

数学基础

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

- 概率测度空间
- 概率测度空间的度量
- 随机过程



概率测度空间

➤ 空间

- 点的集合，比如欧式空间 R^2

➤ 度量空间 X

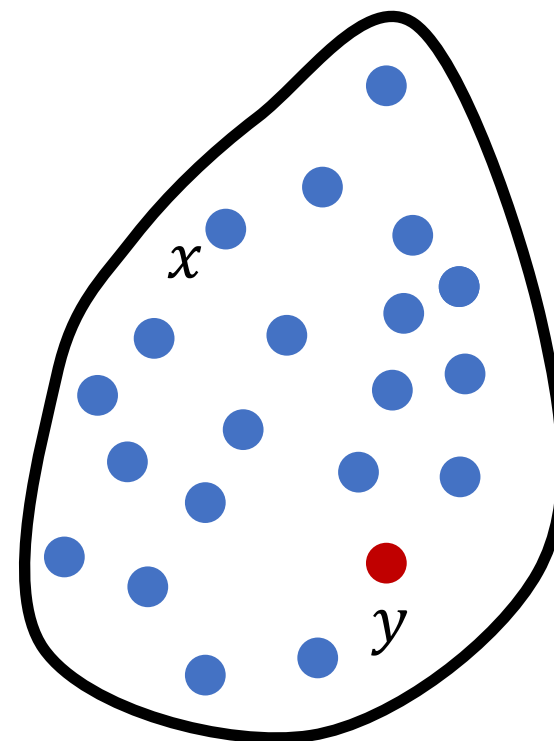
- 定义任何两个元素之间有距离

非负性: $d(x, y) \geq 0$

且 $d(x, y) = 0$ 等价于 $x = y$

对称性: $d(x, y) = d(y, x)$

三角不等式: $d(x, y) \leq d(x, z) + d(z, y)$



度量空间 X

⇒

拓扑空间: 开集、闭集等

连续映射: $f: X_1 \rightarrow X_2$ 连续 \Leftrightarrow 任意 X_2 中的开集 U , $f^{-1}(U)$ 是 X_1 中的开集



概率测度空间

➤ 概率测度空间

$$\mu(A) = P(x \in A) \quad A \subset \mathbb{R}^d$$

离散: $\mu = \sum_i p_i \delta_{x_i}$

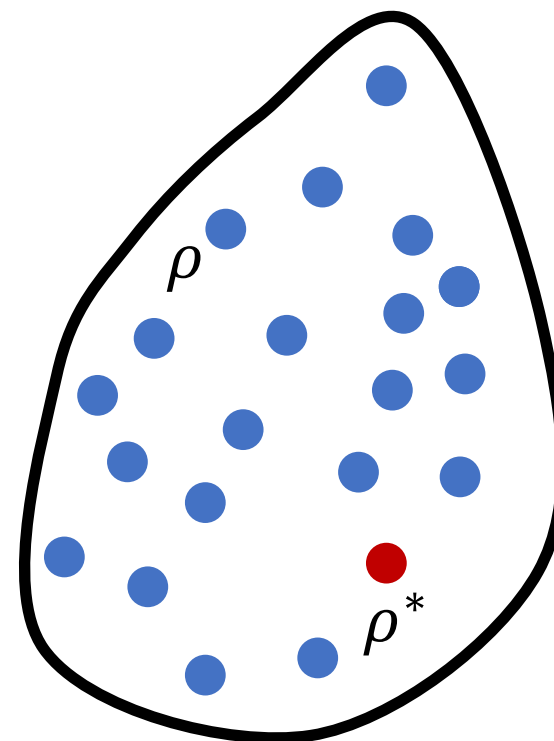
连续有密度: $\mu(dx) = \rho(x)dx$

密度函数的集合: $\{\rho: \mathbb{R}^d \rightarrow \mathbb{R}\}$

- $\rho(x) \geq 0$

- $\int \rho(x)dx = 1$

- $\rho \in C^\infty(\mathbb{R}^d)$



概率测度空间 \mathcal{P}

注: 除非特殊说明, 本课程假设概率测度的密度函数存在



概率测度空间

➤ 参数化的概率测度空间

高斯密度空间 $\{(m, C), C > 0\}$

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

简化的高斯密度空间 $\{(m, \delta), \delta > 0\}$

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

混合高斯密度空间 $\{(w_k, m_k, C_k)_{k=1}^K, C_k > 0, w_k \geq 0\}$

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

.....



概率测度空间

➤ 推前算子 (pushforward operator)

可测函数:

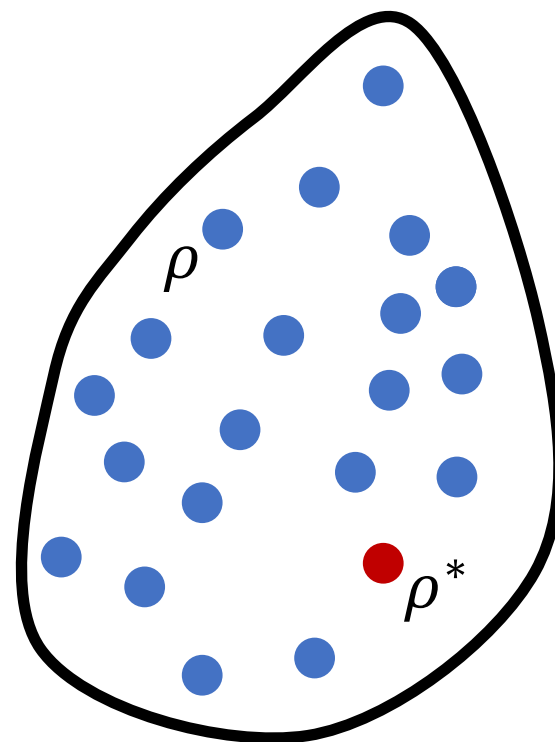
$$T: R^d \rightarrow R^{d'}$$

推前算子:

$$T_{\#}: \mathcal{P} \rightarrow \mathcal{P}'$$

$T_{\#}\rho$ 是一个概率密度函数:

$$\int_{x' \in A} T_{\#}\rho(x') dx' := \int_{Tx \in A} \rho(x) dx$$



概率测度空间 \mathcal{P}



概率测度空间的度量

➤ 距离函数

- 全变差距离 (total variation)

$$d_{TV}(\rho, \rho') = \frac{1}{2} \int |\rho - \rho'| d\theta = \frac{1}{2} \|\rho - \rho'\|_{L_1}$$

