arXiv:2603.07467v1 [stat.ML] 8 Mar 2026

基于Stein方法的概率推断与学习

Qiang Liu[1], Lester Mackey[2], 和 Chris Oates[3,4]

[1]德克萨斯大学奥斯汀分校[2]微软研究院[3]纽卡斯尔大学[4]艾伦·图灵研究所

2026年3月10日

本专著严谨地概述了运用 Stein方法进行概率推断与学习的理论和方法层面。文中给出了如何从 Stein 算子和 Stein集合构造 Stein差异的方法，并讨论了这些差异的可计算性、分离性、收敛检测和收敛控制等性质。此外，文中还详细阐述了 Stein 算子与 Stein 变分梯度下降之间的联系。文中精确陈述了主要定义和结果，并提供了所有证明的参考文献。

目录

1 引言 6

1.1 Stein方法简介 .61.2 将Stein方法作为方法论工具进行阐释 .81.3 本专著的概要 .91.4 致谢.91.5 符号说明.91.6 缩略语 .11

2 背景 12

2.1 概率论与测度论.12

2.1.1 测度 .12

2.1.2 勒贝格空间 .142.1.3 随机变量与独立性 .152.1.4 马尔可夫过程 .16

2.2 泛函分析 .18

2.2.1 再生核希尔伯特空间 .202.2.2 核均值嵌入 .232.2.3 向量值再生核希尔伯特空间 .26

2.3 统计散度 .27

2.3.1 经典散度 .282.3.2 Wasserstein 度量 .292.3.3 最大均值差异 .30

3 斯坦算子34

3.1 Langevin 算子 .353.2 扩散算子 .403.3 约束算子 .423.4 无梯度算子 .443.5 离散算子及其他 .453.5.1 Zanella 算子 .463.5.2 生灭Stein算子 .473.5.3 离散时间马尔可夫链中的算子 .483.5.4 一般Stein算子 .50

4 Stein 差异 **1**

54.1 Fisher散度 .524.2 经典Stein差异 .544.3 图 Stein 差异 .574.4 核 Stein 差异 .594.4.1 详细介绍 Langevin KSD .614.4.2 使用 Langevin KSD 区分分布.624.4.3 使用 Langevin KSD 检测收敛 .664.4.4 使用 Langevin KSD 控制收敛 .664.4.5 约束核 Stein差异 .694.4.6 一般核 Stein 差异 .694.5 随机特征 Stein 差异 .704.5.1 R 的特殊情形ΦSD .724.5.2 选择特征函数 Φ .724.5.3 选择重要性采样分布 ν .734.6 随机 Stein 差异 .754.6.1 使用 SSD 检测收敛 .764.6.2 使用 SSD 控制收敛 .76

5 Stein 动力学 **78**

5.1 动态质量传输.785.2 Kullback–Leibler散度的梯度流 .795.3 梯度流与 Stein 差异 .805.4 作为梯度流的朗之万扩散 .835.5 Stein 变分梯度 .84

6 应用 **85**

6.1 概述.856.2 衡量近似质量.876.2.1 量化采样方法的性能 . . . . . 876.2.2 拟合优度检验 .896.3 基于粒子的近似算法 .926.3.1 序贯算法 .936.3.2 批处理算法 .956.3.3 Stein重要性采样 .996.3.4 稀疏近似 .1016.3.5 积分近似.1026.4 生成模型训练 .1056.4.1 Stein对比散度.1056.4.2 最小Stein差异估计量 .1076.4.3 SteinGAN .109

6.4.4 后验近似的变分方法 .1106.4.5 变分自编码器的学习 .1126.5 梯度估计 .
113 6.5.1 使用离散 Stein 算子的梯度估计 . . . . . 114 6.5.2 策略梯度的控制变量 .
116

7 结论 **119**

1 引言

50多年来，Stein方法一直是概率论中一种强大的理论工具，能够为随机变量收敛到其极限分布提供明确的上界 [Stein， 1972]。若想了解这一研究方向的概述，可参阅几部优秀的参考文献：Barbour 和 Chen [2005],Chen 等人 [2010a],Ross [2011],Ley 等人 [2017]。我们关注的则是 Stein 方法作为一种方法论工具，如何催生用于概率推断和学习的全新而强大的算法。本文的阐述受到Gorham 和 Mackey[2015], 对可计算 Stein 差异的介绍所启发，但我们探讨的应用可追溯到Stein[1986], Stein 等人 [2004]的早期工作。关于这些进展的最新高层次综述见Anastasiou 等人 [2023], ，除此之外，主要成果散见于各类文献。因此，本文旨在为 Stein 方法在概率推断与学习中的相关严格定义和结果提供一份综合参考。

1.1 Stein方法概述

假设 P 是定义在适当空间 $\mathcal{X}$上的一个感兴趣的概率分布；我们尤其关注这样一些情形：在 P 下的大多数期望都难以轻易计算。Stein方法可以看作一种配方，用于衡量在给定的替代分布 P 下的积分对 Q下积分的近似程度。例如， P 可以表示贝叶斯分析中出现的难以处理的后验分布，而 Q 则可以表示一个由离散点构成的样本近似 $Q = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$，这些点为 $\{x_i\}_{i=1}^{n} \subset \mathcal{X}$。为衡量 Q 对 P的近似程度，理想情况下，我们会计算一个积分概率度量Müller， 1997] 它衡量最大差异

$$\sup_{f\in\mathcal{F}} |\mathbb{E}_Q[f(X)] - \mathbb{E}_P[f(X)]| = \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(x_i) - \int f(x)\, \mathrm{d}P(x) \right|$$

在一类测试函数 $\mathcal{F}$ 上，样本与目标期望之间的差异。然而，大多数标准的积分概率度量（包括第 2.3 节介绍的 Wasserstein 度量、Dudley 度量和最大均值差异）都需要在P下进行显式积分，因此无法用于我们感兴趣的目标分布。为应对这一挑战，我们将采用 Charles M. Stein（1920—2016）提出的巧妙思路，用以量化中心极限定理（CLT）中的近似误差。需要指出的是，我们的论述是针对 Stein方法的方法论应用而调整的，其侧重点与上述关于 Stein 方法理论应用的参考文献中的论述截然不同。

其中强调的方面与上述关于 Stein方法理论应用的参考文献中的论述截然不同。

*Stein*方法的核心是一个*Stein*集合$\mathcal{G}$ 和一个*Stein* 算子$\mathcal{T}_P$: $\mathcal{G} \to \mathcal{L}^1(P)$，二者共同生成在 P 下的零均值期望:

$$\int (\mathcal{T}_P g)(x)\, \mathrm{d}P(x) = 0 \quad \text{for all} \quad g \in \mathcal{G}. \tag{1.1}$$

这里$\mathcal{L}^1(P)$表示相对于 P可积的函数集合，以确保（1.1）中的积分有定义。例如，若 P 是维度为 $d = 1$的标准正态分布，且$\mathcal{G}$ 是所有单项式 $g(x) = x^m$ 的集合，其中 $m \in \mathbb{N}$，那么

$$(\mathcal{T}_P g)(x) \triangleq g'(x) - xg(x) \tag{1.2}$$

这定义了一个有效的 Stein 算子。事实上，我们可以显式地计算该期望

$$\begin{aligned}\int (\mathcal{T}_P g)(x)\, \mathrm{d}P(x) &= \int m x^{m-1} - x^{m+1}\, \mathrm{d}P(x) \\ &= \left\{ \begin{array}{ll} 0 & m \text{ is even} \\ m(m-2)!! - m!! & m \text{ is odd} \end{array} \right\} = 0\end{aligned}$$

其中 $n!!$ 表示双阶乘，即从 1 到 n 中所有与 n 奇偶性相同的数的乘积。

给定任意有效的 Stein 集合$\mathcal{G}$ 和 Stein 算子$\mathcal{T}_P$的组合，可以构造一个*Stein* 差异

$$\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}) \triangleq \sup_{g \in \mathcal{G}} \left| \int (\mathcal{T}_P g)(x)\, \mathrm{d}Q(x) \right| \tag{1.3}$$

用于衡量 Q 与 P之间的差异。Stein 差异所诱导的拓扑在很大程度上取决于G 和$\mathcal{T}_P$的选择，而一个重要挑战是理解如何构造实用的 Stein 差异（1.3）。更确切地说，我们可以探讨是否能选择$\mathcal{G}$ 和$\mathcal{T}_P$，使得以下部分（或全部）期望目标得以实现:

- **分离性**：$\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}) = 0$当且仅当 Q与 P相等（即 Stein 差异是一种有效的统计散度）。
- **收敛检测**：$\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}) \to 0$当 Q_n收敛到 P时成立，具体含义将在后文说明。
- **收敛控制**：$\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}) \nrightarrow 0$当 Q_n未能收敛到 P时成立，具体含义将在后文说明。
- **可计算性**：对于任意具有有限支撑集$\{x_i\}_{i=1}^n \subset \mathcal{X}$的分布，其 Stein 差异$\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G})$均可显式计算。

过去十年间，关于 Stein差异性质的若干正面和负面结果陆续发表，但这些结果散见于概率论、统计学和机器学习领域的文献中。本手稿的一个目标是将严谨的定义和结果汇集于一份参考文献中，并为所有结果提供证明的出处。

1.2 将 Stein方法作为方法论工具的示例说明

作为 Stein方法方法论潜力的初步示例，我们考虑在 $\mathcal{X} = \mathbb{R}^d$ [Gelman等人， 2014] 中进行贝叶斯后验推断的标准情形。给定先验密度π 和似然 $\mathcal{L}$，后验分布 P 具有密度 p，可通过贝叶斯定理在归一化常数意义下刻画：

$$p(x) \propto \pi(x)\mathcal{L}(x). \tag{1.4}$$

遗憾的是，在 大多数情况下，归一化常数（也称为边际似然，

$$\int \pi(x)\mathcal{L}(x)\,\mathrm{d}x \tag{1.5}$$

无法以闭式表示，而且对于大多数感兴趣的函数，在 p 下进行精确积分是不可行的。最大似然估计 [Geyer， 1991], 后选择推断 [Tian等人， 2016], 以及更广义的概率推断 [Neal， 1993]等情形中也自然会出现同样的不可行问题。为了近似 P 及其期望，如今已有大量实用方法可供选择，包括马尔可夫链蒙特卡洛（MCMC） [Brooks等人， 2011],近似 MCMC [（例如参见） Welling和Teh， 2011], 变分推断 [Blei等人， 2017],以及求积。然而，直到最近，从业者仍缺乏适用于衡量和比较这些种类繁多且可能彼此不一致的样本近似质量的工具。

为弥补这一空缺，Gorham 和 Mackey [2015]提出了 Stein 差异的概念，并展示了它作为实用样本质量度量的潜力。他们首先推导出一种多元且以 P为目标的推广形式，扩展了 Stein 最初针对标准正态分布提出的算子（1.2）：

$$(\mathcal{T}_P g)(x) \triangleq (\nabla \cdot g)(x) + g(x) \cdot (\nabla \log p)(x). \tag{1.6}$$

这个*Langevin* Stein 算子在 P 满足温和条件时会生成均值为零的函数（见第3章），即使归一化常数（1.5）未知，也可以计算。事实上，算子（1.6）仅通过 P 依赖于***Stein***得分，

$$s_p(x) \triangleq (\nabla \log p)(x) = (\nabla \log \pi)(x) + (\nabla \log \mathcal{L})(x),$$

它仅取决于先验和似然的导数。值得注意的是，当它与一个合适的向量场 Stein集合 $\mathcal{G}$ $g : \mathbb{R}^d \to \mathbb{R}^d$结合时，所得到的 Stein 差异

同时具备上述分离性、收敛检测、收敛控制和可计算性等所有性质（见第 4章）。

如今，Stein差异已被广泛用于衡量样本质量，但Stein方法的方法论意义更为广泛，为梯度估计、生成模型训练、拟合优度检验、偏差校正、样本质量改进和抽样等任务带来了新的有效解决方案；第6章概述了当前使用Stein方法解决的概率推断与学习领域中的若干不同问题。

1.3 专著概述

本专著旨在系统、严谨地阐述如何使用Stein方法进行概率推断与学习。我们假设读者修读过本科阶段的实分析课程，但不要求读者熟悉概率论、测度论或泛函分析；所需概念将在第2章介绍。高阶读者或许可以先跳过第2 章，仅在需要时再回头阅读。第3 章详细讨论Stein算子，第4章讨论与之相关的Stein差异。第5章介绍Stein动力学的最新进展。尽管无法涵盖Stein方法在概率推断与学习领域的所有最新及新兴应用，我们仍选取了一些值得关注的实例，在第6章加以介绍。

1.4 致谢

QL 获得了 NSF CAREER 1846421、机器学习基础研究所（IFML）和海军研究办公室的支持。
CJO 获得了 EPSRC EP/W019590/1、艾伦·图灵研究所和 Philip Leverhulme奖 PLP-2023-004 的支持。

1.5 记号

下表中，通用函数 f 是标量值的， g 是向量值的，而 A 和 B 是矩阵值的。

$\triangleq$	定义为相等
$[n]$	$\{1, \ldots, n\}$
$\binom{[n]}{m}$	$\{\sigma \subseteq [n] : \lvert\sigma\rvert = m\}$， $[n]$的所有大小为 m 的子集
$\lVert\cdot\rVert$	$\mathbb{R}^d$上的一般范数
$\mathbb{I}_S$	示性函数 $\mathbb{I}_S(x) = 1$ 当 $x \in S$ 时，否则为 $\mathbb{I}_S(x) = 0$
$\mathcal{P}$	上下文中将会明确的可测空间上所有概率测度的集合

$\|\cdot\|_p$	范数 $\|x\|_p \triangleq (\|x_1\|^p + \cdots + \|x_d\|^p)^{1/p}$ 在 $\mathbb{R}^d$, $p \in [1, \infty)$
$\|\cdot\|_\infty$	范数 $\|x\|_\infty \triangleq \max\{\|x_1\|, \ldots, \|x_d\|\}$ 在 $\mathbb{R}^d$
$\|\cdot\|_F$	Frobenius范数 $\|A\|_F \triangleq (\sum_{i,j} A_{i,j}^2)^{1/2}$
$\|\cdot\|^*$	对偶范数 $\|x\|^* \triangleq \sup_{y \in \mathbb{R}^d,\ \|y\|=1} \langle x,\ y \rangle$ 在 $\mathbb{R}^d$或 $\|A\|^* \triangleq \sup_{y \in \mathbb{R}^d,\ \|y\|=1} \|Ay\|^*$ 在 $\mathbb{R}^{d \times d}$
$\|\cdot\|_{\mathrm{op}}$	算子范数 $\|L\|_{\mathrm{op}} \triangleq \sup_{\|x\|_{\mathcal{X}} \le 1} \|L(x)\|_{\mathcal{Y}}$，对于从赋范空间（$\mathcal{X}$ $\|\cdot\|_{\mathcal{X}}$）到半赋范空间（$\mathcal{Y}$ $\|\cdot\|_{\mathcal{Y}}$）的连续线性算子 $L : \mathcal{X} \to \mathcal{Y}$
$a \cdot b$	a, $b \in \mathbb{R}^d$之间的点积
$A : B$	双点积 $\mathrm{tr}(AB^\top)$
e_j	$\mathbb{R}^d$ 的第 j 个基向量
$\partial_j f$	$x \mapsto \partial_{x_j} f(x)$ 的第 j 个偏导数
$\partial^\alpha f$	混合偏导数 $\partial_1^{\alpha_1} \ldots \partial_d^{\alpha_d} f$
∇f	梯度 $[\nabla f]_i \triangleq \partial_i f$
∇g	梯度 $[\nabla g]_{i,\ j} \triangleq \partial_j g_i$
$\nabla \cdot g$	散度 $\sum_i \partial_i g_i$
$\nabla \cdot A$	向量散度 $[\nabla \cdot\ A]_j \triangleq \nabla \cdot (A^\top e_j)$
Δf	拉普拉斯算子 $\nabla \cdot \nabla f$
$(\nabla^k f)(x)$	k阶梯度在 $x \in \mathbb{R}^d$处，即一个函数（v_1，.....，v_k） $\mapsto (\nabla^k f)(x)[v_1, \ldots, v_k]$
$C(\mathbb{R}^d)$	连续 f 的集合： $\mathbb{R}^d \to \mathbb{R}$
$C(\mathbb{R}^d,\ \mathbb{R}^p)$	连续 g 的集合： $\mathbb{R}^d \to \mathbb{R}^p$
$C_0(\mathbb{R}^d)$	在 $\mathbb{R}^d$ 上无穷远处趋于零的连续 $f : \mathbb{R}^d \to \mathbb{R}$ 的集合
$C^m(\mathbb{R}^d)$	对所有 $k \le m$ 均使 $\nabla^k f$ 连续的 $f : \mathbb{R}^d \to \mathbb{R}$ 的集合
$C^{(m,m)}(\mathbb{R}^d)$	使得对满足 $\|\alpha\|_1, \|\beta\|_1 \le m$ 的多重指标 α, β， $\partial_x^\alpha \partial_y^\beta A(x, y)$ 连续的矩阵值 A 的集合
$C_b^m(\mathbb{R}^d)$	m 时间连续可微 f 的集合： $\mathbb{R}^d \to \mathbb{R}$ 其所有阶导数 $\le m$ 均有界
$C_b^{(m,m)}\ \mathbb{R}^d\ (\ )$	矩阵值 A的集合，其中 $\partial_x^\alpha \partial_y^\beta A(x, y)$对于满足 $\|\alpha\|_1, \|\beta\|_1 \le m$的多重指标 α, β 均有界且连续
$M_0(g)$	$\sup_x \|g(x)\|_{\mathrm{op}}$
$M_k(g)$	第 k 阶 Lipschitz常数 $$\sup_{x \neq y} \frac{\|(\nabla^{k-1} g)(x) - (\nabla^{k-1} g)(y)\|_{\mathrm{op}}}{\|x - y\|_2}$$
$F_k(A)$	$\sup_{x, \|v_1\|_2=1, \ldots, \|v_k\|_2=1} \|(\nabla^k A)(x)[v_1, \ldots, v_k]\|_F$
$M_1^*(A)$	tle uda psctz 常数

$$\sup_{x \neq y} \frac{\|A(x) - A(y)\|_{\mathrm{op}}^*}{\|x - y\|_2}$$

$\mathcal{F}(f)$	$f : \mathbb{R}^d \to \mathbb{C}$ 的广义傅里叶变换

1.6 缩略词

使用以下缩写:

对比散度 对比散度 中心极限定理 中心极限定理 证据下界 证据下界 GAN 生成对抗网络 核 Stein 差异 核 Stein 差异 最大后验 最大后验 马尔可夫链蒙特卡洛 马尔可夫链蒙特卡洛 最大均值差异 最大均值差异 再生核希尔伯特空间 再生核希尔伯特空间 随机微分方程 随机微分方程 SSD 随机 Stein 差异 SVGD Stein变分梯度下降 ULA 未校正朗之万算法 变分自编码器 变分自编码器

2 背景

本专著假设读者熟悉实分析，但不要求具备概率论、测度论或泛函分析方面的知识；本章将介绍所需概念。经验丰富的读者可以跳过本章，仅在需要时再回头查阅。

2.1 概率论与测度论

虽然理解我们在第3、4和第5章中呈现的大多数结果并不需要熟悉测度论，但我们仍简要介绍其核心概念，以便准确陈述主要定义。

2.1.1 测度

我们的起始点是一个（非空）集合 Ω。

定义 2.1 (σ-代数)。 Ω 的子集族 $\mathcal{S}$ 满足

1. $\Omega \in \mathcal{S}$ 2. $\Omega \setminus S \in \mathcal{S}$ 当 $S \in \mathcal{S}$ 时 3. $S_1 \cup S_2 \cup \cdots \in \mathcal{S}$ 当 $S_1, S_2, \cdots \in \mathcal{S}$ 时

称为 Ω 上的 σ-代数。

二元组 $(\Omega, \mathcal{S})$ 称为可测空间。若 σ-代数在上下文中是明确的，我们可以仅用 Ω 来指称该可测空间。

定义 2.2 (测度). 设 $\mathcal{S}$ 为 Ω 上的 σ-代数。映射 $\mu : \mathcal{S} \to [0, \infty]$ 称为测度 如果

1. $\mu(\emptyset) = 0$ 2. $\mu(S_1 \cup S_2 \cup \dots) = \mu(S_1) + \mu(S_2) + \dots$ 当集合 $S_1, S_2, \dots$ 两两不相交时。

三元组 $(\Omega, \mathcal{S}, \mu)$ 称为测度空间。如果上下文已明确 σ-代数和测度，我们可以仅用 Ω 来指代该测度空间。若集合 $S \in \mathcal{S}$ 满足 $\mu(S) = 0$，则称其为零测集。如果测度 μ 还满足 $\mu(\Omega) = 1$，则称其为概率测度，并称 $(\Omega, \mathcal{S}, \mu)$ 为概率空间。

例 2.1 (原子测度)。 设 (Ω, S) 为一个可测空间，并固定 $\omega \in \Omega$。那么

$$\delta_\omega(S) \triangleq \begin{cases} 1 & \omega \in S \\ 0 & \omega \notin S \end{cases}$$

在(Ω, S) 上定义了一个概率测度，称为 $\omega \in \Omega$处的原子测度 .

σ-代数决定了我们可以测度哪些事件 $S \subset \Omega$。例如，相对于平凡 σ-代数$\mathcal{S} = \{\emptyset, \Omega\}$ 构造的概率测度 μ 仅包含 $\mu(\emptyset) = 0$ 与 $\mu(\Omega) = 1$这一空洞信息。

例 2.2 （离散分布）。 设Ω 为可数集，并设 S 为由Ω的全部子集构成的 σ-代数。则概率测度由每个 $\omega \in \Omega$对应的取值$\mu(\omega)$ 唯一确定，我们称 $\omega \mapsto \mu(\omega)$ 为概率质量函数。

若 Ω 带有额外的数学结构，这可能蕴含对σ-代数的一种自然选择。这里我们将 Ω 视为豪斯多夫拓扑空间，回顾其含义：任意两个不同的元素 $\omega_1, \omega_2 \in \Omega$都可以被开集$\Omega_1, \Omega_2 \subset \Omega$分离，且满足 $\omega \in \Omega_1, \omega_2 \in \Omega_2$ 与$\Omega_1 \cap \Omega_2 = \emptyset$。进一步地，若对每个 $\omega_1 \in \Omega$都存在紧集$\Omega_1 \subset \Omega$使得 $\omega_1 \in \Omega_1$，则称这样的空间 Ω 为局部紧的。容易验证，局部紧豪斯多夫空间的例子包括 $\Omega = \mathbb{R}^d$ 与 $\Omega = [0, 1]^d$。

例 2.3 （博雷尔测度）。 设Ω 是一个局部紧豪斯多夫空间，B_Ω 是包含Ω的开集的最小 σ-代数，称为博雷尔 σ-代数。测度μ ： $S \to [0, 1]$ 称为博雷尔 测度。

博雷尔 σ-代数可能并不充分，因为零博雷尔测度集合的子集未必是博雷尔可测集。在$\Omega = \mathbb{R}^d$的情形下，勒贝格测度将完备化博雷尔测度，并构成第2.1.2节勒贝格积分的基础。对于盒体$B = (a_1, b_1) \times \cdots \times (a_d, b_d) \subset \mathbb{R}^d$，$\mathrm{vol}(B) \triangleq (b_1 - a_1) \ldots\ldots (b_d - a_d)$ 表示该盒体的体积。

例 2.4 （勒贝格测度）。 对于子集 $S \subset \Omega = \mathbb{R}^d$， 勒贝格外测度定义为

$$\lambda^*(S) \triangleq \inf \left\{ \sum_{i=1}^{\infty} vol(B_i) : (B_i) \ a \ sequence \ of \ boxes \ with \ S \subset \cup_i B_i \right\}.$$

勒贝格 σ-代数 L_Ω 中的元素是满足卡拉泰奥多里准则的集合 S： $\lambda^*(A \cap S) + \lambda^*(A \cap S^c) = \lambda^*(A)$ 对所有 $A \subset \mathbb{R}^d$均成立。对于每个 $S \in L_\Omega$，勒贝格测度 $\lambda(S)$ 定义为 $\lambda(S) = \lambda^*(S)$。

定义 2.3 （可测函数）。 若函数 $f : \Omega_1 \to \Omega_2$ 从可测空间 (Ω_1, S_1) 到 (Ω_2, S_2) 满足可测条件，即 $f^{-1}(S) = \{\omega \in \Omega_1 : f(\omega) \in S\} \in S_2$ 对任意$S \in S_1$均成立，则称该函数为可测函数。

2.1.2 勒贝格空间

设 $(\Omega, \mathcal{S}, \mu)$ 是一个测度空间。若存在$a_1, \ldots\ldots, a_n \in [0,\infty]$、$S_1, \ldots\ldots, S_n \in \mathcal{S}$，使得函数 $\phi: \Omega \to [0,\infty]$ 称为简单函数，则

$$\phi(\omega) = \sum_{i=1}^{n} a_i 1_{S_i}(\omega)$$

对所有 $\omega \in \Omega$，我们定义一个积分

$$\int \phi \, \mathrm{d}\mu \triangleq \sum_{i=1}^{n} a_i \mu(S_i).$$

定义 2.4（勒贝格积分）。对于正可测函数 $f: \Omega \to [0,\infty]$，它将 $(\Omega, \mathcal{S})$ 映射到 $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$，其勒贝格积分 定义为

$$\int f \, \mathrm{d}\mu \triangleq \sup \left\{ \int \phi \, \mathrm{d}\mu : \phi \text{ is a simple function and } \phi \leq f \right\},$$

w而如果 f 也取负值，则我们将其勒贝格积分定义为 s

$$\int f \, \mathrm{d}\mu \triangleq \int f_+ \, \mathrm{d}\mu - \int f_- \, \mathrm{d}\mu$$

其中 $f_\pm = f \cdot 1_{S_\pm}$和 $S_\pm = f^{-1}([0, \pm\infty])$，只要这些积分中至少有一个是有限的。

勒贝格积分的名称源于其最初的构造，该构造涉及勒贝格测度 λ，见例2.4；但这里我们将其推广到一般测度 μ。在勒贝格测度 λ 作用于 $\Omega = \mathbb{R}^d$的特殊情形下，我们遵循标准约定，通常写作 $\int f(x) \, \mathrm{d}x$，作为 $\int f \, \mathrm{d}\lambda$的另一种记法。有时强调积分变量会更方便，此时我们写作 $\int f(\omega) \, \mathrm{d}\mu(\omega)$。另一方面，简写形式

$$\mu(f) \triangleq \int f \, \mathrm{d}\mu \tag{2.1}$$

将在第 3 章中广泛使用。将勒贝格积分的定义扩展到函数也很方便 $f: \Omega \to \mathbb{R}^d$，我们通过将上述构造应用于每个坐标函数来实现这一点 $f_i, \, i = 1, .., d.$

对于可测空间 $(\Omega, \mathcal{S})$ 上的 μ 和 ν 概率测度，如果对于每个满足$\nu(S) = 0$的集合 $S \in \mathcal{S}$，都有 $\mu(S) = 0$，则称 μ 关于 ν 绝对连续（记作 $\mu \ll \nu$）。

定理 2.1 (拉东–尼科迪姆)。设 μ 和 ν 是可测空间 $(\Omega, \mathcal{S})$ 上的概率测度，且 $\mu \ll \nu$。那么存在可测函数 $p: \Omega \to [0,\infty)$，使得$\mu(S) = \int p(\omega) 1_S(\omega) \, \mathrm{d}\nu(\omega)$ 对任意 $S \in \mathcal{S}$ 均成立。

函数 p 在定理 2.1中出现，称为 *Radon–Nikodym* 导数，它是 μ 相对于 ν 的导数，记作 $\mathrm{d}\mu/\mathrm{d}\nu$。在特殊情形下，若 ν 是例 λ2.4中的勒贝格测度，我们便得到熟悉的概率密度函数概念：

定义 *2.5*（连续分布）。设 $\Omega = \mathbb{R}^d$。令 μ 为一个概率测度，使得 $\mu(S) = \int p(\omega)1_S(\omega)\, d\lambda(\omega)$，其中 $p : \mathbb{R}^d \to [0,\infty)$ 是某个可测函数，且 λ 是例 *2.4* 中的勒贝格测度。则 $p = d\mu/d\lambda$ 称为相对于 λ的 μ 概率密度函数，或简称为 μ的（勒贝格）密度。

设 $1 \le s < \infty$。考虑由函数 $f : \Omega \to \mathbb{R}$ 构成的实向量空间 $\mathcal{L}^s(\mu)$，其中这些函数满足勒贝格积分

$$|f|_{\mathcal{L}^s(\mu)} \triangleq \left(\int |f|^s \, \mathrm{d}\mu \right)^{1/s} < \infty$$

是良定义的。赋予范数 $|\cdot|_{\mathcal{L}^s(\mu)}$ 后，空间 $\mathrm{L}^s(\mu)$ 是一个半赋范向量空间，但不是赋范空间，因为 $|f|_{\mathcal{L}^s(\mu)} = 0$ 仅能推出 f 在一个 μ-零测集之外处处为零，而不能推出 $f = 0$ 在 Ω 上处处为零。（关于函数空间范数的基础讨论见第 2.2 节。）本专著中采用的一种实用约定是，对于向量值或矩阵值函数 f，我们用简写 $f \in \mathrm{L}^s(\mu)$ 表示 f 的所有分量都属于 $\mathrm{L}^s(\mu)$。当 Ω 上的测度 μ 明确无歧义时，我们也可以写作 $\mathrm{L}^s(\Omega)$。勒贝格空间通过将仅在勒贝格零测集上不同的函数视为相同而得到：

定义 2.6（勒贝格空间）。设 $1 \le s < \infty$。勒贝格空间 $L^s(\mu)$ 是一个向量空间，其元素是 $L^s(\mu)$ 的等价类 $[f] \triangleq \{g \in L^s(\mu) : g \sim f\}$，这里的等价关系规定：$f \sim g$ 当且仅当 $|f - g|_{\mathcal{L}^s(\mu)} = 0$。空间 $L^s(\mu)$ 配备以下范数后成为赋范向量空间：

$$\|[f]\|_{L^s(\mu)} \triangleq |g|_{\mathcal{L}^s(\mu)}, \quad g \in [f],$$

后者对所有 g 都相同，因此定义良好。

当 μ 是 $\mathbb{R}^d$ 上的勒贝格测度 λ 时，通常将 $\mathcal{L}^s(\lambda)$ 写作 $\mathcal{L}^s(\mathbb{R}^d)$，将 $L^s(\lambda)$ 写作 $L^s(\mathbb{R}^d)$。

2.1.3 随机变量与独立性

概率测度的一种（几乎）等价且通常更直观的表示方式是通过随机变量来实现，现对随机变量作出精确定义：

定义 2.7（随机变量）。设$(\Omega, S\ \mu)$ 为一个概率空间，称为样本空间；设$(X, S_{\mathcal{X}})$ 为一个可测空间，称为状态空间。取值于这些样本空间和状态空间的随机变量 是一个可测函数 $X : \Omega \to X$。

随机变量 X 与状态空间$\mathcal{X}$上的概率测度 P 相关联，称为 X的分布 ，定义为：对于每个可测集 $S \in \mathcal{S}_{\mathcal{X}}$， $P(S) = \mu(X^{-1}(S))$；我们将$X \sim P$ 记作简写。随机变量 X 的期望 定义为积分 $\mu(X)$，通常记作 $\mathbb{E}[X]$。由于期望完全由分布确定，通常也记作 $\mathbb{E}_{Y\sim P}[Y]$，其中 Y可视为任意一个服从分布P的随机变量。集合 $\mathcal{S}_X$ 由形如 $X^{-1}(S)$、 $S \in \mathcal{S}_{\mathcal{X}}$ 的集合构成，其自身满足一个σ-代数的公理，因此称为由 X生成 的 σ-代数。

