta\psi)$$

Wassserstein-2度量矩阵：

$$M^{W_2}(\rho)\sigma=\psi$$

其中σ和ψ满足：

$$-\nabla_\theta\cdot(\rho\nabla_\theta\psi)=\sigma \qquad g_\rho^{W_2}(\sigma,\sigma)=\int\psi\sigma d\theta$$

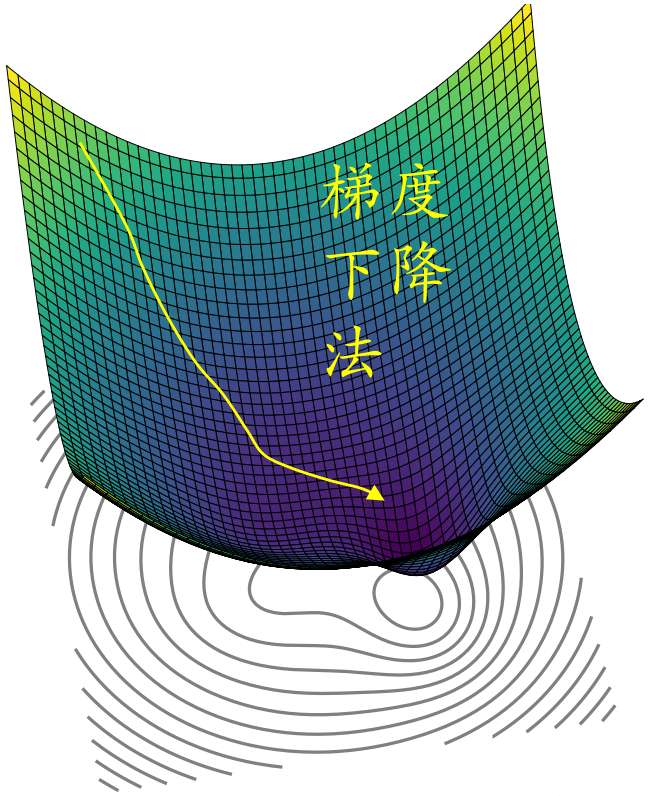
即

$$M^{W_2}(\rho)^{-1}\psi=\sigma=-\nabla_\theta\cdot(\rho\nabla_\theta\psi)$$

Wasserstein变分推理

- Wasserstein梯度流

$$\frac{\partial \rho_t}{\partial t} = -\nabla^{W_2}\mathcal{E}(\rho_t)$$
$$= -M^{W_2}(\rho_t)^{-1}\frac{\delta \mathcal{E}}{\delta \rho}(\rho_t)$$
$$= \nabla_\theta \cdot \left(\rho_t \nabla_\theta \frac{\delta \mathcal{E}}{\delta \rho}(\rho_t)\right)$$
$$= \nabla_\theta \cdot (\rho_t \nabla_\theta (\log \rho_t - \log \rho^*))$$



Wasserstein梯度流是高维偏微分方程！难以求解！

- 设计新的算法
- 收敛性分析（研究像Langevin 动力系统这样的Monte Carlo 方法的收敛性）

24

Langevin 动力学采样

Langevin 动力学

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

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

还有其它动力系统，对应的Fokker Planck 方程是这样吗？

Metropolis-Hastings Langevin 算法

- Metropolis-Hastings 算法

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

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（MALA）

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

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

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

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

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

高斯变分推理

高斯近似的Wasserstein 梯度流

我们对ρ_t进行高斯近似$\rho_t \approx \rho_{a_t} = \mathcal{N}(m_t, C_t)$

$$\frac{dm_t}{dt} = \int \nabla_\theta \cdot \left[\rho_{a_t} \nabla_\theta \Phi_R + \nabla_\theta \rho_{a_t}\right] \theta \, d\theta = -\mathbb{E}_{\rho_{a_t}}[\nabla_\theta \Phi_R]$$

$$\frac{dC_t}{dt} = \int \nabla_\theta \cdot \left[\rho_{a_t} \nabla_\theta \Phi_R + \nabla_\theta \rho_{a_t}\right] (\theta - m_t)(\theta - m_t)^T \, d\theta$$

$$= 2I - \mathbb{E}_{\rho_{a_t}}[\nabla_\theta \nabla_\theta \Phi_R] C_t - C_t \mathbb{E}_{\rho_{a_t}}[\nabla_\theta \nabla_\theta \Phi_R]$$