- 海林格 (Hellinger) 距离

$$d_H(\rho, \rho') = \left(\frac{1}{2} \int |\sqrt{\rho} - \sqrt{\rho'}|^2 d\theta \right)^{\frac{1}{2}} = \frac{1}{\sqrt{2}} \|\sqrt{\rho} - \sqrt{\rho'}\|_{L_2}$$

➤ 练习

验证 d_{TV} 和 d_H 是距离



概率测度空间的度量

➤ 练习

- 良定性

$$0 \leq d_{TV}(\rho, \rho') \leq 1 \quad 0 \leq d_H(\rho, \rho') \leq 1$$

- 等价性

$$\frac{1}{\sqrt{2}} d_{TV}(\rho, \rho') \leq d_H(\rho, \rho') \leq \sqrt{d_{TV}(\rho, \rho')}$$

- 其它估计

$$d_{TV}(\rho, \rho') = \frac{1}{2} \sup_{|f|_\infty \leq 1} |\mathbb{E}_\rho[f] - \mathbb{E}_{\rho'}[f]|$$

$$|\mathbb{E}_\rho[f] - \mathbb{E}_{\rho'}[f]| \leq 2 \|f\|_\infty d_{TV}(\rho, \rho')$$

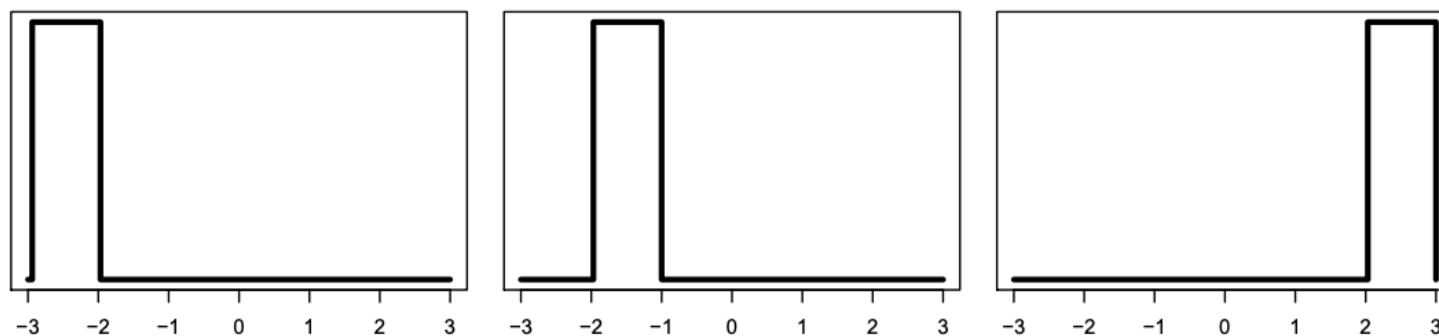
$$|\mathbb{E}_\rho[f] - \mathbb{E}_{\rho'}[f]| \leq 2 (\mathbb{E}_\rho[f^2] + \mathbb{E}_{\rho'}[f^2])^{\frac{1}{2}} d_H(\rho, \rho')$$