定义 2.8（独立性）。 设 （Ω，$\mathcal{S}$ μ） 是一个概率空间，并且每个 $\mathcal{X}_i$ 都是一个可测空间，其中索引 i 遍历一个可能不可数无限的集合 I。如果随机变量集合（ X_i ） $_{i\in I}$ X_i 个随机变量 $X_i : \Omega \to \mathcal{X}_i$ 满足以下条件，则称其为独立 的：

$$\mu\left(\cap_{i\in I} S_i\right) = \prod_{i\in I} \mu(S_i)$$

对于所有满足 $S_i \in \mathcal{S}_{\mathcal{X}_i}$ 且对每个 $i \in I$均成立的 （S_i） $_{i\in I}$ 。

定义 2.9（条件期望）。 设 $X : \Omega \to \mathbb{R}^d$ 和 $Y : \Omega \to \mathcal{Y}$ 是定义在同一概率空间 （Ω，$\mathcal{S}$ μ）上的随机变量。给定 Y的 X 的条件期望 是一个可测函数，记为 $\mathbb{E}[X|Y]$，从（Ω，$\mathcal{S}_Y$） 映射到 $\mathbb{R}^d$，并满足：

$$\int \mathbb{E}[X|Y] \cdot 1_S \, \mathrm{d}\mu = \int X \cdot 1_S \, \mathrm{d}\mu$$

对于所有 $S \in \mathcal{S}_Y$。

事件 $S \in \mathcal{S}_X$ 给定 Y的条件概率 为 $\mu(S|Y) = \mathbb{E}[1_S|Y]$。

定义 2.10（条件独立性）。 设 （Ω，$\mathcal{S}$ μ） 是一个概率空间，并且 $\mathcal{X}_i$和 $\mathcal{Y}$ 都是可测空间，其中索引 i 遍历一个可能不可数无限的集合 I。如果随机变量集合（X_i） $_{i\in I}$ X_i 个随机变量 $X_i : \Omega \to \mathcal{X}_i$ 在给定随机变量 $Y : \Omega \to \mathcal{Y}$ 的条件下满足以下条件，则称其为条件独立的：对于 μ-几乎处处，

$$\mu\left(\cap_{i\in I} S_i \mid Y\right) = \prod_{i\in I} \mu(S_i|Y)$$

对于所有$(S_i)_{i\in I}$，其中 $S_i \in \mathcal{S}_{\mathcal{X}_i}$ 对每个 $i \in I$ 均成立。

简写 $X_i \perp\!\!\!\perp X_j|Y$ 用于表示随机变量 X_i 和 X_j 在给定 Y 的条件下相互独立。

2.1.4 马尔可夫过程

在处理随机变量集合 $(X_i)_{i\in I}$ 时，指标集 I 的结构以及随机变量之间的条件独立性关系，可用于表示不同的关注情形。

定义 2.11（随机过程）。 设 Ω 是一个概率空间，$\boldsymbol{X}$ 是一个可测空间。随机过程 是一组随机变量 $(X_t)_{t\in T}$，其中随机变量 $X_t : \Omega \to \boldsymbol{X}$ 的指标 t 遍历集合 T。

在本专著中，我们主要关注两种情况；(1) 指标集 T 是离散的，例如 $t = 0$，1，2，.....，其中指标 t 被解释为与随机变量 X_t 相关联的时间；在这种情况下，该随机过程称为离散时间过程；(2) 时间指标是连续的，例如 $t \in [0, \infty)$；在这种情况下，该随机过程称为连续时间过程。在这两种情况下，指标集 T 都是有序的。

定义 2.12（马尔可夫过程）。具有如下性质的随机过程 $(X_t)_{t\in T}$

$$X_r \perp\!\!\!\perp X_t | X_s$$

对所有 $r < s < t$，均称其为马尔可夫过程。

直观而言，这一性质意味着，给定当前状态后，随机过程的未来状态与过去状态条件独立，因此，了解当前状态就足以预测未来状态。

与每个随机变量 X_t 相关联的分布记为 P_t；回顾一下，其定义为$P_t(S) \triangleq \mu(X_t^{-1}(S))$。

定义 2.13（马尔可夫核）。 一个马尔可夫过程是时间齐次的，如果对所有 $t \geq 0$，存在一个马尔可夫核 $K_t : \boldsymbol{X} \times \boldsymbol{S}_{\mathcal{X}} \to [0, \infty]$，使得

1. $x \mapsto K_t(x, S)$ 对每个 $S \in \boldsymbol{S}_{\mathcal{X}}$ 是可测的
2. $S \mapsto K_t(x, S)$ 对每个 $x \in \boldsymbol{X}$ 是一个概率测度
3. 成立如下

$$P_{s+t}(S) = \int K_t(x, S)\, \mathrm{d}P_s(x)$$

对所有 $s \geq 0$ 与 $S \in \boldsymbol{X}_{\mathcal{S}}$。

本专著中所考虑的马尔可夫过程均为时间齐次的，因而完全由其初始分布及其马尔可夫核所刻画。在用其马尔可夫核描述时间齐次马尔可夫过程时，我们常常会隐含地假定时间齐次性。

定义 2.14（不变性）。 如果马尔可夫核 K_t 使概率测度 P保持不变，则称其为

$$P(S) = \int K_t(x, S)\, \mathrm{d}P(x)$$

对所有 $t > 0$ 以及所有 $S \in \boldsymbol{S}_{\mathcal{X}}$。

例 2.5 （过阻尼朗之万扩散 I）。设 P 是定义在 $\mathbb{R}^d$上的连续分布，其在 $\mathbb{R}^d$上的密度 p 为正且可微。由随机微分方程（随机微分方程）定义的连续时间过程 $(X_t)_{t\geq 0}$

$$\begin{aligned} \mathrm{d}X_t &= \frac{1}{2}(\nabla \log p)(X_t)\,\mathrm{d}t + \mathrm{d}W_t && t > 0 \\ X_0 &\sim P_0 \end{aligned}$$

是一个时间齐次马尔可夫过程，其马尔可夫核满足 P 为不变分布，称为具有初始分布 P_0 的过阻尼朗之万扩散 ∇。这里 $p\log\ \mathbb{R}^d \to \mathbb{R}^d$表示$\log p$ 的梯度 $\mathbb{R}^d \to \mathbb{R}$，而$(W_t)_{t\geq 0}$ 表示定义在 d上的标准 $\mathbb{R}^d$维 *Wiener*过程；参见*Oksendal [2013]*第 *5* 章。

定义 2.15 (马尔可夫过程的生成元**)**。设$(X_t)_{t\geq 0}$ 是一个连续时间马尔可夫过程，其马尔可夫核为 K_t。若其存在，则该过程的无穷小生成元定义为算子

$$(Af)(x) \triangleq \lim_{t\to 0} \frac{1}{t}\left\{\int f\,\mathrm{d}K_t(x,\cdot) - f(x)\right\}$$

作用于足够正则的函数 $f : X \to \mathbb{R}$。

例 2.6 (过阻尼朗之万扩散 II**)**。例*2.5* 中的过阻尼朗之万扩散的无穷小生成元是微分算子 $(Af)(x) = (\Delta f)(x) + (\nabla f)(x)\cdot(\nabla \log p)(x)$；参见*Oksendal [2013]*第 *7.3* 节。

定义 2.16 （遍历过程）。 如果随机过程$(X_t)_{t\in T}$ 的相关测度 P_t 在 $t \to \infty$时收敛到极限 P，则称该过程是遍历的；收敛的含义需另行说明。

第 2.3 节讨论了几种不同意义下的收敛。也可以考虑更精细的遍历性概念，例如明确规定收敛速率。有关马尔可夫过程遍历性的更多背景，可参见Gallegos-Herrada 等人 [2024]。

2.2 泛函分析

泛函分析研究函数空间的数学性质。本节中，我们令$\mathcal{X}$ 为一个集合。

定义 2.17 （函数空间）。 本书中的函数空间 *F* 是一个向量空间（在实数集上），其元素是形如 $f : X \to \mathbb{R}^d$ 的函数，其中 $d \in \mathbb{N}$。

我们将遇到的最常见函数空间由标量值函数（即 $d = 1$）组成，但在第 4.4 节中，我们还需要向量值函数；此类函数空间的示例将在第 2.2.3节中讨论。

例 2.7（多项式空间 I）。 一个熟悉的函数空间是 $X = \mathbb{R}$ 上最高阶数固定的多项式空间：它是集合 $F_p \triangleq \{f(x) = a_0 + a_1 x + \cdots + a_p x^p : a_0, a_1, \ldots, a_p \in \mathbb{R}\}$，并配备逐点加法和数乘，这意味着 $(\alpha f + \beta g)(x) = \alpha f(x) + \beta g(x)$ 对所有 $f, g \in F_p$, $\alpha, \beta \in \mathbb{R}$ 和 $x \in \mathbb{R}$ 均成立。

函数空间通常可以赋予额外的数学结构，这些结构有助于理论研究和计算。特别地，当函数空间配备了范数 $\|\cdot\|_{\mathcal{F}} : \mathcal{F} \to [0, \infty)$ 时，我们称其为赋范空间；当函数空间配备了内积 $\langle \cdot, \cdot \rangle_{\mathcal{F}} : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ 时，我们称其为内积空间。所有内积都会导出相应的诱导范数，其定义为 $\|f\|_{\mathcal{F}} = \sqrt{\langle f, f \rangle_{\mathcal{F}}}$，但并非所有范数都由内积诱导。

例 2.8（多项式空间 II）。 多项式空间 $\mathcal{F}_p$（例 2.7 中的）在配备了 $\langle f, g \rangle_{\mathcal{F}_p} = a_0 b_0 + \cdots + a_p b_p$ 其中 $f(x) = a_0 + \cdots + a_p x^p$ 以及 $g(x) = b_0 + \cdots + b_p x^p$。

例 2.9（上确界范数和有界 Lipschitz 范数）。 设 ρ 是 X 上的一个度量。对于一个函数空间 F，其元素是连续有界函数 $f : X \to \mathbb{R}^d$，上确界范数和 Lipschitz 半范数分别定义为

$$\|f\|_\infty \triangleq \sup_{x \in \mathcal{X}} \|f(x)\|, \qquad |f|_{Lip,\rho} \triangleq \sup_{x \neq y,\, x,y \in \mathcal{X}} \frac{\|f(x) - f(y)\|}{\rho(x, y)}.$$

有界 Lipschitz 范数定义为 $\|f\|_{BL,\ \rho} \triangleq \|f\|_\infty + |f|_{Lip,\ \rho}$。

在度量 ρ 明确无歧义的情况下，我们将简单地用 $|f|_{\text{Lip}}$ 表示 $|f|_{\text{Lip},\ \rho}$。

赋范空间 $\mathcal{F}$ 中的元素序列 $(f_i)_{i \in \mathbb{N}}$ 称为 柯西 序列，如果对于所有 $\epsilon > 0$，都存在 $n \in \mathbb{N}$，使得对于所有 $i,\ j > n$，均有 $\|f_i - f_j\|_{\mathcal{F}} < \epsilon$。若赋范空间 $\mathcal{F}$ 中的所有柯西序列在 $\mathcal{F}$ 中也都有极限（即存在 $f \in \mathcal{F}$，使得当 $i \to \infty$ 时，$\|f_i - f\|_{\mathcal{F}} \to 0$），则称赋范空间 $\mathcal{F}$ 为 完备。

定义 2.18（希尔伯特空间）。 完备的内积空间（相对于由内积诱导的范数）称为希尔伯特空间。

例 2.10（多项式空间 III）。 例 p 中的多项式空间 2.7 是希尔伯特空间。事实上，映射 φ: $\mathbb{R}^{p+1} \to \mathcal{F}_p$ 将系数向量 $a \in \mathbb{R}^{p+1}$ 映射为相应多项式 $f(x) = a_0 + \cdots + a_p x^p$ 是 $\mathbb{R}^{p+1}$ 与 $\mathcal{F}_p$ 之间的等距映射，在该映射下，$\mathbb{R}^{p+1}$ 的完备性得以保留。

例 2.11 (勒贝格空间 $L^2(\mu)$)。 勒贝格空间 $L^p(\mu)$ 在定义 2.6 中引入，严格来说并非函数空间，因为其元素本身并不是函数，而是函数等价类。尽管如此，在 $p = 2$ 的情形下，勒贝格空间具有自然的内积

$$\langle [f], [g] \rangle_{L^2(\mu)} = \int f(\omega) g(\omega)\, \mathrm{d}\mu(\omega),$$

其诱导范数为 $\|\cdot\|_{L^2(\mu)}$，并且可以验证， $L^2(\mu)$ 配备此内积后是希尔伯特空间。

希尔伯特空间中最重要的数学结果之一是$Riesz\ representertheorem$:

定理 2.2 (Riesz 表示)。 设 H 为希尔伯特空间，且 $L: H \to \mathbb{R}$ 是一个连续线性泛函，即 L 是线性的，并且存在常数 C，使得 $|L(f)| \leq C\|f\|_{\mathcal{H}}$对所有 $f \in H$ 均成立。那么，对某个 $g \in H$，有 $L(f) = \langle f, g\rangle_{\mathcal{H}}$。

定理 g 中的元素2.2 称为*Riesz* 表示元，属于 L 空间。

2.2.1 再生核希尔伯特空间

例 $L^2(\mu)$ 中的勒贝格空间 2.11 是希尔伯特空间，但其元素并非可逐点求值的函数。在本专著的大部分内容中，我们希望利用希尔伯特空间便于使用的数学结构，同时排除某些希尔伯特空间，例如 $L^2(\mu)$，因为这类空间的元素更为抽象。为此，我们将只考虑这样的希尔伯特空间 $\mathcal{H}$： 其中逐点求值 $f \mapsto f(x)$ 是 H 上的连续线性泛函$\mathcal{H}$，因此对所有元素 $f(x)$，其各个函数值均有定义$f \in \mathcal{H}$。

定义 2.19 （再生核希尔伯特空间）。设 X 为一个集合。再生核希尔伯特空间是由 X 上实值函数构成的希尔伯特空间 H，并且对每个 $x \in X$，映射 $f \mapsto f(x)$ 都是 H 上的连续线性泛函。

由定理 2.2 可知， $f \mapsto f(x)$ 有一个 Riesz 表示元，记作$k_x \in$H。再生核希尔伯特空间（RKHS）的再生核定义为二元函数 k: X $\times$ X $\to \mathbb{R}$，满足 $k(x, y) = k_x, k_y\mathcal{H}$。由此可知，$k_x(y) = k_y, k_x\mathcal{H} = k(y, x)$，因此我们将在文中交替使用记号 k_x和 $k(\cdot\ x)$。不难验证，再生核 k 是对称且半正定的，因此属于核的一种:

定义 2.20 （核）。 二元函数 $k: X \times X \to \mathbb{R}$ 若满足以下条件，则称其为集合 X 上的核

1. 对称： k（x, y） $= k$（y, x） 对所有 $x, y \in X$ 均成立

2. 正定 半定： 对于所有 $w_1, \ldots, w_n \in \mathbb{R}$, $x_1, \ldots, x_n \in X$，以及 $n \in \mathbb{N}$，均有 t

$$\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j k(x_i, x_j) \geq 0.$$

核的第二个性质与线性代数中的一个熟悉概念密切相关：如果矩阵 $K \in \mathbb{R}^{n\times n}$ 满足半正定，则称其为半正定矩阵，其中 $w^\top K w \geq 0$ 对于所有 $w \in \mathbb{R}^n$均成立，并记作 $K \succeq 0$。如果对于所有 $w \neq 0$，不等式均为严格不等式，则称矩阵 K 为正定，记作 $K \succ 0$。因此，定义 2.24 中的第二个性质指出，对于所有 $x_1, \ldots, x_n \in \mathcal{X}$和$n \in \mathbb{N}$的取值，由 $K_{i,j} = k(x_i, x_j)$ 构成的所有矩阵K都是半正定的。如果所有此类矩阵都是（严格）正定的，则称 k 是一个正定核。

泛函分析中的一个基本结果是，再生核希尔伯特空间完全由其核刻画：

定理 2.3 （存在唯一性）。 对于所有核 $k : X \times X \to \mathbb{R}$，存在唯一的再生核希尔伯特空间，使得 k 为其再生核，记为 H_k。该希尔伯特空间被刻画为满足以下条件的唯一希尔伯特空间

1. $k_x \in H_k$ 对所有 $x \in X$，2. $f(x) = \langle f, k_x \rangle_{\mathcal{H}_k}$ 对所有 $x \in X$， $f \in H_k$，