高斯积分

收敛性分析

Log-Sobolev 不等式

给定ρ^*, 存在$\alpha > 0$， 对于任何ρ

$$\int \rho \left\| \nabla_\theta \log \frac{\rho}{\rho^*} \right\|_2^2 d\theta \geq 2\alpha \int \rho \log \frac{\rho}{\rho^*} d\theta$$

那么沿着Wasserstein梯度流

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

我们有指数收敛

$$\mathrm{KL}[\rho_t \parallel \rho^*] \leq e^{-2\alpha t} \mathrm{KL}[\rho_0 \parallel \rho^*]$$

收敛性分析

Bakry-Émery 定理

如果 ρ^* 是对数凹(log-concave)的密度函数，

$$-\nabla_\theta \nabla_\theta \log \rho^* \geq \alpha I$$

那么 ρ^* 满足Log Sobolev 不等式

$$\exists \alpha > 0, \int \rho \left\| \nabla_\theta \log \frac{\rho}{\rho^*} \right\|_2^2 d\theta \geq 2\alpha \int \rho \log \frac{\rho}{\rho^*} d\theta, \forall \rho$$

收敛性分析

- 凸函数

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

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

- 对数凹密度函数

$$-\nabla_\theta \nabla_\theta \log \rho^* \geq \alpha I \quad (\nabla_\theta \nabla_\theta \Phi_R \geq \alpha I)$$

KL散度在Wasserstein 度量下是强凸的。即从ρ_0出发方向为σ_0到ρ_1的测地线满足：

$$\mathrm{KL}[\rho_1 \| \rho^*] \geq \mathrm{KL}[\rho_0 \| \rho^*] + g_{\rho_0}^{W_2}(\nabla^{W_2} \mathrm{KL}[\rho_0 \| \rho^*], \sigma_0) + \frac{\alpha}{2} W_2^2(\rho_0, \rho_1)$$

Wasserstein变分推理

Monge 问题 (1781)
Kantorovich 问题 (1942)
动力学观点 (Benamou, Brenier, 2000)
Wasserstein-2 度量 (Otto Calculus, 2001)
Wasserstein 梯度流

- 设计变分推理算法
- 收敛性分析（研究像Langevin 动力系统这样的 Monte Carlo 方法的收敛性）

31

Fisher-Rao度量

黎曼几何表述 (Rao, 1945)

给定Fisher-Rao黎曼度量：

$$g_\rho^{FR}(\sigma,\sigma)=\int\frac{\|\sigma\|_2^2}{\rho}d\theta$$

Fisher-Rao度量矩阵：

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

其中σ和ψ满足：

$$g_\rho^{FR}(\sigma,\sigma)=\int\psi\sigma d\theta=\int\|\psi\|_2^2\,\rho d\theta$$

即

$$M^{FR}(\rho)^{-1}\psi=\sigma:=\rho\psi-\mathbb{E}_\rho[\psi]$$

Fisher-Rao度量

微分同胚不变性 (diffeomorphism invariance)