概率测度空间的度量

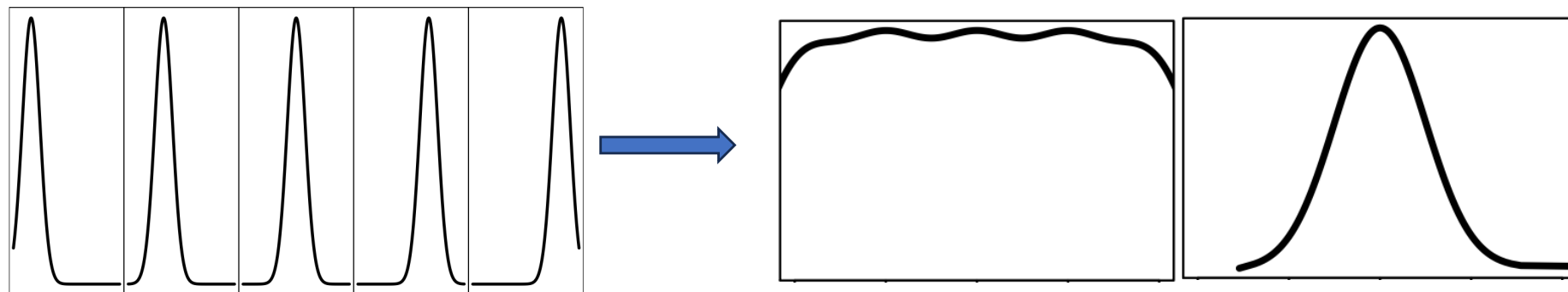
➤ 这些距离函数的问题

- 并没有考虑概率测度空间的几何性质



$$d(\rho_1, \rho_2) = d(\rho_1, \rho_3)$$

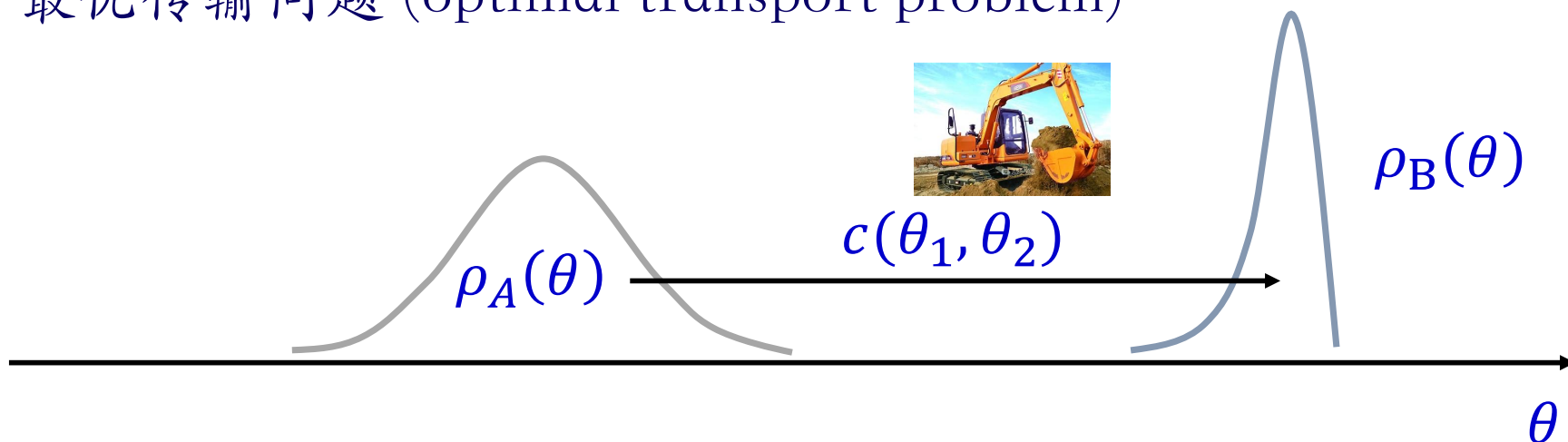
- 当我们对不同的对象 (如概率分布或图像) 求平均 (重心), 希望得到的结果仍然是一个相似的对象





最优传输问题

➤ 最优传输问题 (optimal transport problem)



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

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

对于一般测度,
解不一定存在!

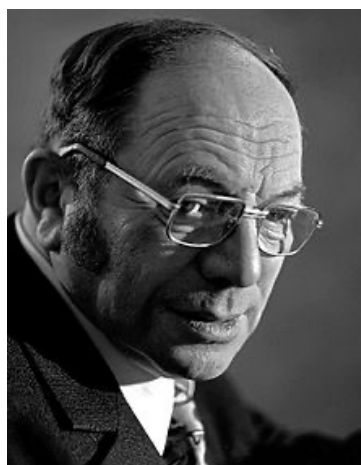
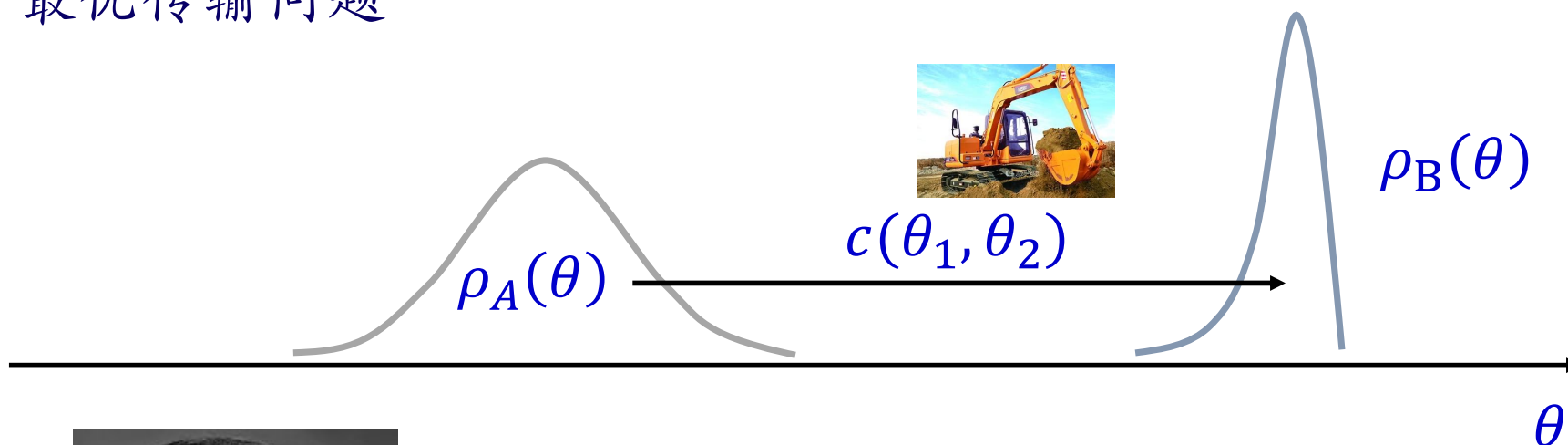
Monge 问题 (1781)

T^* : 最优传输映射 (optimal transport map)



最优传输问题

➤ 最优传输问题



Kantorovich 问题 (1942)
 γ^* : 最优耦合 (optimal coupling,
optimal transport plan)