后者称为该核的再生性质 。

定理 2.3 的证明超出了本专著的范围；参见 Steinwart 和 Christmann [2008] 的第 4.2 节。该证明本身是构造性的，并表明 $\mathcal{H}_k$ 的元素是黎曼表示元的有限线性组合（的极限）；

$$\mathcal{H}_k \triangleq \overline{\operatorname{span}\{k_x \mid x \in \mathcal{X}\}}$$

其中 b ar 符号表示关于内积取完备化 t

$$\left(\sum_{i=1}^{n} u_i k_{x_i}, \sum_{j=1}^{m} v_j k_{y_j}\right) \mapsto \sum_{i=1}^{n}\sum_{j=1}^{m} u_i v_j k(x_i, y_j).$$

判断一个一般的二元函数 k 是否为核的一种方法是考察其泰勒级数；我们不再进一步讨论这一点，但可参见Steinwart 和 Christmann [2008]中的引理 4.8。定义在 $\mathcal{X} = \mathbb{R}^d$ 上的常见核包括以下几种：

例 2.12 (高斯核)。 高斯分布 核的形式为 $k(x, y) = \exp(-\|x - y\|^2/2\ell^2)$，其中 $\ell > 0$ 为某个值。

例 2.13 (逆多二次核)。 该 逆多二次核的形式为$k(x, y) = (\ell^2 + \|x - y\|^2)^{-\beta}$，其中 $\ell > 0$， $\beta \in (0, \infty)$。

由高斯分布核和逆多二次核再生的希尔伯特空间$\mathcal{H}_k$ 是“较小的”，也就是说，其中的元素都是各阶导数均存在的函数。若要再生更大的希尔伯特空间——例如包含只有有限阶导数的函数的空间——我们可以采用其他核，例如 Matérn 核：

例 2.14 （Matérn 核）。 *Matérn* 核的形式为

$$k(x,y)=\exp\left(-\frac{\|x-y\|}{\ell}\right)\frac{s!}{(2s)!}\sum_{i=0}^{s}\frac{(s+i)!}{i!(s-i)!}\left(2\frac{\|x-y\|}{\ell}\right)^{s-i}$$

对于某些 $\ell>0$、 $s\in\mathbb{N}_0$。

由 Matérn 核再生的希尔伯特空间中的元素，其最高至 s 阶的偏导数在弱意义下存在，作为 $L^2(\mathbb{R})$ 的元素。所有这些核中出现的参数 ℓ 称为核的带宽 或长度尺度 。

例 2.12 至 2.14 中的每个核都有这样的性质： $k(x,y)$的值仅通过差 $x-y$取决于 x 和 y；这样的核称为 平移不变核。Bochner 的一个著名定理刻画了实数集 $\mathbb{R}^d$上的所有连续平移不变核:

定理 2.4 (Bochner 定理；Wendland， 2004，定理 6.6）。连续函数k： $\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$ 是平移不变核，当且仅当存在一个有限的非负（*Borel*）测度 μ ，定义在 $\mathbb{R}^d$上，称为谱测度，使得

$$k(x,y)=\frac{1}{(2\pi)^d}\int e^{-\mathrm{i}\langle x-y,\omega\rangle}\,\mathrm{d}\mu(\omega) \tag{2.2}$$

对所有 x、 $y\in\mathbb{R}^d$。

也就是说，所有平移不变核 k 都可以视为某个非负测度的傅里叶变换 μ。此外，可以证明，连续的平移不变核 k 是正定的，只要载体，即谱测度 μ （见（2.2））的载体，定义为

$$\operatorname{carrier}(\mu)\triangleq\mathbb{R}^d\setminus\{U:U\text{ is open and }\mu(U)=0\},$$

包含一个开子集；参见 Chang [1996]以及定理 6.8（Wendland [2004]）。

利用再生核的显式公式，我们便可在再生核希尔伯特空间中进行计算，例如计算范数:

$$\begin{aligned}\left\|\sum_{i=1}^{n}w_ik(\cdot,x_i)\right\|_{\mathcal{H}_k}&=\sqrt{\left\langle\sum_{i=1}^{n}w_ik(\cdot,x_i),\sum_{j=1}^{n}w_jk(\cdot,x_j)\right\rangle_{\mathcal{H}_k}}\\&=\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}w_iw_j\langle k(\cdot,x_i),k(\cdot x_j)\rangle_{\mathcal{H}_k}}\\&=\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}w_iw_jk(x_i,x_j)}.\end{aligned}$$

这是再生核希尔伯特空间的一个相当显著的特点，因为通常很难直接根据核显式刻画这些空间中的元素。以下是一些值得注意的例外，其中再生核希尔伯特空间和核都具有显式形式:

例 2.15 (多项式空间 IV)。 例 $\mathcal{F}_P$ *2.7* 节中的 多项式空间 是一个再生核希尔伯特空间，其核为 $k(x,y) = 1 + (xy) + \cdots + (xy)^p$。

例 2.16 （高斯核 II）。 例 k 中的高斯核 *2.12* （为简单起见，维度为 $d = 1$ ）再生的希尔伯特空间为

$$\mathcal{H}_k = \left\{ f = \sum_{i=0}^{\infty} c_i \phi_i \ : \ \|f\|_{\mathcal{H}_k}^2 \triangleq \sum_{i=1}^{\infty} c_i^2 < \infty \right\},$$

其中

$$\phi_i(x) \triangleq \frac{1}{\ell^i \sqrt{i!}} x^i \exp\left(-\frac{x^2}{2\ell^2}\right).$$

参见*Steinwart* 等人 *[2006]*。

2.2.2 核均值嵌入

再生核希尔伯特空间所具有的便利数学结构，与概率测度集合上可用的有限结构形成鲜明对比；一般而言，概率测度集合并非向量空间。尽管如此，仍可将（具有足够正则性的）概率测度嵌入再生核希尔伯特空间，从而利用其数学结构。本节中， $k : \mathrm{X} \times \mathrm{X} \to \mathbb{R}$是定义在可测空间 X 上的正定核。假设对于所有 $x \in \mathrm{X}$， k_x 均可测；这等价于假设 H_k 中的元素均可测 [Steinwart 和 Christmann，2008，引理4。24]。令 $\mathrm{P}_{\mathcal{H}_k}$ 为 X 上概率分布 μ 的集合，其中映射 $\mathrm{I}_\mu : \mathrm{H}_k \to \mathbb{R}$ 由 $\mathrm{I}_\mu(f) = \int f \, \mathrm{d}\mu$ 给出，且为连续线性泛函。

定义 2.21 （核均值嵌入）。 对于 $\mu \in P_{\mathcal{H}_k}$， I_μ 的 *Riesz* 表示元称为核均值元，记作 $\phi_\mu \in H_k$。映射

$$\begin{aligned} \phi : \mathcal{P}_k &\to \mathcal{H}(k) \\ \mu &\mapsto \phi_\mu \end{aligned}$$

称为 $P_{\mathcal{H}_k}$ 到 H_k 的核均值嵌入 。

根据定理 2.2，此定义是良定义的。核均值嵌入使我们能够直接利用 $\mathcal{H}_k$ 的希尔伯特空间结构，对 $\mathcal{P}_{\mathcal{H}_k}$进行分析和计算。例如，我们可以通过连接一对测度 μ 与 ν 的核均值元 $\phi_\mu - \phi_\nu$ 和 ϕ_μ 的向量的模来衡量它们之间的不相似度； ϕ_ν我们将在第 2.3.3 节详细讨论这一思路。

为了计算与概率测度 μ相关联的核均值元，我们可以在Riesz表示陈述$\mathrm{I}_\mu(f) = \langle f, \phi_\mu \rangle_{\mathcal{H}_k}$中取$f = k_x$，从而立即看出

$$\phi_\mu(x) = \int k(x,y) \, \mathrm{d}\mu(y). \tag{2.3}$$

对于某些概率测度对 μ 与核 k，例如高斯分布与高斯核配对时，可以解析地计算核均值元 [Briol 等人， 2025]。在这些可处理的情形之外，可以使用数值方法来近似核均值元。例如，若 $\int k(x,x)\,\mathrm{d}\mu(x) < \infty$ 则该蒙特卡洛估计量

$$\phi_\mu^n(\cdot) \triangleq \frac{1}{n}\sum_{i=1}^n k(\cdot, Y_i), \qquad Y_1,\ldots,Y_n \overset{\text{iid}}{\sim} \mu \tag{2.4}$$

为 ϕ_μ提供了强相合近似，其含义是 $\|\phi_\mu^n - \phi_\mu\|_{\mathcal{H}_k} \to 0$ 几乎必然成立，当 $n \to \infty$。此外，值得注意的是 ϕ_μ^n 可以解释为与近似测度 $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$相关联的精确的核均值元。关于核均值元的蒙特卡洛估计，更多讨论见第2.3.3节。人们还开发了多种更复杂的核均值元近似方法，例如准蒙特卡洛、行列式点过程、梯度流和核驱赶；此处不打算详细综述这些方法，但第6.3节将介绍其中一部分与 Stein方法尤为相关的算法。

一般而言，无法将所有概率测度嵌入同一个希尔伯特空间；对于给定的核 k ，集合 $\mathcal{P}_{\mathcal{H}_k}$ 决定了哪些测度可以安全地嵌入。现在将讨论 $\mu \in \mathcal{P}_{\mathcal{H}_k}$ 的充分条件。回顾第$\mathcal{L}^p(\mu)$ 表示函数 $f: \mathcal{X} \to \mathbb{R}$ 所构成的向量空间，其中半范数$|f|_{\mathcal{L}^p(\mu)} = \int |f|^p \,\mathrm{d}\mu$ 有限，见第2.1.2节。本文此处的阐述遵循Barp 等人 [2024, 附录C]。

定义 2.22 (标量可积)。 若映射 $\Phi\colon \mathcal{X} \to \mathcal{H}_k$ 满足以下条件，则称其为 标量可积 μ-映射： $\{x \mapsto \langle h, \Phi(x)\rangle_{\mathcal{H}_k} : h \in \mathcal{H}_k\} \subset \mathcal{L}^1(\mu)$。

以下证明中使用的论证可追溯至 Dunford [1937], ，我们的论述基于 Schwabik 和 Ye [2005, 引理 2.1。1]。

命题 2.1。 若 $\Phi\colon X \to H_k$ 是标量可积的，则

$$\begin{aligned} \mathrm{I}_\Phi : \mathcal{H}_k &\to \mathbb{R} \\ h &\mapsto \int \langle h, \Phi(x)\rangle_{\mathcal{H}_k}\,\mathrm{d}\mu(x) \end{aligned}$$

是连续线性泛函。

证明。 首先，我们断言线性映射的图

$$\begin{aligned} T : \mathcal{H}_k &\to \mathcal{L}^1(\mu) \\ h &\mapsto (x \mapsto \langle h, \Phi(x)\rangle_{\mathcal{H}_k}). \end{aligned}$$

是闭的。为此，令 $h_n \to h$ 属于 $\mathcal{H}_k$ ，并假设 $T(h_n) \to g$ 属于 $\mathcal{L}^1(\mu)$。要证明的是， g 和 $T(h)$ 在 $\mathcal{L}^1(\mu)$ 中相等。由于在 $\mathcal{L}^1(\mu)$ 中收敛的每个序列都有一个几乎处处收敛的子序列，因此存在一个子序列 $(h_{n_i})_{i\in\mathbb{N}}$ ，使得

$$\langle h_{n_i}, \Phi(x)\rangle_{\mathcal{H}_k} \to g(x)$$

对 μ-几乎所有 $x \in \mathbf{X}$ 都成立。由于对所有 $x \in \mathbf{X}$，$h_n, \Phi(x)_{\mathcal{H}_k} \to h, \Phi(x)^{\mathcal{H}_k}$，因此对 μ-几乎所有 $x \in \mathbf{X}$，$g(x) = h, \Phi(x)_{\mathcal{H}_k} = T(h)(x)$。因此，$g$和 $T(h)$ 在 $\mathrm{L}^1(\mu)$ 中相等，并且 T 的图确实是闭的。闭图定理的条件现已得到验证。该定理指出，巴拿赫空间之间的线性映射 T 有界，当且仅当其图是闭的。因此，T 是连续的，并且

$$\|T\|_{\mathrm{op}} = \sup_{\|h\|_{\mathcal{H}_k} \leq 1} \int |T(h)| \, \mathrm{d}\mu < \infty.$$

因此我们可以得出，I_Φ 是一个连续线性泛函，因为

$$\begin{aligned} |\mathrm{I}_\Phi(h)| = \left| \int \langle h, \Phi(x) \rangle_{\mathcal{H}_k} \, \mathrm{d}\mu(x) \right| &\leq \int |\langle h, \Phi(x) \rangle_{\mathcal{H}_k}| \, \mathrm{d}\mu(x) \\ &= \int |T(h)| \, \mathrm{d}P\mu \leq \|T\|_{\mathrm{op}} \|h\|_{\mathcal{H}_k}, \end{aligned}$$

正如所要求的那样。 □

这些结果给出了以下优雅的刻画：哪些概率测度构成的集合$\mathcal{P}^{\mathcal{H}_k}$ 能够安全地嵌入$\mathcal{H}_k$：

命题 2.2 （$\mathcal{P}_k$的刻画）．$\mu \in \mathcal{P}^{\mathcal{H}_k}$ 当且仅当 $\mathcal{H}_k \subset \mathcal{L}^1(\mu)$。

证明。取 $\Phi(x) = k(\cdot\, x)$ 为典范特征映射，因此 k 可测意味着标量函数 $x \mapsto h, \Phi(x)$ $_{\mathcal{H}_k} = h(x)$ 可测，且 $\mathrm{I}^\mu = \mathrm{I}_\Phi$。若 $\mathrm{H}_k \subset \mathrm{L}^1(\mu)$，则 $x \mapsto k(\cdot\, x)$ 是标量 μ-可积的，而命题 2.1 表明 I^μ 是一个连续线性泛函。反之，若 $\mathrm{H}_k \not\subset \mathrm{L}^1(\mu)$，则显然 I^μ 不连续，因此 $\mu \notin \mathrm{P}^{\mathcal{H}_k}$。

□

因此，可嵌入哪些测度的问题归结为 $\mathcal{H}_k \subset \mathcal{L}^1(\mu)$ 是否成立。以下结果给出了一个著名的充分条件：

命题 2.3。如果 $\int \sqrt{k(x, x)} \mathrm{d}\mu(x) < \infty$ 那么 $\mathcal{H}_k \subset \mathcal{L}^1(\mu)$。

证明。对于 $h \in \mathcal{H}_k$，由再生性质和柯西–施瓦茨不等式，

$$\int |h(x)| \, \mathrm{d}\mu(x) = \int |\langle h, k(\cdot, x) \rangle| \, \mathrm{d}\mu(x) \leq \|h\|_{\mathcal{H}_k} \int \|k(\cdot, x)\|_{\mathcal{H}_k} \, \mathrm{d}\mu(x),$$

其中，再次由再生性质可得 $\|k(\cdot\, x)\|_{\mathcal{H}_k} = \sqrt{k(x, x)}$，证毕。 ———

□

在对 k 和 $\mathcal{X}$ 作出温和假设的情况下，可以放宽上述可积条件：

命题 2.4。如果 k 连续，X 可分，并且

$$\iint |k(x, y)| \, \mathrm{d}\mu(x) \mathrm{d}\mu(y) < \infty,$$

那么 $H_k \subset L^1(\mu)$。

证明。由于 k 可测且取实值， k 按照 Carmeli 等人 [2006, 第 3 节1]的意义是强可测的。此外，由于 k 是强可测的，且 $\int\int |k(x,y)|\, \mathrm{d}\mu(x)\mathrm{d}\mu(y) < \infty$，那么 k 按照 Carmeli 等人 [2006, 定义 4.1]的意义是 ∞-有界的；参见 Carmeli 等人 [2006, 推论 4.3]。由于 k 连续且 X 可分，因此 H_k 是可分的（Carmeli 等人，2006年，推论 5.2]）。由于 H_k 是可分的，k 为 ∞-有界等价于 $\mathrm{H}_k \subset \mathrm{L}^1(\mu)$ [（Carmeli 等人，2006年，命题 4.4]）。 □