给定任意微分同胚：

$$T: R^d \to R^d$$

Fisher-Rao 度量

$$g_\rho^{FR}(\sigma,\sigma) = \int \frac{\| \sigma \|_2^2}{\rho} d\theta$$

是唯一(相差一个常数)的度量满足

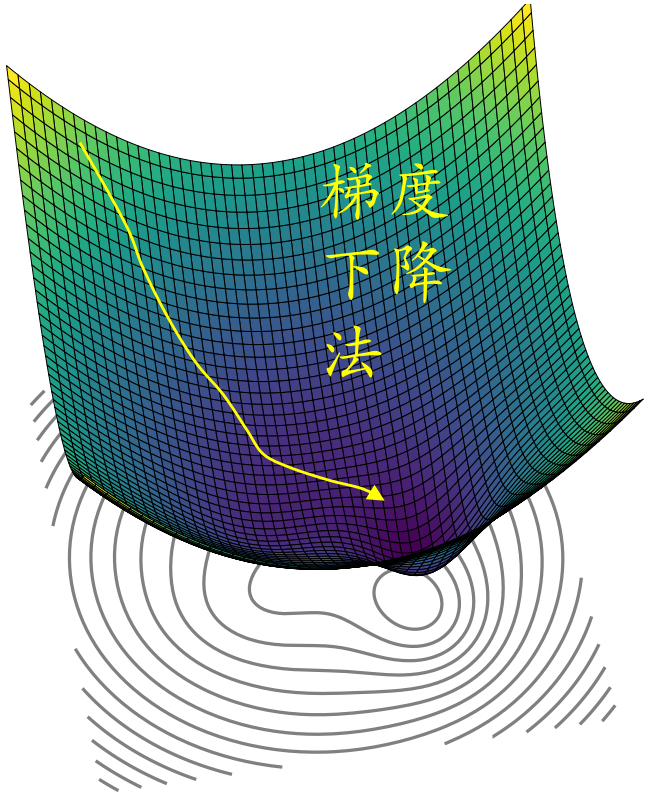
$$g_{T_\# \rho}^{FR}(T_\# \sigma, T_\# \sigma) = g_\rho^{FR}(\sigma,\sigma)$$

Martin Bauer, Martins Bruveris, and Peter W Michor. Uniqueness of the Fisher – Rao metric on the space of smooth densities. Bulletin of the London Mathematical Society, 48(3):499 – 506, 2016.

Fisher-Rao变分推理

- Fisher-Rao梯度流

$$\frac{\partial \rho_t}{\partial t} = -\nabla^{FR}\mathcal{E}(\rho_t)$$
$$= -M^{FR}(\rho_t)^{-1}\frac{\delta \mathcal{E}}{\delta \rho}(\rho_t)$$
$$= \rho_t \frac{\delta \mathcal{E}}{\delta \rho}(\rho_t) - \mathbb{E}_{\rho_t}[\frac{\delta \mathcal{E}}{\delta \rho}(\rho_t)]$$
$$= \rho_t(\log \rho^* - \log \rho_t)$$
$$-\rho_t \mathbb{E}_{\rho_t}[\log \rho^* - \log \rho_t]$$



- 收敛性分析
- 设计新的算法

34

收敛性分析

对偶梯度支配 (dual gradient dominance, Carrillo 等, 2024)

沿着Fisher-Rao梯度流

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

我们有

$$-\frac{d}{dt}(\mathrm{KL}[\rho_t \| \rho^*] + \mathrm{KL}[\rho^* \| \rho_t]) \geq \mathrm{KL}[\rho_t \| \rho^*] + \mathrm{KL}[\rho^* \| \rho_t]$$