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

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

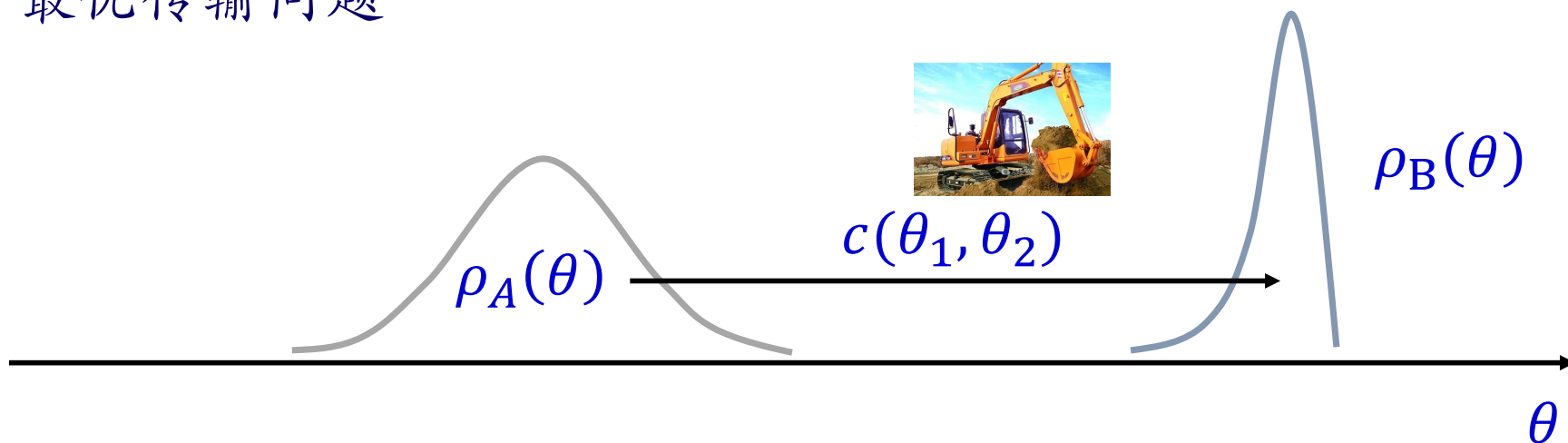
$$\int \gamma(\theta_1, \theta_2) d\theta_1 = \rho_B(\theta_2)$$

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



最优传输问题

➤ 最优传输问题



Kantorovich 对偶问题

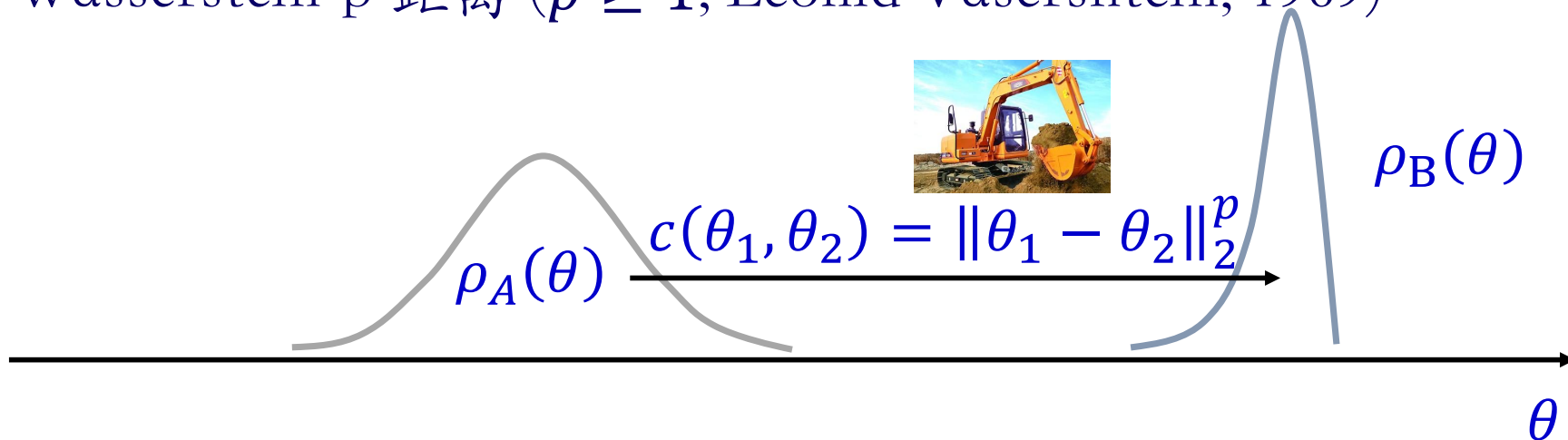
$$\sup_{f, g} \int \rho_A(\theta) f(\theta) + \rho_B(\theta) g(\theta) d\theta$$

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



概率测度空间的度量

➤ Wasserstein-p 距离 ($p \geq 1$, Leonid Vasershtein, 1969)



$$W_p(\rho_A, \rho_B) = \inf_{\gamma} \left(\iint \gamma(\theta_1, \theta_2) \|\theta_1 - \theta_2\|_2^p d\theta_1 d\theta_2 \right)^{1/p}$$

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

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



概率测度空间的度量

- Wasserstein-1距离 (Earth Mover distance)

$$c(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|_2$$

成本函数是欧几里得距离，更贴近物理运输成本

- Wasserstein-2距离

$$c(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|_2^2$$

成本函数是平方距离，导致它的优化目标更倾向于短距离的调整，更适用于平滑变形。多用于描述流体动力学、几何优化



Wasserstein-1 距离

Kantorovich 对偶问题

一般的最优输运问题可以转化为

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

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

当 $c(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|_2$, 我们有

$$W_1(\rho_A, \rho_B) = \sup_h \int \rho_A(\theta) h(\theta) - \rho_B(\theta) h(\theta) d\theta$$

$$|h(\theta_1) - h(\theta_2)| \leq \|\theta_1 - \theta_2\|_2$$



Wasserstein-2 距离

Brenier 定理

ρ_A 和 ρ_B 是两个概率密度函数，对于 Kantorovich 问题，

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

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

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

那么存在一个凸函数 $\varphi: R^d \rightarrow R$ ，使得 $(I, \nabla\varphi)_\# \rho_A = \gamma^*$ 。
且如果存在凸函数 $\psi: R^d \rightarrow R$ ，使得 $\nabla\psi_\# \rho_A = \rho_B$ ，那么几乎处处 $\nabla\varphi = \nabla\psi = T$ (Monge 问题)。