2.2.3 向量值再生核希尔伯特空间

本专著后文中多元 Stein 差异的发展依赖于将再生核希尔伯特空间推广到向量值函数空间。为了推广有界线性泛函的概念，该概念曾被用于定义（标量值）再生核希尔伯特空间（定义2.19），我们称映射 $L : \mathcal{H} \to \mathbb{R}^d$ 作用于希尔伯特空间 $\mathcal{H}$ 为连续线性算子，如果 L 是线性的，且存在常数 C 使得 $\|L(f)\| \leq C\|f\|_{\mathcal{H}}$ 对所有 $f \in \mathcal{H}$。

定义 2.23 （向量值再生核希尔伯特空间）。 设 X 为一个集合。一个向量值再生核希尔伯特空间是由 X 上的向量值函数构成的希尔伯特空间 H，使得对每个 $x \in X$，映射 $f \mapsto f(x)$ 是 H 上的连续线性算子。

若 $L : \mathrm{H} \to \mathbb{R}^d$ 是连续线性算子，则其各分量 L_i 都是连续线性泛函，因此可以应用 Riesz 表示定理。由此可知，若 H 是一个向量值再生核希尔伯特空间，其元素是形如 $f : \mathrm{X} \to \mathbb{R}^d$ 的函数，那么对每个 $x \in \mathrm{X}$ 和 $i \in \{1, \ldots, d\}$，都存在一个 Riesz 表示元 $K_{x,i} \in \mathrm{H}$，对应于泛函 $f \mapsto f_i(x)$，也就是说，对于所有 $f \in \mathrm{H}$，均有 $f_i(x) = f, K_{x,i}\mathcal{H}$。向量值再生核希尔伯特空间的再生核定义为二元函数 $K : \mathrm{X} \times \mathrm{X} \to \mathbb{R}^{d\times d}$，满足 $K_{i,j}(x,y) = K_{x,i}, K_{y,j}\mathcal{H}$。可以验证，再生核 K 是一种矩阵值核：

定义 2.24 (矩阵值核)。 若二元函数 $K : X \times X \to \mathbb{R}^{d\times d}$，其中 $d \in \mathbb{N}$ 为某个元素，则称其为集合 X 上的核，如果它满足

1. transpose-symmetric; $K(x,y) = K(y,x)^\top$ *for all* $x, y \in \mathcal{X}$
2. 半正定；

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\langle u_i, K(x_i,x_j)u_j\rangle \geq 0 \tag{2.5}$$

对所有 x_1、……、$x_n \in X$，所有 u_1、……、$u_n \in \mathbb{R}^d$，以及所有 $n \in \mathbb{N}$。

为清楚起见，我们强调，$\langle\cdot,\ \cdot\rangle$（2.5）中出现的内积是 $\mathbb{R}^d$ 上通常的欧几里得内积。

令 $K_x \triangleq K(\cdot\ x)$，使得 $K_x : \mathcal{X} \to \mathbb{R}^{d\times d}$ 为矩阵值。类比标量值核的情形，给定一个核，存在唯一的希尔伯特空间由其再生：

定理 2.5 (存在唯一性)。 对于所有核 $K : X \times X \to \mathbb{R}^{d\times d}$，都存在唯一的希尔伯特空间，使得 K 是一个再生核，记作 H_K。该希尔伯特空间的特征是，它是满足以下条件的唯一希尔伯特空间

1. $K_x u \in H_K$ 对所有 $x \in X$， $u \in \mathbb{R}^d$,
2. $\langle f, K_x u\rangle_{\mathcal{H}_K} = \langle f(x), u\rangle$ *for all* $f \in \mathcal{H}_K$, $x \in \mathcal{X}$, $u \in \mathbb{R}^d$,

其中后者称为核的再生性质 。

定理的证明2.5 可见于 Carmeli 等人 [2006, 命题 2。1]。 $\mathcal{H}_K$ 的元素是形如 $K_x u$的向量值函数有限线性组合的（极限），其中 $u \in \mathbb{R}^d$；也就是说,

$$\mathcal{H}_K \triangleq \overline{\operatorname{span}\{K_x u : x \in \mathcal{X}, u \in \mathbb{R}^d\}}$$

其中，此处的完备化是相对于内积

$$\left(\sum_{i=1}^{n} K_{x_i} u_i, \sum_{j=1}^{m} K_{y_j} v_j\right) \mapsto \sum_{i=1}^{n}\sum_{j=1}^{m} \langle u_i, K(x_i, y_j) v_j\rangle.$$

矩阵值核中最简单的例子是由标量值核构造而成的核，对应的向量值再生核希尔伯特空间可以明确表征为:

例 2.17 （对角矩阵值核）。 设 $k_i : X \times X \to \mathbb{R}$ 对每个 $i \in \{1, \ldots, d\}$都是一个（标量值）核。则 $K(x, y) \triangleq \operatorname{diag}(k_1(x, y) \ldots, k_d(x, y))$ 是一个矩阵值核。其对应的向量值希尔伯特空间的内积为

$$\langle f, g\rangle_{\mathcal{H}_K} = \sum_{i=1}^{d} \langle f_i, g_i\rangle_{\mathcal{H}_{k_i}}.$$

例 2.18 (张量积矩阵值核)。 设 $k : X \times X \to \mathbb{R}$ 是一个（标量值）核，且 $B \in \mathbb{R}^{d\times d}$ 是一个对称正定矩阵。那么 $K(x, y) \triangleq k(x, y)B$ 是一个矩阵值核。其对应的向量值希尔伯特空间的内积为

$$\langle f, g\rangle_{\mathcal{H}_K} = \sum_{i=1}^{d}\sum_{j=1}^{d} B_{i,j}\langle f_i, g_j\rangle_{\mathcal{H}_k}.$$

2.3 统计散度

本节介绍用于量化两个概率分布差异程度的工具。其抽象表述是一种统计散度:

定义 2.25 （统计散度）。 设 P 为定义在同一个可测空间（Ω，S）上的概率测度集合。映射 $D: P \times P \to [0,\infty]$ 称为统计散度，若 $D(\mu, \nu) \geq 0$，且当且仅当 $\mu = \nu$ 时取等号。

$\mathcal{P}$ 上的任意度量自动构成一种统计散度，但一般而言，统计散度不必对称，也不必满足三角不等式。接下来我们回顾几种已有的统计散度，并指出，在 Stein 方法的应用背景下，这些散度可能难以显式计算，甚至无法计算。

2.3.1 经典散度

或许最经典的统计散度就是 Kullback–Leibler散度。回顾一下第 2.1.2 节中引入的拉东–尼科迪姆导数概念。

定义 2.26 (Kullback–Leibler散度)。 设 μ 和 ν 是可测空间(Ω,S) 上的概率测度，且满足 $\mu \ll \nu$。Kullback–Leibler散度 定义为

$$\mathrm{KL}(\mu||\nu) \triangleq \int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\mu$$

约定：当积分无定义时，$\mathrm{KL}(\mu||\nu)$ 为无穷大。

Kullback–Leibler散度确实是一种统计散度，因为 $\log(x) \leq x - 1$意味着

$$\mathrm{KL}(\mu||\nu) = -\int \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right) \mathrm{d}\mu \geq -\int \frac{\mathrm{d}\nu}{\mathrm{d}\mu} - 1 \, \mathrm{d}\mu = 0$$

当且仅当对于 ν-几乎所有 $\omega \in \Omega$，$(\mathrm{d}\mu/\mathrm{d}\nu)(\omega) = 1$ 。Kullback–Leibler散度在信息论领域居于核心地位，在该领域中，它被称为从 ν 到 μ的相对熵。与其他统计散度（见下文）相比，Kullback–Leibler散度是一种较强的散度概念，要求 μ 和 ν 具有相同的支撑集。例如，若 $x_n \to x$ 且对所有 $n \in \mathbb{N}$、 $\mu_n = \delta_{x_n}$ 和 $\nu = \delta_x$都有 $x_n \neq x$ ，则 $\mathrm{KL}(\mu_n||\nu) \nrightarrow 0$。这里， δ_x 表示 x处的原子测度，定义见例 2.1。

另一种替代方案且往往更实用的概念是弱收敛。弱收敛有几种等价表述；但从统计散度的角度来看，我们可以使用Dudley 度量来度量弱收敛 [Dudley，2018年，第11]章：

定义 2.27 （Dudley 度量）。设 $(\mathcal{X}\ \rho)$ 是一个赋有 *Borel*σ-代数的度量空间。设 $\|\cdot\|_{BL,\rho}$ 表示例 *2.9* 中的有界 *Lipschitz* 范数。对于 *2.9*。设 μ 和 ν*X* 上的（*Borel*）概率分布，$\mathcal{X}$上的 Dudley （或有界 Lipschitz）度量 定义如下：

$$\mathrm{BL}_\rho(\mu,\nu) \triangleq \sup_{\|f\|_{BL,\rho}\leq 1} \mu(f) - \nu(f),$$

其中我们回顾简写 $\mu(f) = \int f \, \mathrm{d}\mu$ ，该简写在 （2.1）中引入。

若一个概率分布序列 $(\mu_n)_{n\in\mathbb{N}}$ 依分布收敛 到概率分布 ν，当 $\mathrm{BL}_\rho(\mu_n,\nu) \to 0$时，我们将其简写为 $P_n \xrightarrow{d} P$。等价地，设随机变量序列 X_n 定义在同一个概率空间 Ω 上：$\Omega \to \mathcal{X}$，且满足 $X_n \sim P_n$；如果 $P_n \xrightarrow{d} P$，则称其弱收敛 到随机变量 $X : \Omega \to \mathcal{X}$，且满足 $X \sim P$。不难验证，BL_ρ 是 $\mathcal{X}$ 上所有（Borel）概率分布集合上的一个度量，尤其是一个统计散度。

计算 Kullback–Leibler散度与 Dudley 度量，需要同时掌握分布 μ 和 ν 的详细信息，或通过其 Radon–Nikodym 导数，或通过其广义矩。然而，在我们的动机背景中，所涉及的分布之一是通过难以处理的归一化常数隐式定义的。在这种情况下，便需要其他统计散度。

2.3.2 Wasserstein 度量

一个重要的统计散度由 Wasserstein 度量给出。对于度量空间 (X, ρ) 上的概率分布 μ 和 ν，μ和 ν的耦合是 $\mathrm{X} \times \mathrm{X}$ 上的一个概率分布，其边缘分布为 $\mu(S) = \gamma(S \times \mathrm{X})$ 与 $\nu(S) = \gamma(\mathrm{X} \times S)$。令 $\Gamma(\mu,\nu)$ 表示 μ 和 ν 的所有可能耦合构成的集合。令 $\mathrm{P}_s(\mathrm{X})$ 表示可测空间 X 上的概率分布μ 的集合，使得对任意（因而对所有） $x_0 \in \mathrm{X}$，其 s阶矩 $\int \rho(x,x_0)^s \, \mathrm{d}\mu(x)$ 有限。

定义 2.28 (s-Wasserstein 度量)。设 ρ 是 *Radon*[1] 空间 $\mathcal{X}$上的一个度量。对于μ，$\nu \in \mathcal{P}_s(\mathcal{X})$，$s$-Wasserstein 度量定义为

$$\mathrm{W}_\rho^s(\mu,\nu) \triangleq \left(\inf_{\gamma\in\Gamma(\mu,\nu)} \int \rho(x,y)^s \, \mathrm{d}\gamma(x,y) \right)^{1/s}.$$

可以证明，定义 2.28 在集合 $\mathrm{P}_s(\mathrm{X})$ 上定义了一个度量，尤其是一种统计散度 [Clement and Desch, 2008]。 s-Wasserstein 度量常用于最优传输的应用中，因为它天然适用于 X 的度量结构；例如，对于所有 $x, y \in \mathrm{X}$，$\mathrm{W}_\rho^s(\delta_x,\delta_y) = \rho(x,y)$。序列 $(\mu_n)_{n\in\mathbb{N}}$在 $\mathrm{W}_\rho^s(\mu_n,\nu) \to 0$ 意义下收敛到极限 ν，等价于以下陈述：$\mu_n \xrightarrow{d} \nu$，且对于某个 $x_0 \in \mathrm{X}$，有 $\int \rho(x,x_0)^s \, \mathrm{d}\mu_n(x) \to \int \rho(x,x_0)^s \, \mathrm{d}\nu$。因此，一般而言， s-Wasserstein 度量诱导的拓扑比定义 2.27 中的 Dudley 度量更强。在 $s = 1$ 的情形下，这一点尤为明显，此时 1-Wasserstein 度量具有对偶表示

$$\mathrm{W}_\rho^1(\mu,\nu) = \sup_{|f|_{\mathrm{Lip},\rho}\le 1} \mu(f) - \nu(f), \tag{2.6}$$

其中 $|\cdot|_{\mathrm{Lip},\,\rho}$ 是例 2.9中的 Lipschitz 半范数。因此，与 Dudley 度量相比，1-Wasserstein 度量涉及在更大的测试函数集合上取上确界

[1]所有可分完备度量空间都是 Radon 空间。

f 通过放宽有界性要求。显然，由于连续函数在 $\mathbb{R}^d$ 的紧子集上会自动有界，当$\mathcal{X}$ 为紧集时，1-Wasserstein 度量所诱导的拓扑与 Dudley 度量所诱导的拓扑一致。在度量 ρ 明确无歧义的情况下，我们将简写 W^s 为 W^s_ρ。

除标准的1-Wasserstein 度量外，我们还引入一种倾斜版本：给定倾斜函数 $t:\mathbb{R}^d\to(0,\infty)$，定义[2]

$$\mathrm{W}^{1,t}_\rho(\mu,\nu)\triangleq\sup_{|f|_{\mathrm{Lip},\rho}\le 1}\mu(tf)-\nu(tf).$$

倾斜Wasserstein距离之间不存在一般的支配关系；它们所诱导的拓扑各不相同，并取决于倾斜函数 t。请注意，当 $t(x)\equiv 1$时，可还原标准的1-Wasserstein距离（2.6）。与标准Wasserstein距离一样，倾斜Wasserstein距离所诱导的拓扑要弱得多，例如弱于库尔贝克-莱布勒或海林格等散度，因为它们不要求测度具有绝对连续性。

尽管Wasserstein度量优雅且自然，但它们并不适用于我们所关注情境中的计算。事实上，计算 $\mathrm{W}^s_\rho(\mu,\nu)$ 需要对 μ 和 ν进行积分，而对于其中至少一个分布，这通常难以处理。有时可以采用数值近似方案，但这些方案会遇到难以承受的样本复杂度：假设 μ 和 ν 定义在 $\mathbb{R}^d$上，并且我们只能通过数量为 n （可能很大）的独立样本 $X_1,\ldots,X_n\sim\mu$和 $Y_1,\ldots,Y_n\sim\nu$来访问各个分布。近似计算 $\mathrm{W}^s_\rho(\mu,\nu)$的一种自然方法是使用插件估计量 $\mathrm{W}^s_\rho(\mu_n,\nu_n)$，其中 $\mu_n=\frac{1}{n}\sum_{i=1}^n\delta_{X_i}$和 $\nu_n=\frac{1}{n}\sum_{i=1}^n\delta_{Y_i}$。然而，这种近似会产生大小为 $O_P(n^{1/d})$的误差。也就是说，计算Wasserstein度量的难度会随着分布所定义域的维数 d 呈指数级增长；参见 Sriperumbudur 等 [2012]。此外，我们的一个主要动机是使用差异来评估近似采样算法的质量，而进行这种评估又要求有精确样本，因此会陷入先有鸡还是先有蛋的困境。

2.3.3 最大均值差异

本节介绍最大均值差异，它基于第 2.2.2 节介绍的核均值嵌入。对于 1-Wasserstein 度量而言，它提供了另一种选择，其样本复杂度基本上与维度无关。回顾一下，与分布 μ 对应的核均值元记作 ϕ_μ；参见定义 2.21。

定义 2.29 （最大均值差异）。设 $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ 是一个正定核，且 $\mu,\nu\in P_{\mathcal{H}_k}$。两个分布$\mu,\nu\in P_{\mathcal{H}_k}$之间的最大均值差异 为$\mathrm{MMD}_k(\mu,\nu)\triangleq\|\phi_\mu-\phi_\nu\|_{\mathcal{H}_k}$。