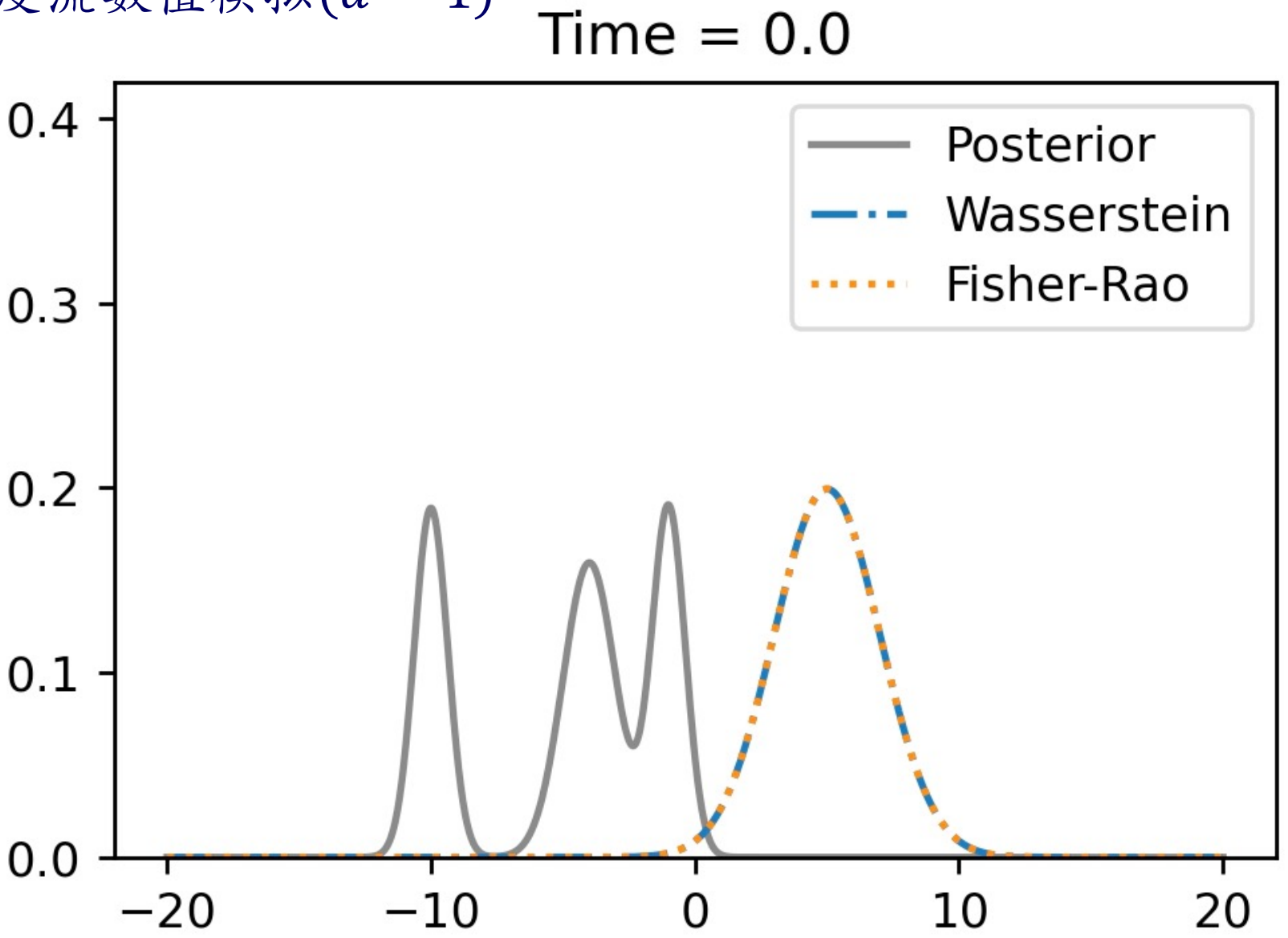
我们有指数收敛

$$\mathrm{KL}[\rho_t \| \rho^*] + \mathrm{KL}[\rho^* \| \rho_t] \leq e^{-t}(\mathrm{KL}[\rho_0 \| \rho^*] + \mathrm{KL}[\rho^* \| \rho_0])$$

对ρ^*没有要求，但需要 $\mathrm{KL}[\rho_0 \| \rho^*] + \mathrm{KL}[\rho^* \| \rho_0] < \infty$

收敛性分析

- 梯度流数值模拟($d = 1$)



生灭过程(birth-death process)

生灭过程

定义

$$\alpha_t(\theta) = \left(-\Phi_R(\theta) - \log \rho_t(\theta)\right) - \mathbb{E}_{\rho_t}[-\Phi_R - \log \rho_t]$$

考虑生灭过程：

假设$\{\theta_0^j\} \sim \rho_0$ 如果$\alpha_t(\theta_t^j) > 0$ (或者$\alpha_t(\theta_t^j) < 0$), θ_t^j 以 $1 - e^{-|\alpha_t|dt}$ 的概率 复制(或者湮灭)。