注：这里 ρ_B 可以换成概率测度 μ_B



Wasserstein-2 距离

测地线

对于 Kantorovich 问题,

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

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

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

定义

$$T_t(\theta_1, \theta_2) = (1-t)\theta_1 + t\theta_2$$

那么

$$\rho_t = T_{t\#}\gamma^*$$

是连接 ρ_A 和 ρ_B 的匀速测地线 (constant-speed geodesic)。



Wasserstein-2 距离

动力学观点 (Dynamic formulation, Benamou, Brenier, 1999)

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

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

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

ρ_t 为连接 ρ_A 和 ρ_B 的匀速测地线。



概率测度空间的度量

➤ 测地线



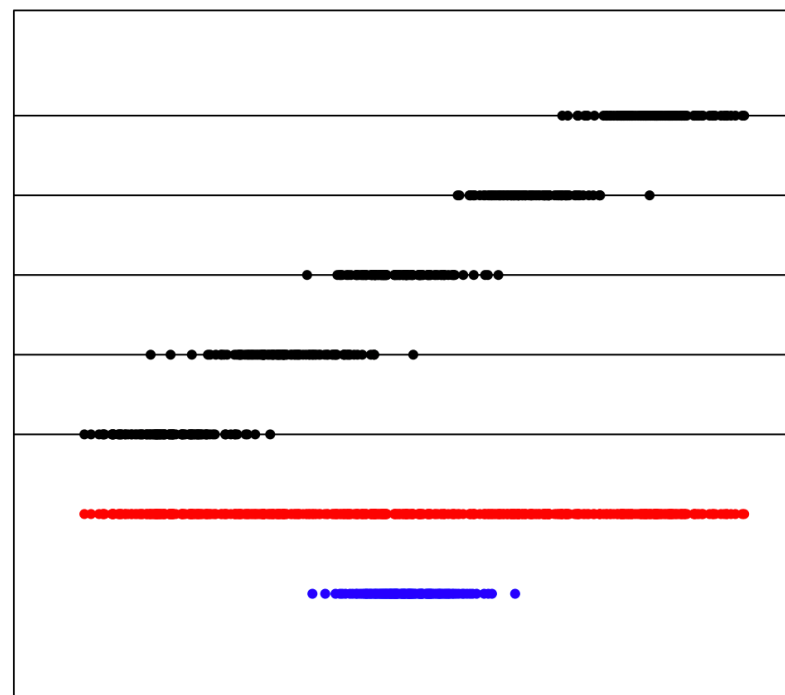
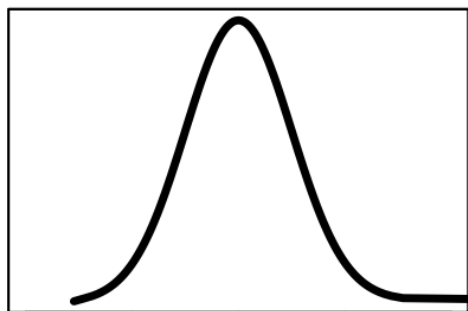
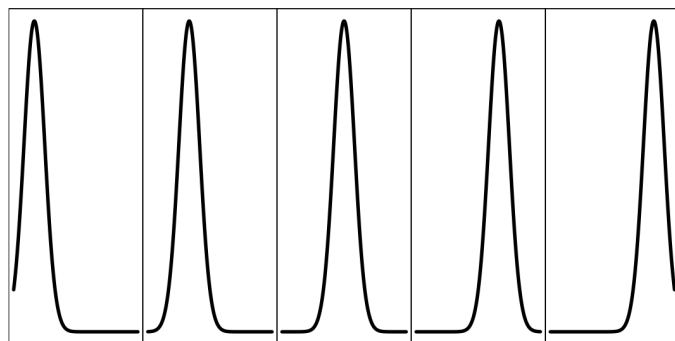
在数据处理中，用于理解和预测数据插值和演变



概率测度空间的度量

➤ Wasserstein 重心 (barycenter)

$$\rho^* = \operatorname{argmin}_{\rho} \sum_{j=1}^m W_p^p(\rho, \rho_j)$$



用于图像平均、分布对齐等



概率测度空间的“伪度量”

➤ 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：

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

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

因此

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



概率测度空间的“伪度量”

➤ KL-散度

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

$$\text{KL}[\rho \parallel Z\rho^*] = \text{KL}[\rho \parallel \rho^*] - \log(Z)$$

➤ 反向 KL-散度

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

➤ χ^2 -距离

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



概率测度空间的“伪度量”

- 最大均值差异(maximum mean discrepancy, MMD)

给点任意函数类 F

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

离散情况:

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

$$F = \{f: |f|_{\infty} \leq 1\}, \quad F = \text{span}\{x, x^2\} \dots$$

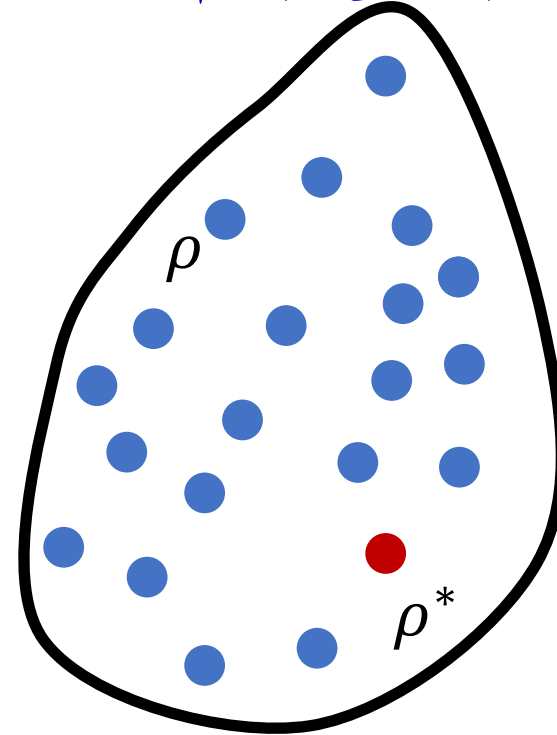


概率测度空间

➤ 概率测度空间

- 密度函数的集合： $\{\rho: R^d \rightarrow R\}$
- $\rho(x) \geq 0$
- $\int \rho(x) dx = 1$
- $\rho \in C^\infty(R^d)$

概率测度空间 \mathcal{P}



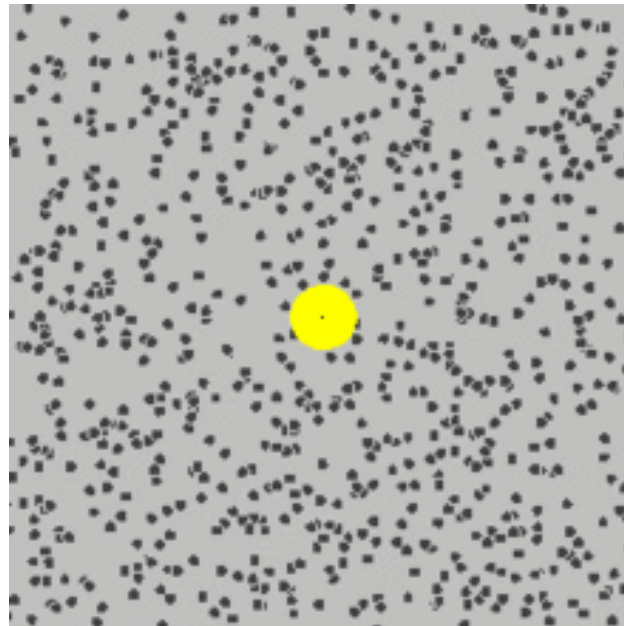
➤ 概率测度空间的度量、伪度量

- 概率分布的比较提供严谨的数学框架（比如收敛性）
- 提供了在高维分布之间进行优化的可能性



随机过程

➤ 布朗运动



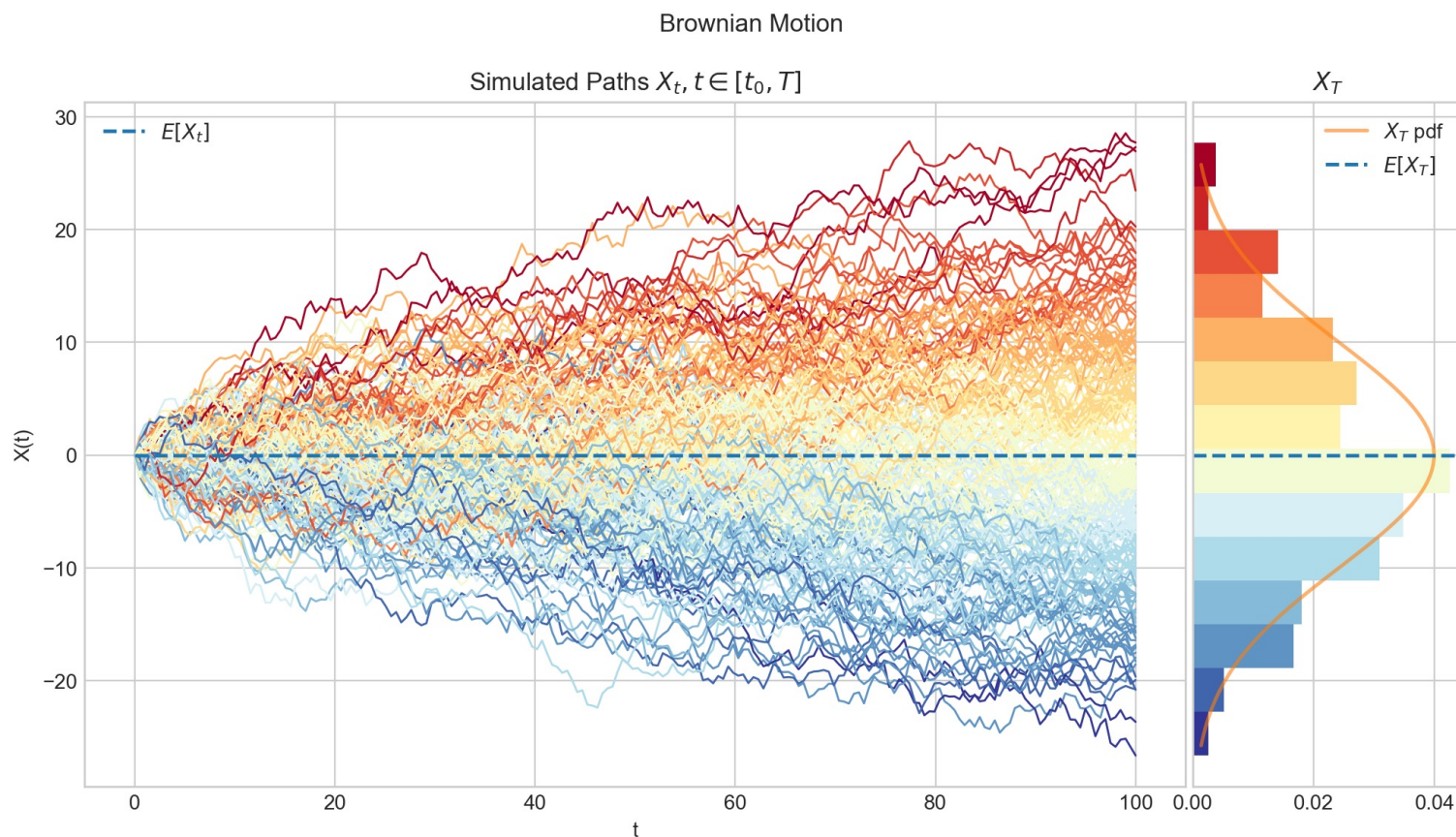
Robert Brown 于 1827 年在显微镜下观察浸泡在水中的植物花粉，首次描述了这一现象。