[2]明确地说，这里 tf表示逐点乘积 $(tf)(x)\triangleq t(x)f(x)$。

为了理解为什么与 Wasserstein 度量相比，最大均值差异（MMD）具有更优的样本复杂度，我们首先推导一种等价形式，使其代数表达更加显式。事实上，

$$\begin{aligned}\mathrm{MMD}_k^2(\mu,\nu) = \|\phi_\mu - \phi_\nu\|_{\mathcal{H}_k}^2 &= \langle \phi_\mu - \phi_\nu, \phi_\mu - \phi_\nu\rangle_{\mathcal{H}_k} \\ &= \langle \phi_\mu, \phi_\mu\rangle_{\mathcal{H}_k} - 2\langle \phi_\mu, \phi_\nu\rangle_{\mathcal{H}_k} + \langle \phi_\nu, \phi_\nu\rangle_{\mathcal{H}_k}\end{aligned} \tag{2.7}$$

例如，考虑$\langle \phi_\mu, \phi_\nu\rangle_{\mathcal{H}_k}$这一项，根据 ϕ_ν是I_ν 的 Riesz 表示元，以及 (2.3)，可得

$$\begin{aligned}\langle \phi_\nu, \phi_\mu\rangle_{\mathcal{H}_k} = \mathrm{I}_\nu(\phi_\mu) &= \int \phi_\mu(y)\, \mathrm{d}\nu(y) \\ &= \iint k(x,y)\, \mathrm{d}\mu(x)\mathrm{d}\nu(y).\end{aligned} \tag{2.8}$$

对 (2.7) 中的三个项都进行类似处理，可得显式表达式

$$\begin{aligned}\mathrm{MMD}_k^2(\mu,\nu) = &\iint k(x,y)\mathrm{d}\mu(x)\mathrm{d}\mu(y) - 2\iint k(x,y)\mathrm{d}\mu(x)\mathrm{d}\nu(y) \\ &+ \iint k(x,y)\mathrm{d}\nu(x)\mathrm{d}\nu(y).\end{aligned} \tag{2.9}$$

在 (2.9) 中的积分无法精确计算，但仍可从 μ 和 μ 中抽样的情况下，可分别用基于独立样本 $X_1,\ldots,X_n \sim \mu$ 和 $Y_1,\ldots,Y_n \sim \nu$ 的蒙特卡洛估计量 ϕ_μ^n 和 ϕ_ν^n，替换难以计算的核均值元 ϕ_μ 和 ϕ_ν，从而近似最大均值差异；这些估计量定义见 (2.4)。根据我们在 (2.4) 后的观察，ϕ_μ^n 可解释为对应于近似测度 $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ 的精确核均值元，因此，插件估计量 $\|\phi_\mu^n - \phi_\nu^n\|_{\mathcal{H}_k}$ 也可作另一种解释：它计算的是两个近似测度 $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ 和 $\nu_n = \frac{1}{n}\sum_{j=1}^n \delta_{Y_j}$ 之间的精确最大均值差异$_k(\mu_n, \nu_n)$。从这个角度看，要得出以这种方式获得的最大均值差异近似具有一致性的条件并不困难。事实上，由 (2.9) 可得

$$\begin{aligned}\mathrm{MMD}_k^2(\mu_n,\nu_n) = &\frac{1}{n^2}\sum_{i=1}^n\sum_{i'=1}^n k(X_i, X_{i'}) - 2\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(X_i, Y_j) \\ &+ \frac{1}{n^2}\sum_{j=1}^n\sum_{j'=1}^n k(Y_j, Y_{j'})\end{aligned} \tag{2.10}$$

其中这三个项分别表示式（2.9）中相应积分的蒙特卡洛近似，其误差以 $O_P(n^{-1/2})$ 的速率收敛。因此，最大均值差异的样本复杂度基本上与维度无关。此外，蒙特卡洛估计量（）（2.10）所涉及的表面计算复杂度，可以通过注意到计算是“易并行的”来缓解。

尽管与 Wasserstein度量相比，最大均值差异在计算方面具有优势，但在我们的动机背景下，它们同样不适合计算。事实上，使用来自 μ 和 ν 的精确样本来近似最大均值差异$_k(\mu,\ \nu)$ 是有问题的，因为我们的主要动机之一正是利用差异来评估近似采样算法的质量，这再次导致了先有鸡还是先有蛋的问题。不过，理解最大均值差异所具有的吸引力十足的理论性质仍然很有启发性，因为我们稍后将寻求构造具有类似吸引力的 Stein差异。

MMD 也称为最坏情形求积误差，这是由于其对偶解释为

$$\mathrm{MMD}_k(\mu,\nu) = \sup_{\|f\|_{\mathcal{H}_k}\leq 1} \mu(f) - \nu(f). \tag{2.11}$$

从这一角度来看，很明显，核 k 通过决定$\mathcal{H}_k$单位球中所包含函数的正则性，决定了由 MMD 诱导的拓扑。特别地，如果 $k(x,\ y) = \varphi(x-y)$ 对于一个利普希茨函数 φ（其 Lipschitz常数为 L_φ）成立，那么$\{f\colon \|f\|_{\mathcal{H}_k} \leq 1\} \subset \{f\colon |f|_{\mathrm{Lip}} \leq L_\varphi\}$，在这种情况下，MMD 的拓扑弱于或等于 1-Wasserstein距离的拓扑，因为 MMD$_k(\mu,\ \nu) \leq L_\varphi \mathrm{W}^1(\mu,\ \nu)$。一个能够复现足够大的希尔伯特空间的核，直观而言，将包含足够多的检验函数，从而能够区分不同的概率测度：

定义 2.30（特征核）。 如果一个核 k 满足：对于所有 μ、$\nu \in P^{\mathcal{H}_k}$，特征条件 MMD_k（$\mu,\ \nu$）$=0$ 蕴含 $\mu = \nu$，则称其为特征核。

例 *2.19*（多项式核不是特征核）。由例 *2.15* 可知，核 $k(x,y) = 1 + (xy) + \cdots + (xy)^p$ 所再生的希尔伯特空间，其元素是定义域 $X = \mathbb{R}$ 上次数至多为 p 的多项式。因此，当且仅当矩 $\int x^i\, d\mu(x)$ 和 $\int x^i\, d\nu(x)$ 对于 $i = 1, \ldots, p$ 都相同时，最大均值差异 $MMD_k(\mu,\nu) = 0$。特别地，k 不是特征核。

例 2.20。 高斯核 $k(x,y) = \exp(-\|x-y\|^2/(2\ell^2))$ 对于所有 $\ell > 0$，都是定义域 $X = \mathbb{R}^d$ 上的特征核。

特征性质并未说明最大均值差异 MMD$_k(\mu,\nu)$ 取值“较小”时的情况，只说明了最大均值差异 MMD$_k(\mu,\nu)$ 恰好为 0 的情形。因此，单凭特征性并不能充分证明使用核 k 来衡量 μ 与 ν 之间差异的合理性。为此，我们现在引入一种更强的性质，在此语境下（颇具讽刺意味地）称为 弱收敛控制。

定义 2.31（弱收敛控制）。核 k 称为具有弱收敛控制，如果最大均值差异$_k(\mu_n,\mu) \to 0$ 蕴含 $\mu_n \xrightarrow{d} \mu$。

收敛控制证明了将最大均值差异用作量化及更一般的分布近似的优化准则是合理的。有关最大均值差异常见应用的概述，请参见 Muandet 等 [2017]。

备注 2.1。 或许令人惊讶的是，对于紧豪斯多夫空间 $\mathcal{X}$，有界且可测的特征核 k 保证具有弱收敛控制。当定义域 $\mathcal{X}$ 非紧时，这种等价关系不再成立，有界特征核可能不具备弱收敛控制；参见 *Simon-Gabriel 等 [2023]*。显然，非特征核不具备弱收敛控制。

例 2.21。 高斯核 $k(x, y) = \exp(-\|x - y\|^2/(2\ell^2))$ 控制了 $\mathcal{X} = [0, 1]^d$ 上概率分布的弱收敛。还可以证明，高斯核控制了 $\mathcal{X} = \mathbb{R}^d$ 上的弱收敛；这可由 *Simon-Gabriel 等 [2023]* 的定理 7 以及 *Sriperumbudur 等 [2011]*中的一般结果推得。

3 Stein 算子

本章介绍 Stein 算子的概念。通俗地说，Stein 算子是一种映射，它生成的函数在目标概率测度下的积分为零P。Stein 算子是我们将在第 4 章中用于构造 Stein差异的关键要素。本章将大量使用简写符号 $P(f) = \int f \mathrm{d}P$，该符号在 (2.1) 中引入。

定义 3.1 (Stein 算子与 Stein 集合)。 设 P 是可测空间 X 上的一个概率分布。我们称集合 G 为一个Stein集合，称线性映射 $T_P: G \to L^1(P)$ 为关于 P 的一个Stein 算子，如果

$$P(\mathcal{T}_P g) = 0 \quad for\ all \quad g \in \mathcal{G}. \tag{3.1}$$

积分 $P(\mathcal{T}_P g)$ 在定义 3.1 中出现，可以看作随机变量 $(\mathcal{T}_P g)(X)$ 的期望，其中 $X \sim P$，从这个意义上说，当 $\mathcal{T}_P$满足以下条件时，它是一个 Stein 算子：对 Stein 集合 g G 中的所有元素，$(\mathcal{T}_P g)(X)$ 的均值为零。

Stein 算子的构造方法有很多，但从实践和理论角度来看，并非所有 Stein 算子都相同。例如，在一个极端情形下，我们有恒为零的 Stein 算子

$$(\mathcal{T}_P g)(x) = 0, \tag{3.2}$$

它易于计算，但完全无法提供有关目标分布 P 的信息。与之相反的是均值重心化 Stein 算子

$$(\mathcal{T}_P g)(x) = g(x) - P(g) \tag{3.3}$$

其 Stein 集合为 $\mathcal{G} = \mathcal{L}^1(P)$。从理论分析的角度来看，这个算子很理想，但只有在 P 下的积分易于处理时才可计算。遗憾的是，对于我们感兴趣的大多数应用而言，在 P 下进行精确积分显然难以处理。事实上，对于许多应用，我们的最终目标是近似 P 下难以处理的积分，或评估这些积分的近似结果。

值得注意的是，有些可计算的 Stein 算子无需在 P下进行显式积分，就能得到有用的差异度量。例如，Stein [1986] 证明了，如果 P 是定义在 $\mathbb{R}$ 上的单变量概率分布，且其概率密度函数为正且可微 p （见定义2.5），那么对于适当的 Stein 集合$\mathcal{G}$中的所有可微函数，

$$(\mathcal{T}_P g)(x) = g'(x) + g(x) \frac{\mathrm{d}}{\mathrm{d}x} \log p(x) \tag{3.4}$$

是 P 的 Stein 算子。事实上，通过分部积分

$$\begin{aligned} P(\mathcal{T}_P g) = \int \mathcal{T}_P g \, \mathrm{d}P &= \int \left[g'(x) + g(x) \frac{\mathrm{d}}{\mathrm{d}x} \log p(x) \right] p(x) \, \mathrm{d}x \\ &= \int \left[g'(x) + \frac{g(x)p'(x)}{p(x)} \right] p(x) \, \mathrm{d}x \\ &= \int g'(x)p(x) + g(x)p'(x) \, \mathrm{d}x = \int (gp)'(x) \, \mathrm{d}x \end{aligned} \tag{3.5}$$

根据微积分基本定理，只要当 $|x| \to \infty$ 趋于无穷时， $g(x)p(x) \to$ 趋于 0，该值就等于零。

算子 (3.4) 称为 Stein 密度方法算子，因为它将对 P (3.3) 的显式积分替换为对数密度导数 $\mathrm{d}\,\mathrm{d}_x \log p(x)$ 的求值。重要的是，即使对 p 的精确积分难以处理，这个对数密度导数通常仍可计算。例如，若 P 是贝叶斯统计分析中得到的后验分布（参见第 1.2 节），那么根据贝叶斯定理， $p(x) \propto \pi(x)\mathrm{L}(x)$，其中 $\pi(x)$ 是先验分布的密度，L 是似然。这两个量通常都有明确表达式，足以用于密度方法，因为

$$\frac{\mathrm{d}}{\mathrm{d}x} \log p(x) = \frac{\mathrm{d}}{\mathrm{d}x} \log \pi(x) + \frac{\mathrm{d}}{\mathrm{d}x} \log \mathcal{L}(x),$$

原则上，只要先验和似然都可以分别求导，就能够计算出来。事实上，无需知道难以处理的边际似然，也能计算对数后验密度的梯度；这一事实是贝叶斯统计背景下所用的多种近似采样算法的基础 [Brooks 等人， 2011]。

本章将探讨适用于多种目标分布、同样具有实用性的 Stein 算子 P。我们从$\mathbb{R}^d$上分布的一个典范 Stein 算子——Langevin Stein 算子——入手，并在第3.1节介绍。Langevin Stein 算子是更一般构造的一种特例，这种构造称为扩散 Stein 算子，我们将在第3.2节介绍。我们经常会遇到定义域为某个子集$\mathcal{X} \subset \mathbb{R}^d$的分布，此时设计 Stein 算子需要特别考虑；针对这种情况，我们将在第3.3节介绍镜像斯坦算子。另一方面，我们可能希望避免计算目标分布密度的梯度P；针对这种情况，人们开发了无梯度斯坦算子，相关讨论见第3.4节。最后，我们将讨论离散状态空间$\mathcal{X}$的情形；针对这种情形，人们开发了若干不同的 Stein 算子，详见第3.5节。

3.1 Langevin 算子

Stein 密度方法算子（3.4）可应用于一元分布 P，但我们希望为贝叶斯统计和概率推断中常见的多元分布开发类似的算子

其中 P 的支撑集为$\mathcal{X} = \mathbb{R}^d$。为此，我们将首先探讨 Barbour 提出的一个精妙想法 [1988, 1990],Gotze [1991], ，如今称为生成元方法。Barbour 和 Götze 注意到，如果能确定一个（时间齐次）马尔可夫过程，使得 P 是不变的（定义2.14），那么在温和条件下，该马尔可夫过程的无穷小生成元（定义2.15）就是 P的 Stein 算子。直观而言，这很合理：如果我们将马尔可夫过程初始化为从 P中抽取的状态，那么只要该过程具有 P 不变性，后续状态的分布也将是 P。马尔可夫过程的生成元描述了检验函数期望值的变化率；对于从不变分布初始化的过程，这一变化率为零。Gorham 和 Mackey [2015] 将这一想法应用于一种特定的马尔可夫过程——例2.5中介绍的过阻尼朗之万扩散。

假设 3.1 （ P的正则性）。 设 P 是定义在 $\mathcal{X} = \mathbb{R}^d$ 上、具有正且可微密度 p 的概率分布。

在假设 3.1下，过阻尼朗之万扩散的无穷小生成元是微分算子 $u \mapsto (\Delta u) + (\nabla u) \cdot (\nabla \log p)$；参见例 2.6。将 ∇u 替换为更一般的向量值函数 g，便得到如今称为 Langevin Stein 算子的算子：

定义 3.2 （Langevin Stein 算子；Gorham 和 Mackey，2015年）。 在假设 *3.1*的设定下，Langevin Stein 算子 对于 P 满足

$$(\mathcal{T}_P g)(x) = (\nabla \cdot g)(x) + g(x) \cdot (\nabla \log p)(x) \tag{3.6}$$

对于适当正则的向量值函数 g： $\mathbb{R}^d \to \mathbb{R}^d$ 以及所有 $x \in \mathbb{R}^d$。

当$d = 1$时，Langevin Stein 算子恰好还原密度方法算子（3.4）。现在，我们更精确地说明定义 3.2中出现的向量值函数 g 所需满足的正则性条件。回顾第 2.1.2 节，对于向量值函数 g，我们用简写 $g \in \mathcal{L}^1(P)$ 表示 g 的所有分量都属于 $\mathcal{L}^1(P)$。接下来，我们需要一个适用于可能无界定义域的散度定理。为此，我们先完整给出散度定理的几个有用版本及其证明。这样做是为了强调，散度的连续性——大多数教材在论述散度定理时通常都会假设这一条件——并非散度定理结论成立所必需的条件：