那么$\{\theta_t^j\} \sim \rho_t$， ρ_t满足

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

$\{\theta_t^j\}$ 位置不能改变， $\mathrm{KL}[\rho^* \| \rho_0] = \infty$，不能收敛？

高斯变分推理

自然梯度下降法

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

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

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

高斯变分推理

高斯变分推理

$$\min_a \mathrm{KL}[\rho_a \,\|\, \rho^*]$$

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

极小值点满足

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

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

当Φ_R是强凸函数时，极小值点唯一。

高斯变分推理

梯度下降方法

导数满足

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

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

梯度下降方法

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

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

高斯变分推理

- 高斯近似的Wasserstein变分推理：

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

- 梯度下降方法：

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

- 自然梯度下降方法：

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

自然梯度下降法

- 参数密度空间的Fisher-Rao度量

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

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

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

$$\mathfrak{g}_a(v_1, v_2) = g^{FR}_{\rho_a}(\nabla_a \rho_a \cdot v_1, \nabla_a \rho_a \cdot v_2) = v_1^T \mathfrak{M}(a) v_2$$

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？

自然梯度下降法

- Fisher 信息矩阵 (Fisher information matrix)

$$\mathfrak{M}(a) = \mathrm{FIM}(a) := \int \frac{\nabla_a \rho_a \cdot \nabla_a \rho_a}{\rho_a} d\theta$$

$$\mathrm{KL}[\rho_{a+da} \| \rho_a] \approx \frac{1}{2} da^T \mathrm{FIM}(a) da$$

- 自然梯度下降法 (natural gradient descent)

$$\frac{\partial a_t}{\partial t} = -\mathrm{FIM}(a_t)^{-1} \nabla_a \mathrm{KL}[\rho_{a_t} \| \rho^*]$$

$$v = \mathrm{argmin}_v \frac{\nabla_a \mathrm{KL}[\rho_{a_t} \| \rho^*] \cdot v}{\sqrt{v^T \mathfrak{M}(a) v}} = \mathrm{FIM}(a_t)^{-1} \nabla_a \mathrm{KL}[\rho_{a_t} \| \rho^*]$$

自然梯度下降法

- 指数分布族

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

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

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

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

自然梯度下降法

- 练习

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

期望：

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

Fisher信息矩阵：

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

自然梯度下降法

- 练习

计算Gaussian 分布

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

的Fisher信息矩阵。

高斯变分推理

- 练习

高斯近似的Wasserstein变分推理：

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

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

梯度下降方法：

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

自然梯度下降法：

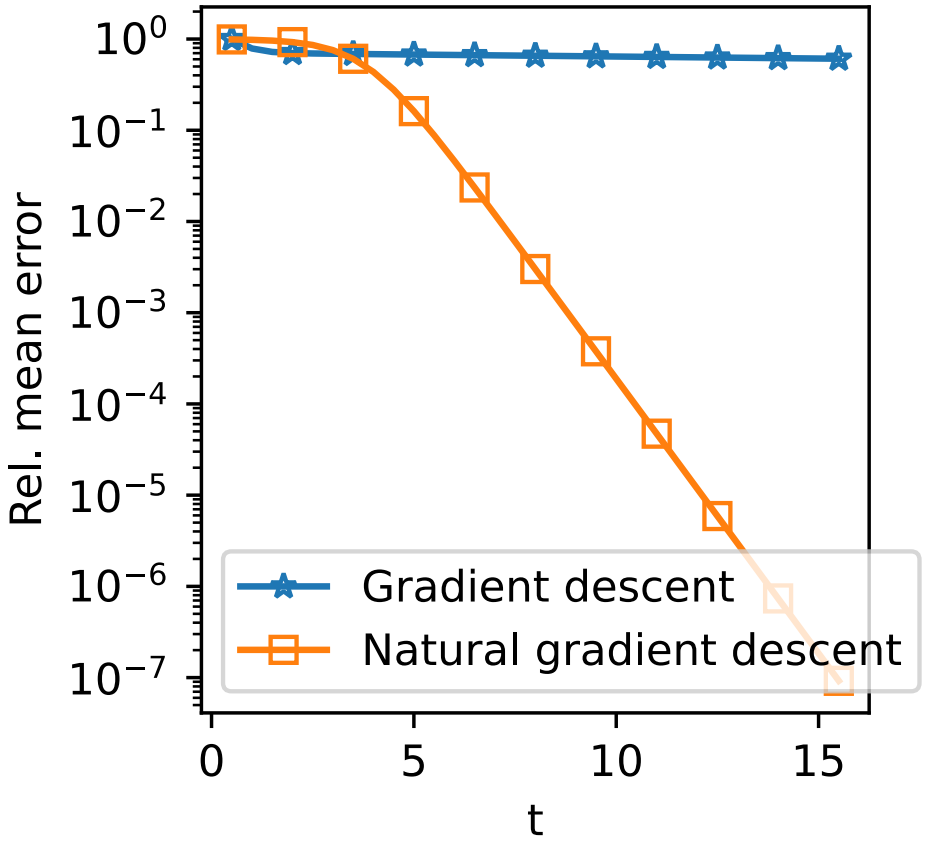
$$\frac{dm_t}{dt} = -C_t\mathbb{E}_{\rho_{a_t}}[\nabla_\theta \Phi_R] \qquad \frac{dC_t}{dt} = C_t - C_t\mathbb{E}_{\rho_{a_t}}[\nabla_\theta \nabla_\theta \Phi_R]C_t$$