随机过程

➤ 随机过程

- 随机过程是描述系统在时间上随机变化的数学模型
- 如果系统状态服从一个概率测度，概率测度在概率密度函数空间的演化

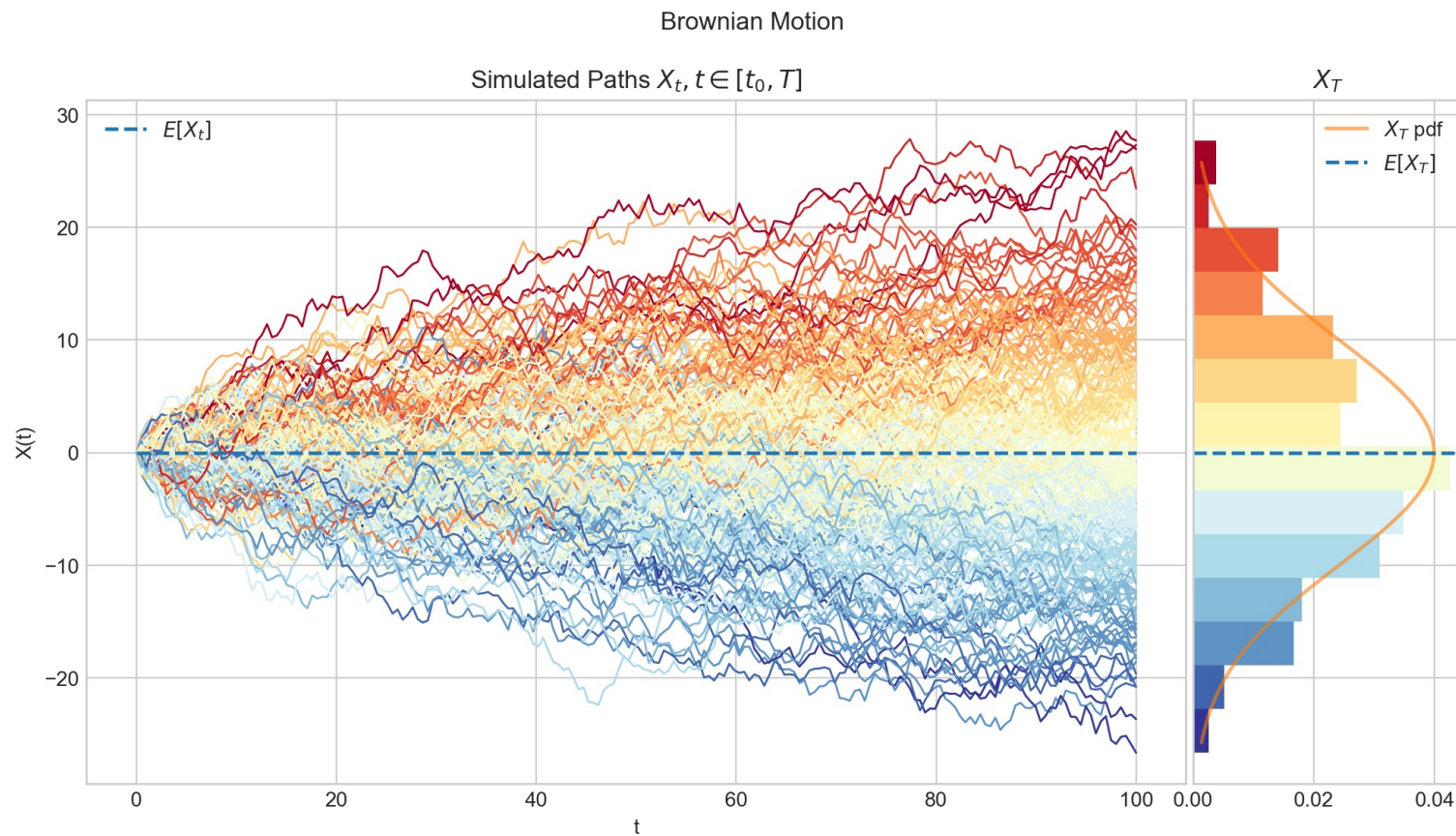




随机过程

➤ 随机过程、高斯过程

- 欧式空间的概率密度 $\rho: R^d \rightarrow R$
- 函数空间的概率密度 $\rho: C[0,1] \rightarrow R$





随机过程

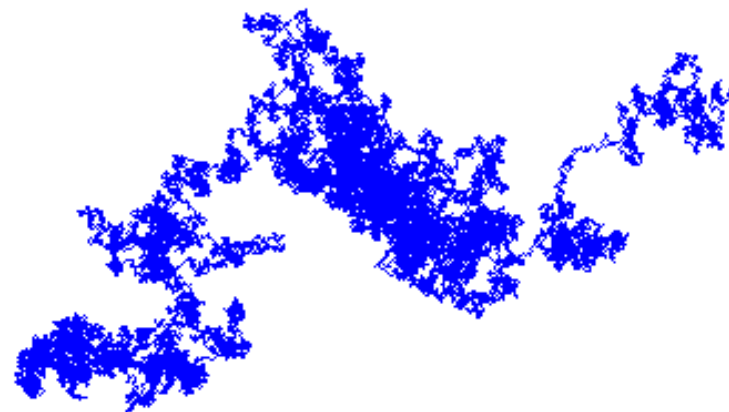
➤ 布朗运动 $(B_t)_{t \geq 0} \in R^d$

1. $B_0 = 0$

2. 独立增量：对于任意 $0 \leq t_1 < \dots < t_k$ ，随机变量 $(B_{t_1}, B_{t_2} - B_{t_1}, \dots, B_{t_k} - B_{t_{k-1}})$ 相互独立

3. 稳定增量和正态性：对于任意 $0 \leq s < t < \infty$ ，
$$B_t - B_s \sim \mathcal{N}(0, t - s)$$

4. $(B_t)_{t \geq 0}$ 关于 t 几乎处处连续





随机过程

➤ 随机微分方程 $(\theta_t)_{t \geq 0} \in \mathbb{R}^d$

$$d\theta_t = b(\theta_t, t)dt + \sigma(\theta_t, t)dB_t \quad \theta_0 = \theta(0)$$

其中 $b : \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R}^d$, $\sigma : \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R}^{d \times d}$, dB_t 满足

$$dB_t \sim \mathcal{N}(0, dt)$$

数值计算: $\theta_{t+\Delta t} - \theta_t \approx b(\theta_t, t)\Delta t + \sigma(\theta_t, t)(B_{t+\Delta t} - B_t)$
 $\approx b(\theta_t, t)\Delta t + \sigma(\theta_t, t)\mathcal{N}(0, \Delta t)$

$$\mathcal{N}(0, \Delta t) \approx \sqrt{\Delta t}$$



随机过程

伊藤 (Itô) 公式

$(\theta_t)_{t \geq 0} \in R^d$ 满足 $d\theta_t = b(\theta_t, t)dt + \sigma(\theta_t, t)dB_t$, 定义 $X_t = f(t, \theta_t)$, 那么 $(X_t)_{t \geq 0} \in R$ 满足