定理 3.1 （ $[-1,1]^d$上的散度定理）。假设向量值函数 $v\colon \mathcal{X} \mapsto \mathbb{R}^d$的散度 $\nabla \cdot v$ 在 $\mathcal{X} \triangleq [-1,1]^d$ 上处处存在。如果 $\nabla \cdot v \in L^1(\mathcal{X})$ ，则

$$\int_{\mathcal{X}} (\nabla \cdot v)(x)\, \mathrm{d}x = \oint_{\partial \mathcal{X}} v(x) \cdot \mathrm{n}(x)\, \mathrm{d}x$$

其中，$\mathrm{n}(x)$ 是 $x \in \partial\mathcal{X}$ 处的外向单位法向量。

证明。对于每个维度 $j \in [d]$ 和 $x \in \mathbf{X}$，令 x_{-j} 表示从 x 中移除第 j 个坐标后得到的子向量。由于 $\nabla \cdot v \in \mathrm{L}^1(\mathbf{X})$，根据 Fubini 定理 [Rudin, 1987, 定理 8.8]，可知对于每个 $j \in [d]$，几乎处处的 $x_{-j} \in [-1,1]^{d-1}$ 都有 $x_j \mapsto \nabla_j v_j(x) \in \mathrm{L}^1([-1,1])$。此外，由于 $\nabla \cdot v$ 在 $\mathbf{X}$ 上处处存在，根据微积分基本定理 [Rudin, 1987, 定理 7.21]，可得

$$v_j(x)\Big|_{x_j=1} - v_j(x)\Big|_{x_j=-1} = \int_{-1}^{1} \nabla_j v_j(x)\ \mathrm{d}x_j$$

对每个 $j \in [d]$，几乎处处的 $x_{-j} \in [-1,1]^{d-1}$ 均成立。因此我们得到

$$\begin{aligned}
\oint_{\partial \mathcal{X}} v(x) \cdot \mathrm{n}(x)\ \mathrm{d}x &= \sum_{j=1}^{d} \int_{[-1,1]^{d-1}} v_j(x)\Big|_{x_j=1} - v_j(x)\Big|_{x_j=-1}\ \mathrm{d}x_{-j} \\
&= \sum_{j=1}^{d} \int_{[-1,1]^{d-1}} \int_{-1}^{1} \nabla_j v_j(x)\ \mathrm{d}x_j\ \mathrm{d}x_{-j} \\
&= \int_{\mathcal{X}} (\nabla \cdot v)(x)\ \mathrm{d}x,
\end{aligned}$$

其中最后一步再次用到了 Fubini 定理。 □

定理 3.2 （$\mathbb{R}^d$上的散度定理）。 设向量值函数 $v\colon \mathbb{R}^d \to \mathbb{R}^d$ 的散度 $\nabla \cdot v$ 在 $\mathbb{R}^d$ 上处处存在。若 $\nabla \cdot v \in \boldsymbol{L}^1(\mathbb{R}^d)$ 则

$$\int (\nabla \cdot v)(x)\ \mathrm{d}x = \lim_{r \to \infty} \oint_{\partial \mathcal{X}_r} v(x) \cdot \mathrm{n}_r(x)\ \mathrm{d}x$$

其中，对每个 $r \in \mathbb{R}$，$\boldsymbol{X}_r \triangleq [-r,r]^d$，$\mathrm{n}_r(x)$ 是 $\boldsymbol{X}_r$ 上 $x \in \partial$ 处的外向单位法向量。此外，若 $v \in \boldsymbol{L}^1(\mathbb{R}^d)$，则

$$\int (\nabla \cdot v)(x)\ \mathrm{d}x = 0.$$

证明。 由于 $\nabla \cdot v \in \mathcal{L}^1(\mathbb{R}^d)$ 在 $\mathbb{R}^d$ 上处处存在，因此第一个结论由 Lebesgue 控制收敛定理 [Rudin, 1987, 定理 1 得出。34] 以及散度定理（定理 3.1）：

$$\begin{aligned}
\int (\nabla \cdot v)(x)\ \mathrm{d}x &= \int \lim_{r \to \infty} \mathbb{I}\left[x \in \mathcal{X}_r\right] (\nabla \cdot v)(x)\ \mathrm{d}x \\
&= \lim_{r \to \infty} \int_{\mathcal{X}_r} (\nabla \cdot v)(x)\ \mathrm{d}x = \lim_{r \to \infty} \oint_{\partial \mathcal{X}_r} v(x) \cdot \mathrm{n}_r(x)\ \mathrm{d}x.
\end{aligned}$$

现在再假设 $v \in \mathcal{L}^1(\mathbb{R}^d)$ 也成立，并且对每个 $j \in [d]$ 和 $r \in \mathbb{R}$，定义函数

$$f_j(r) \triangleq \int |v_j(x)|\Big|_{x_j=r}\ \mathrm{d}x_{-j} \geq \int_{\|x_{-j}\|_\infty \leq |r|} |v_j(x)|\Big|_{x_j=r}\ \mathrm{d}x_{-j}$$

其中 x_{-j} 是通过去掉 x 的第 j 个坐标得到的子向量。此定义给出了以下界

$$\begin{aligned}\left|\int (\nabla \cdot v)(x)\ \mathrm{d}x\right| &= \lim_{r\to\infty}\left|\oint_{\partial\mathcal{X}_r} v(x)\cdot \mathrm{n}_r(x)\ \mathrm{d}x\right| \\ &= \liminf_{r\to\infty}\left|\oint_{\partial\mathcal{X}_r} v(x)\cdot \mathrm{n}_r(x)\ \mathrm{d}x\right| \\ &= \liminf_{r\to\infty}\left|\sum_{j=1}^{d}\int_{\|x_{-j}\|_\infty\le r} v_j(x)\Big|_{x_j=r} - v_j(x)\Big|_{x_j=-r}\ \mathrm{d}x_{-j}\right| \\ &\le \liminf_{r\to\infty}\sum_{j=1}^{d} f_j(r)+f_j(-r).\end{aligned}$$

最后，我们将通过反证法证明该不等式右侧为零。

为此，任取 $j\in[d]$ 和 $\epsilon>0$，并假设 $\liminf_{r\to\infty} f_j(r)+f_j(-r)\ge\epsilon$。那么，存在一个 $r'>0$，使得 $\inf_{r\ge r'} f_j(r)+f_j(r)\ge\epsilon/2$，从而 $\int_0^\infty f_j(r)+f_j(-r)\,\mathrm{d}r \ge \int_{r'}^\infty f_j(r)+f_j(-r)\,\mathrm{d}r \ge \int_{r'}^\infty \epsilon/2\,\mathrm{d}r=\infty$。这与 Fubini 定理 [Rudin，1987，定理 8.8] 以及 v 的可积性假设相矛盾，因为它们意味着

$$\int_0^\infty f_j(r)+f_j(-r)\ \mathrm{d}r = \int |v_j(x)|\ \mathrm{d}x < \infty.$$

因此，必有 $\sum_{j=1}^d \liminf_{r\to\infty} f_j(r)+f_j(-r)=0$，进而 $\left|\int(\nabla\cdot v)(x)\,\mathrm{d}x\right|=0$，正如所述。□

定理 3.3（凸集上的散度定理）。设向量值函数$v: X\to\mathbb{R}^d$在凸集$X\subseteq\mathbb{R}^d$上可微。若 $\nabla\cdot v\in L^1(X)$ 则

$$\int_{\mathcal{X}} (\nabla\cdot v)(x)\ \mathrm{d}x = \lim_{r\to\infty}\oint_{\partial\mathcal{X}_r} v(x)\cdot \mathrm{n}_r(x)\ \mathrm{d}x$$

其中，对于每个 $r\in\mathbb{R}$，$X_r \triangleq X\cap[-r,r]^d$，$\mathrm{n}_r(x)$ 是$x\in\partial X_r$处的外向单位法向量。

证明。由于 $\mathcal{X}$ 是凸的，每个 $\mathcal{X}_r$ 有界且凸，因此具有有限周长[Buttazzo 等人，1995，引理 2。4]此外，由于 v 在 $\mathcal{X}$ 上可微，该结果由 Lebesgue 控制收敛定理得出 [Rudin，1987，定理 1。34] 以及Pfeffer [2012, 的 Gauss-Green 定理（定理 5.19 与命题 5）。8]:

$$\begin{aligned}\int_{\mathcal{X}} (\nabla\cdot v)(x)\ \mathrm{d}x &= \int \lim_{r\to\infty}\mathbb{I}\left[x\in\mathcal{X}_r\right](\nabla\cdot v)(x)\ \mathrm{d}x \\ &= \lim_{r\to\infty}\int_{\mathcal{X}_r}(\nabla\cdot v)(x)\ \mathrm{d}x = \lim_{r\to\infty}\oint_{\partial\mathcal{X}_r} v(x)\cdot \mathrm{n}_r(x)\ \mathrm{d}x.\end{aligned}$$

□

命题 3.1（朗之万算子的定义域）。 在假设 *3.1* 的设定下，设 $g\colon \mathbb{R}^d \to \mathbb{R}^d$ 可微。以下结论对 $\mathcal{T}_P$ 成立，其定义见 （3.6）。

1. 若 $g \cdot \nabla \log p, \nabla \cdot g \in L^1(P)$，则 $T_P g \in L^1(P)$。
2. 若 $g, T_P g \in L^1(P)$，则 $P(T_P g) = 0$。

证明。 第一个结论可由（3.6）直接得到。对于第二个结论，

$$\int (\mathcal{T}_P g)(x)\, \mathrm{d}P(x) = \int \frac{1}{p(x)} (\nabla \cdot (pg))(x)\, \mathrm{d}P(x) = \int (\nabla \cdot (pg))(x)\, \mathrm{d}x.$$

在我们的假设下，向量场 $v = pg$ 满足 $v \in \mathcal{L}^1(\mathbb{R}^d)$ 且 $\nabla \cdot v \in \mathcal{L}^1(\mathbb{R}^d)$，因此我们可以在 $\mathbb{R}^d$ （定理 3.2）上应用散度定理来证明该结论。 □

人们已为支撑于光滑流形上的分布构造了类似的 Stein 算子 [Le等人，2024，Barp 等人， 2022], 他们可以构造朗之万扩散的适当推广。这需要对散度定理作流形上的推广；关于所涉及数学内容的技术讨论，请参见 Pigola and Setti [2014]。

为了在第 5.5 节进一步讨论 Stein 变分梯度，我们还将引入一个算子，它作用于标量值函数 h并生成向量值函数 $\mathcal{A}_P h$，使得 $\mathcal{A}_P h$ 的每个分量在 P下的期望均为零。

定义 3.3 (向量值 Langevin Stein 算子；Liu 等人，2016）。 在假设 *3.1*的条件下，向量值朗之万Stein算子对于 P 满足

$$(\mathcal{A}_P h)(x) = (\nabla h)(x) + h(x) \cdot (\nabla \log p)(x) \tag{3.7}$$

对于适当正则的标量值函数 $h\colon \mathbb{R}^d \to \mathbb{R}$ 以及所有 $x \in \mathbb{R}^d$。

我们可以用与推导其标量对应算子时大致相同的方法，为向量值朗之万算子推导出一个合适的定义域。

命题 3.2 (向量值朗之万算子的定义域)。 在假设 *3.1*的条件下，设 $h\colon \mathbb{R}^d \to \mathbb{R}$ 可微。对于$\mathcal{A}_P$ （定义见(3.7)），以下结论成立。

1. 若 $h \cdot \nabla \log p, \nabla h \in L^1(P)$，则 $A_P h \in L^1(P)$。
2. 若 $h, A_P h \in L^1(P)$，则 $P(A_P h) = 0$。

证明。第一个结论直接由定义 (3.7) 得出。我们将通过归结为定义 3.2 中的标量朗之万算子 T_P 的情形来推导第二个结果。固定任意坐标 $j \in \{1, \ldots, d\}$，并定义向量值函数 $g(x) = e_j h(x)$，其中 e_j是 $\mathbb{R}^d$ 中的第 j 个标准基向量。于是 $(\mathrm{A}_P h)_j = \mathrm{T}_P g$，且 g 满足命题 3.1 的前提条件，因此 $P((\mathrm{A}_P h)_j) = P(\mathrm{T}_P g) = 0$。 □

3.2 扩散算子

在上一节中，我们看到某个特定马尔可夫过程的生成元可用于为目标分布构造 Stein 算子 P，前提是该马尔可夫过程是 P-不变的。然而，满足这一要求的马尔可夫过程有无穷多个！根据我们需要 Stein 差异的具体情境，没有理由认为 Langevin Stein 算子的选择是最优的。幸运的是，Gorham 等人 [2019] 证明可以将第3.1节中的构造推广到一类被称为 *Itô*扩散。

定义 3.4 （Itô扩散）。具有起始点$x \in \mathbb{R}^d$ 且漂移系数: $\mathbb{R}^d \to \mathbb{R}^d$ 和扩散系数: $\mathbb{R}^d \to \mathbb{R}^{d\times m}$满足局部Lipschitz条件和线性增长条件的（时间齐次）Itô扩散 是一个随机过程$(Z_{t,x})_{t\geq 0}$，它满足 *Itô*随机微分方程

$$\mathrm{d}Z_{t,x} = b(Z_{t,x})\,\mathrm{d}t + \sigma(Z_{t,x})\,\mathrm{d}W_t \quad with \quad Z_{0,x} = x, \tag{3.8}$$

其中$(W_t)_{t\geq 0}$ 表示 $\mathbb{R}^d$上的标准 d维 *Wiener*过程。

为了像利用朗之万扩散那样利用 Itô扩散，我们需要对漂移系数 b 和扩散系数 σ 施加约束，以确保过程 P保持不变。下面的结果见 Gorham 等人 [2019, 定理 2]，其内容源自 Ma 等人 [2015, 定理 2] 和 Pavliotis [2016, 第 4.6]，节，它完整刻画了所有使 P 保持不变的 Itô扩散。回顾一下，对于矩阵值函数 $M(x)$，记号 $\nabla\cdot M$表示一个向量，其第 j个分量为 $\nabla\cdot(M^\top e_j)$，其中e_j 是 $\mathbb{R}^d$的第 j个基向量。

定理 *3.4*（ P-不变扩散的完整方案）。在假设 *3.1* 的条件下，考虑漂移系数为 $b \in C^1(\mathbb{R}^d, \mathbb{R}^d)$、扩散系数为$\sigma \in C^1(\mathbb{R}^d, \mathbb{R}^{d\times m})$ 的 *Itô*扩散（*3.8*），并定义协方差系数 $a(x) \triangleq \sigma(x)\sigma(x)^\top$。当且仅当满足以下条件时，该扩散是$P$-不变的：

$$b(x) = \frac{1}{2}\frac{1}{p(x)}(\nabla\cdot(pa))(x) + f(x) \tag{3.9}$$

对于一个非可逆分量 $f \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ 满足 $\nabla\cdot(pf) = 0$。此外，若 b满足(3.9) 且 $f \in \mathcal{L}^1(P)$，则[2]

$$b(x) = \frac{1}{2}\frac{1}{p(x)}(\nabla\cdot(p(a+c)))(x) \tag{3.10}$$

[1] b 和 σ 满足局部Lipschitz及线性增长假设，确保随机微分方程（3.8）存在唯一的[Ikeda和Watanabe，2014年，第 IV 章，定理 3。1] 且不会爆炸的 [Oksendal，2013，定理 5.2。1] 解。[2]为避免歧义，特此强调，下文中的 $p(a+c)$ 表示函数 $x \mapsto p(x)(a(x)+c(x))$，而不是“将 p 应用于 $a+c$”。