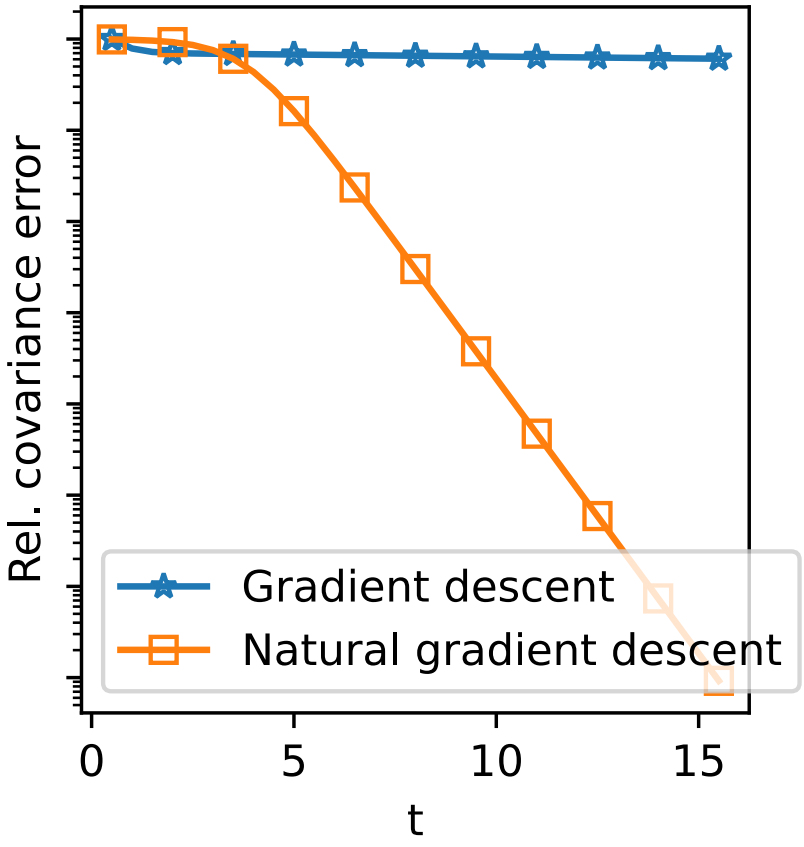
$$\frac{dC_t^{-1}}{dt} = -C_t^{-1} + \mathbb{E}_{\rho_{a_t}}[\nabla_\theta \nabla_\theta \Phi_R]$$

高斯变分推理

- 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$m_0 = [0;0]$，$C_0 = \mathrm{diag}\{1, 1\}$。