$$dX_t = \partial_t f(\theta_t, t)dt + \nabla_{\theta} f(\theta_t)^T b_t dt + \frac{1}{2} \nabla_{\theta}^2 f(\theta_t, t) : \sigma_t \sigma_t^T dt + \nabla_{\theta} f(\theta_t, t)^T \sigma(\theta_t, t) dB_t$$



随机过程

Fokker Planck 方程

$(\theta_t)_{t \geq 0} \in R^d$ 满足 $d\theta_t = b(\theta_t, t)dt + \sigma(\theta_t, t)dB_t$, 若 $\theta_t \sim \rho_t$, 那么

$$\dot{\rho}_t = -\nabla_{\theta} \cdot (b_t \rho_t) + \sum_i \sum_j \frac{\partial^2}{\partial_i \partial_j} [D_{ij} \rho_t]$$

$$\text{其中 } D = \frac{1}{2} \sigma_t \sigma_t^T$$

系统（参数）在时间上随机变化 \Leftrightarrow 概率测度的演化

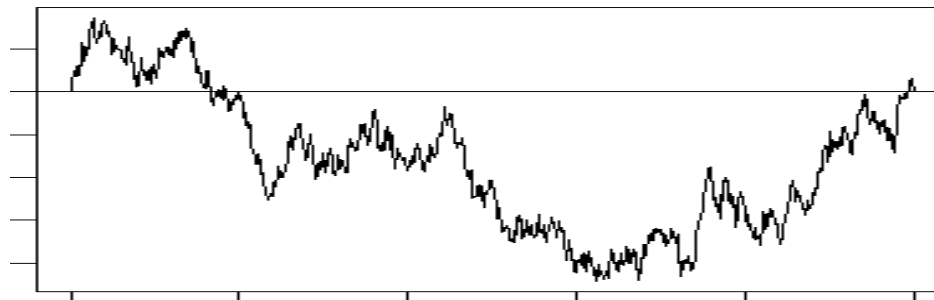


随机过程

➤ 布朗桥 $(\beta_t)_{t \geq 0} \in \mathbb{R}^d$

$$B_t | B_T = 0$$

直接构造:



$$\beta_t = B_t - \frac{t}{T} B_T, \quad (\beta_{t_1} \beta_{t_2} \cdots \beta_{t_k})^d = (B_{t_1} B_{t_2} \cdots B_{t_k} | B_T)$$

Karhunen-Loève展开:

$$\beta_t = \sum_{i=0}^{\infty} a_i \frac{\sqrt{2} \sin \frac{k\pi t}{T}}{k\pi} \quad a_i \sim \mathcal{N}(0,1)$$

随机常微分方程:

$$dX_t = -\frac{X_t}{T-t} dt + dB_t$$



扩展阅读

➤ 最优输运问题

理论:

Filippo Santambrogio, “Optimal transport for applied mathematicians”, Springer, 2015.

Lénaïc Chizat and Luca Nenna, “Introduction to Optimal Transport Theory”, <https://lchizat.github.io/ot2020orsay.html>”.

Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. "Statistical Optimal Transport", <https://arxiv.org/pdf/2407.18163>.

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).