仿射不变性

仿射不变性

给定梯度流

$$\frac{\partial \rho_t}{\partial t} = F(\rho_t, \rho^*)$$

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

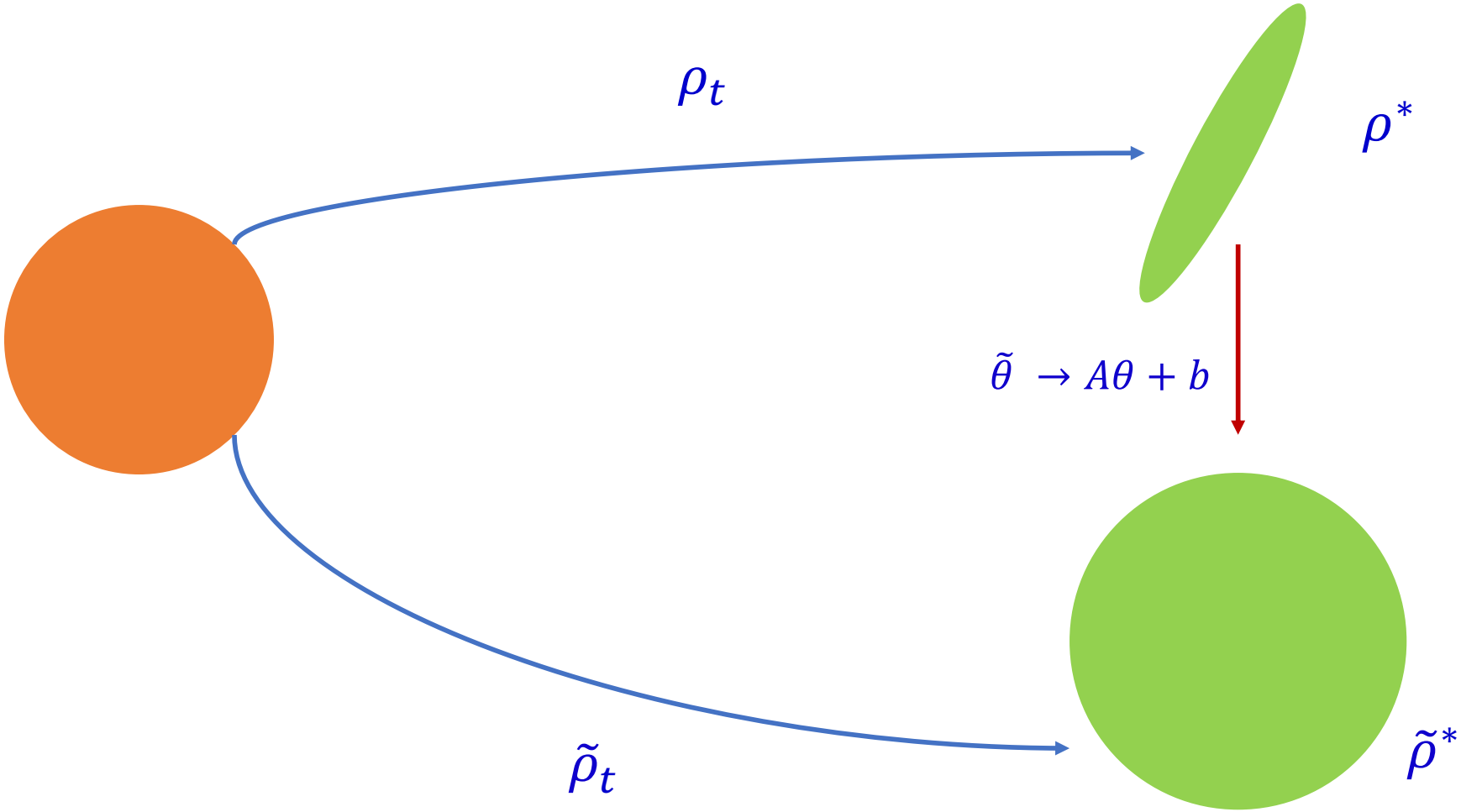
$$\frac{\partial \tilde{\rho}_t}{\partial t} = F(\tilde{\rho}_t, \tilde{\rho}^*)$$

其中

$$\tilde{\rho}_t = T_{\#}\tilde{\rho}_t \qquad \tilde{\rho}^* = T_{\#}\tilde{\rho}^*$$

仿射不变性

- 仿射不变性


ρ_t
ρ^*
$\tilde{\theta} \rightarrow A\theta + b$
$\tilde{\rho}_t$
$\tilde{\rho}^*$

仿射不变性

- 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

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

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

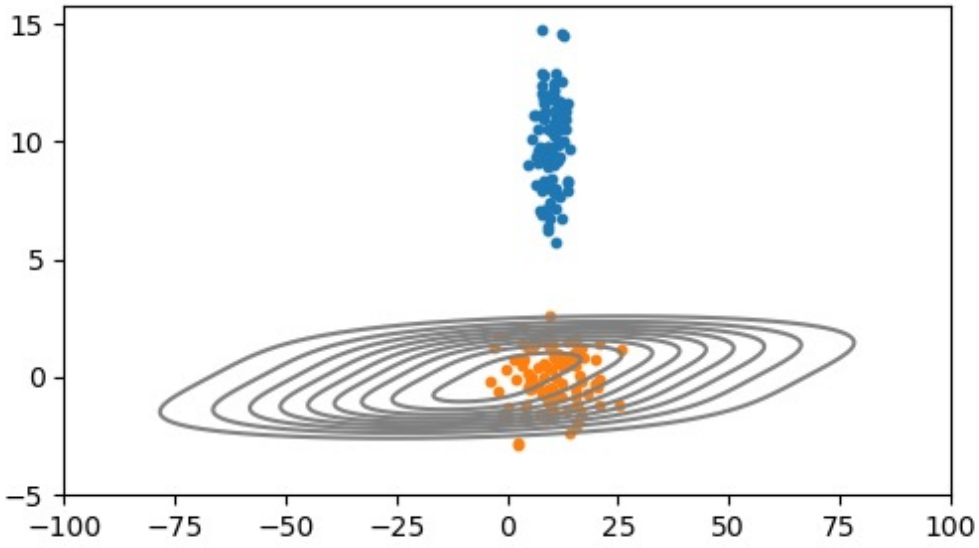
仿射不变性

- 练习

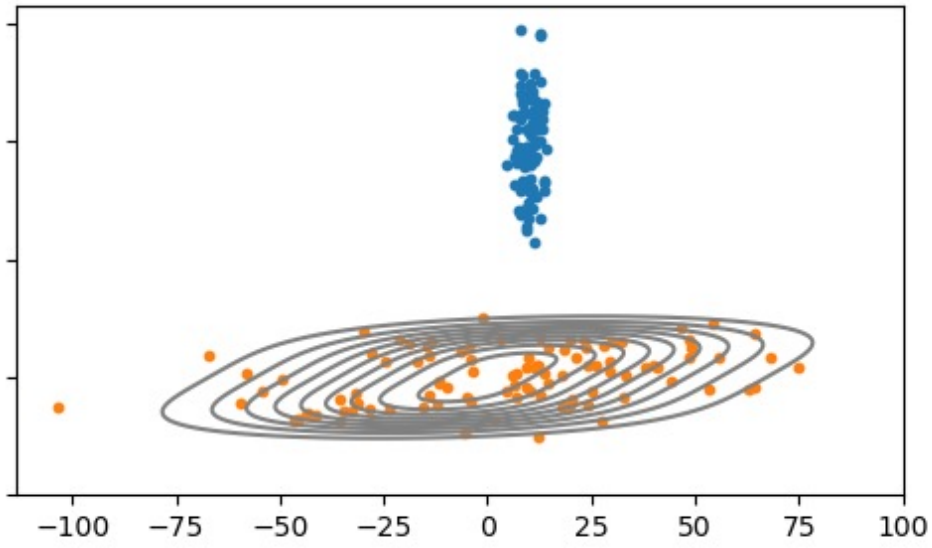
$$\rho^*(\theta) = \frac{1}{Z} e^{-\Phi_R(\theta)}$$

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

Langevin 动力学



预处理 Langevin 动力学



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.

贝叶斯反问题

	计算速度	近似方法	导数计算
重要性采样	慢	粒子逼近	不需要
卡尔曼方法	快	高斯近似	一般不需要
正规化流	求解优化问题	神经网络逼近	不需要
基于随机游走、重要性采样的马氏链蒙特卡洛	慢	遍历性	不需要
基于Langevin的马氏链蒙特卡洛	中	遍历性	需要
高斯变分推理（自然梯度下降方法）	中	高斯近似	需要（甚至二阶导数）
非参数化的变分推理	中	粒子逼近	一般需要