对于一个可微的 P-可积反对称 $d \times d$ 矩阵值函数 c，称为流系数 *[Conca 和 Vanninathan, 2007年，Landim 等人， 1998]*。在这种情况下，扩散的无穷小生成元具有如下形式

$$u \mapsto \frac{1}{2p}\nabla \cdot (p(a+c)(\nabla u))$$

对所有适当正则的 $u \in C^2(\mathbb{R}^d, \mathbb{R})$。

这些观察促使Gorham 等人 [2019] 提出 Langevin Stein 算子的如下推广：

定义 *3.5*（扩散 *Stein* 算子；*Gorham* 等人，*2019*，*(8)*）。在假设 *3.1* 的设定下，设 $a(x) \in \mathbb{R}^{d\times d}$ 为对称半正定，且 $c(x)$ 为斜对称，并且 $x \mapsto p(x)(a(x) + c(x))$ 在 $x \in \mathbb{R}^d$ 上可微。则 P 的 (a, c)-扩散 *Stein* 算子满足

$$(\mathcal{T}_P g)(x) = \frac{1}{p(x)}(\nabla \cdot (p(a+c)g))(x) \tag{3.11}$$

对所有适当正则的向量值函数 $g : \mathbb{R}^d \to \mathbb{R}^d$ 以及所有 $x \in \mathbb{R}^d$。

对协方差系数 a 与流系数 c 在（3.11）中的选择可以依赖于具体应用，从而赋予基于扩散 Stein 算子的 Stein差异一系列不同性质。若干重要特例由

1. 预条件朗之万扩散，其中 $c \equiv 0$ 且 $a = \sigma\sigma^\top$ 扩散系数为常数 $\sigma \in \mathbb{R}^{d\times m}$ [Stuart等人, 2004];

2. 黎曼 *Langevin* 扩散，其中 $c \equiv 0$ 且 $a(x) = G^{-1}(x)$，其中 G 是一个正定度量张量 [Patterson和Teh, 2013, Xifara等人, 2014, Ma等人, 2015];

3. 非可逆预条件朗之万扩散，其中 $c \not\equiv 0$ 且 $a = \sigma\sigma^\top$ 对于常数 $\sigma \in \mathbb{R}^{d\times m}$ [Ma等人, 2015, Duncan等人, 2016, Rey-Bellet和Spiliopoulos, 2015]；以及

4. 欠阻尼朗之万扩散 [Horowitz, 1987], 其目标分布为增广分布 $P \times \mathcal{N}(0, I)$ 在 $\mathbb{R}^{2d}$ 上，且

$$c \equiv 2\begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix} \quad \text{and} \quad a \equiv 2\begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix}.$$

进一步讨论推迟至第4.4.2节。

下面的结果是对Gorham 等人 [2019, 命题 3], 的推广，它阐明了我们对向量值函数 g 所要求的最低正则性： $\mathbb{R}^d \to \mathbb{R}^d$回顾

对于 $g\colon \mathbb{R}^d \to \mathbb{R}^d$，记号 $\nabla g\colon \mathbb{R}^d \to \mathbb{R}^{d\times d}$ 表示 $[\nabla g]_{i,\,j} = \partial_j g_i$；而对于矩阵 A 和 B，双点积为 $A : B = \mathrm{tr}(AB^\top)$。特别地，对于矩阵值函数 $M(x)$，乘积法则写作 $\nabla \cdot (Mg) = (\nabla \cdot M) \cdot g + M : \nabla g$。

命题 3.3（扩散算子的定义域）。在假设 3.1下，设$\mathcal{T}_P$ 为 $(a,\ c)$-扩散 *Stein* 算子，见（3.11）；设 b 定义如（3.10）所示；并设$g\colon \mathbb{R}^d \to \mathbb{R}^d$ 可微。则以下结论成立

1. 若 $b \cdot g,\ (a+c)\colon \nabla g \in L^1(P)$，则 $\mathcal{T}_P g \in L^1(P)$。
2. 若 $(a+c)\,g$，则 $\mathcal{T}_P g \in L^1(P)$，且 $P(\mathcal{T}_P g) = 0$。

证明。证明与命题 3.2的证明类似。对于第一个结论，由于 g 可微，且 $b \cdot g,\ (a+c)\colon \nabla g \in \mathcal{L}^1(P)$，因此立即可得

$$\mathcal{T}_P g = 2b \cdot g + (a+c) : \nabla g \in \mathcal{L}^1(P).$$

对于第二个结论，

$$\int (\mathcal{T}_P g)(x)\,\mathrm{d}P(x) = \int (\nabla \cdot (p(a+c)g))(x)\,\mathrm{d}x.$$

在我们的假设下，向量场 $v = p(a+c)\,g$满足 $v \in \mathcal{L}^1(\mathbb{R}^d)$，且 $\nabla \cdot v \in \mathcal{L}^1(\mathbb{R}^d)$，因此我们可以再次应用定理 3.2 来证明该结论。

□

最后，我们指出，由于只有复合 $\mathcal{T}_P g$（由 Stein 算子与函数g 构成）出现在Stein方法中，因此可以灵活地同时修改 Stein 算子 $\mathcal{T}_P$ 与函数 g，同时保持复合 $\mathcal{T}_P g$ 不变。由此，可以将扩散 Stein 算子视为等价于把 Langevin Stein 算子（定义 3.2）应用于形如 $(a+c)\,g$ 的函数；这一视角使基于 Langevin Stein 算子的 Stein 差异的相关结果（我们在章4中给出）能够转移到基于扩散 Stein 算子的 Stein 差异。

3.3 约束算子

到目前为止，我们只针对支撑集位于 $\mathbb{R}^d$上的分布介绍了朗之万 Stein 算子和扩散 Stein 算子，但我们往往希望考虑支撑集位于子集$\mathcal{X} \subset \mathbb{R}^d$上的分布。遗憾的是，取一个$\mathcal{T}_P$ 算子（它是为定义在 $\mathbb{R}^d$ 上的分布而设计的），再将其限制到$\mathcal{X}$上并不可行；定义 $P(\mathcal{T}_P g) = 0$3.1中的条件将不再成立，因为边界项

$$\oint_{\partial\mathcal{X}} p(x)g(x) \cdot \mathrm{n}(x)\,\mathrm{d}x$$

在散度定理中出现的边界项通常不会消失。回想一下， $\partial\mathcal{X}$ 表示集合$\mathcal{X}$的边界；根据定义，当$\mathcal{X} = \mathbb{R}^d$时， $\partial\mathcal{X}$ 为空。接下来有一种选择。

一种做法是仅考虑满足以下条件的函数 g ：通过构造使边界项为零；这可以通过要求满足边界处消失条件来实现 [Gorham 和 Mackey， 2015]

$$g(x) \cdot \mathrm{n}(x) = 0 \quad \text{for all} \quad x \in \partial\mathcal{X}. \tag{3.12}$$

这种构造与第4.2节定义的经典 Stein 差异、第4.3节定义的图 Stein 差异，以及第4.4节定义的核 Stein 差异相兼容。

第二种做法是采用针对$\mathcal{X}$调整的 Stein 算子。镜像 *Langevin* 扩散[Zhang 等人，2020年，Chewi 等人， 2020]，

$$Z_{t,x} = \nabla\psi^*(\eta_{t,x}), \quad \mathrm{d}\eta_{t,x} = \nabla \log p(Z_{t,x})\mathrm{d}t + \sqrt{2}\nabla^2\psi(Z_{t,x})^{1/2}\mathrm{d}B_t$$
$$\text{for} \quad \eta_{0,x} = \nabla\psi(x), \tag{3.13}$$

提供了一个优雅的 P-不变马尔可夫过程示例，可用于在定义域 $\mathcal{X}$ 为凸且闭集的情况下构造 Stein 算子。这里使用镜像映射$\nabla\psi$ 将输入变量变换为镜像变量 $\eta_{0,\ x}$镜像变量$\eta_{t,\ x}$ 按照随机微分方程（3.13）演化，最后 $\eta_{t,\ x}$ 通过逆镜像映射 $\nabla\psi^*$映射回原始约束样本空间 $\mathcal{X}$ 。

需要明确的是，我们并未将讨论范围限制在有界定义域 $\mathcal{X}$ 内。一个无界凸闭集$\mathcal{X}$的简单例子是$\{x \in \mathbb{R}^d\colon\ x_1, \ldots, x_d \geq 0\}$。对于边界 $\partial\mathcal{X}$ 可在局部表示为 $F(x) = 0$ 的闭凸集，其外单位法向量满足

$$n(x) \in \left\{ \pm \frac{\nabla F(x)}{\|\nabla F(x)\|_2} \right\}.$$

Zhang 等人 [2020] 指出，镜像扩散是度量张量为 $G = \nabla^2\psi$的黎曼 Langevin 扩散的一个实例，其中函数 $\psi\colon\ \mathcal{X} \to \mathbb{R} \cup \{\infty\}$需满足

1. $(\nabla^2\psi)^{-1}$ 在$\mathcal{X}$上满足利普希茨条件且可微，2. $\|(\nabla\psi)(x_n)\|_2 \to \infty$ 对任意 $x_n \to x \in \partial\mathcal{X}$均成立，3. 存在某个 $c > 0$, $\nabla^2\psi(x) \succcurlyeq cI$ ，使得对所有 $x \in \mathcal{X}$.

假设 3.2 （ P的正则性）。 设 P 是凸闭集 $X \subset \mathbb{R}^d$ 上的一个概率分布，其具有有限均值，且在 X 上具有正可微密度 p 。

对镜像 Langevin 扩散的无穷小生成元的考察促使Shi 等人[2022提出]了以下 Stein 算子：

定义 3.6（镜像 Stein 算子；Shi 等人，2022a）。 考虑一个凸闭集 $\mathcal{X} \subset \mathbb{R}^d$，并令 ψ 为具有上述性质的函数。在假设 3.2的设定下，镜像 Stein 算子对 P 满足

$$(\mathcal{T}_{P,\psi} g)(x) = g(x) \cdot (\nabla^2 \psi)(x)^{-1} (\nabla \log p)(x) + \nabla \cdot ((\nabla^2 \psi)(x)^{-1} g(x)), \tag{3.14}$$

对于足够正则的向量值函数 $g: X \to \mathbb{R}^d$ 以及所有 $x \in X$。

由于镜像 Stein 算子是扩散 Stein 算子的特例，因此可以通过类似于命题 3.3[的论证，推导出镜像 Stein 算子的定义域，参见 Shi 等人，2022a，命题 1]。这里唯一真正的区别在于，必须注意确保 $(\nabla^2 \psi)^{-1}$ 能够抵消 p 在边界 $\mathcal{X}$上的增长。

回顾一下，对于矩阵值函数 $M(x)$，记号 $\nabla \cdot M$ 表示一个向量，其第 j 个分量为 $\nabla \cdot (M^\top e_j)$，其中 e_j 是 $\mathbb{R}^d$的第 j 个基向量。

命题 3.4（镜像算子的定义域）。 在假设 3.2的条件下，设 $g: \mathcal{X} \to \mathbb{R}^d$ 可微，且 $\psi: \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ 足够正则，使得以下各量均存在。则有以下结论：

1. *If* $g \cdot (\nabla^2 \psi)^{-1} (\nabla \log p),\ \nabla \cdot ((\nabla^2 \psi)^{-1} g) \in \mathcal{L}^1(P)$ *then* $\mathcal{T}_{P,\psi} g \in \mathcal{L}^1(P)$.
2. 若 $\mathcal{T}_{P,\psi} g \in L^1(P)$，$v \triangleq p(\nabla^2 \psi)^{-1} g$ 可微，且

$$\lim_{n \to \infty} \oint_{\partial \mathcal{X}_n} v(x) \cdot \mathrm{n}(x)\, \mathrm{d}x = 0 \tag{3.15}$$

其中 $X_n \triangleq \{x \in X: \ \|x\|_\infty \le n\}$，则 $P(\mathcal{T}_P g) = 0$。

证明。 第一个结论由式（3.14）立即得出。第二个结论由凸集的散度定理（定理 3.3）和式（3.15）得出，因为 $\nabla \cdot v \in \mathcal{L}(\mathcal{X})$。 □

举例来说，考虑概率单纯形 $\mathcal{X} = \{x \in [0,1]^d: \ \sum_{i=1}^d x_i \le 1\}$，例如，我们可能感兴趣的是狄利克雷目标分布 P。取 ψ 为负熵$\psi(x) = \sum_{i=1}^{d+1} x_i \log x_i$，其中 $x_{d+1} \triangleq 1 - \sum_{i=1}^d x_i$，镜像 Langevin 扩散与*Wright–Fisher* 扩散 [Ethier, 1976]。

3.4 无梯度算子

前述 Stein 算子的例子基于这样一个思想：梯度 $\nabla \log p$ 以及该梯度的某些线性变换，在 P 下的积分为零。然而，在某些应用中，计算梯度的计算成本高得难以承受，甚至可能无法求得梯度。在这类情形下，我们可以借鉴重要性采样的思想，转而利用易于处理的近似分布 Π 的梯度。

假设 3.3 （ P的正则性）。 设 P 和 Π 是 $\mathbb{R}^d$ 上的概率分布。设 P 有正密度 p，且 Π 在 $\mathbb{R}^d$ 上有正且可微的密度 π 。

下列 Stein 算子被称为无梯度，因为它不涉及目标分布 P 的密度函数的梯度:

定义 3.7 (无梯度 Stein 算子；Han 和 Liu，2018年)。 在假设 3.3的条件下，基于 P，针对 Π 的无梯度 *Langevin Stein* 算子满足

$$(\mathcal{T}_{P,\Pi}\, g)(x) = \frac{\pi(x)}{p(x)}\left[(\nabla \cdot g)(x) + g(x) \cdot (\nabla \log \pi)(x)\right] \tag{3.16}$$

对于适当正则的向量值函数 g： $\mathbb{R}^d \to \mathbb{R}^d$ 以及所有 $x \in \mathbb{R}^d$。

若 P 和 Π 相等，则无梯度 Stein 算子与 (3.11) 中的 Langevin Stein 算子 $\mathcal{T}_P$ 一致。若 P 和 Π 不相等，则容易看出，$P(\mathcal{T}_{P,\,\Pi}\, g) = 0$ 当且仅当 $\Pi(\mathcal{T}_{\Pi} g) = 0$。因此，类似的论证可得以下结果:

命题 3.5 (无梯度算子的定义域)。 在假设 3.3的条件下，设 g： $\mathbb{R}^d \to \mathbb{R}^d$ 可微。对于 $\mathcal{T}_{P,\,\Pi}$ 如 (3.16)所定义，以下结论成立。

1. 如果 $g \cdot \nabla \log \pi, \nabla \cdot g \in L^1(\Pi)$，那么 $T_{P,\Pi} g \in L^1(P)$。
2. 如果 $g, T_{P,\Pi} g \in L^1(P)$，那么 $P(T_{P,\Pi} g) = 0$。

*Pr*证明。 第一条结论由（3.16）立即得出。对于第二条结论 ，

$$\int (\mathcal{T}_{P,\Pi} g)(x)\, \mathrm{d}P(x) = \int (\nabla \cdot (\pi g))(x)\, \mathrm{d}x.$$

在我们的假设下，向量场 $v = \pi g$ 满足 $v \in \mathcal{L}^1(\mathbb{R}^d)$ 以及 $\nabla \cdot v \in \mathcal{L}^1(\mathbb{R}^d)$，因此我们可以运用定理 3.2 来证明该结论。 □

无梯度 Stein 算子最初是在变分采样算法的背景下提出的 [Han 和 Liu， 2018], 相关内容将在第 6.3.2 节讨论。在特殊情形下，该算子也可视为扩散 Stein 算子，其中 $a + c = \pi/p$ 在（3.11）中。

3.5 离散算子及其他

本章最后一部分主要讨论目标 P 具有离散支撑集 $\mathcal{X}$，这意味着$\mathcal{X}$ 是一个可数集，例如在分析计数数据、文本数据、网络数据等时遇到的情形。不过，我们还会在第3.5.3节说明如何将这些思想扩展到更一般的状态空间，例如流形或无限维状态空间。

在离散情形下， P 可由其概率质量函数在 $p(x)$ 处的取值表示（示例 2.2）。

假设 3.4 （P的正则性）。 设 P 是定义在可数集 X 上、概率质量函数 p 处处为正的概率分布。

正如Henderson [1997], Shi 等人 [2022b], BarbourGotze[1988, 1990], [1991]提出的生成元方法也可应用于离散状态空间。离散状态空间上的（时间齐次）马尔可夫过程的生成元 A X（定义2.15）可以表示为一个转移率矩阵 Q，其元素为

$$Q_{x,y} = (Af_y)(x), \qquad f_y(x) = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases},$$

描述过程从状态 x 转移到状态 y 的无穷小速率，其中x， $y \in \mathcal{X}$。下面将给出几个例子。

3.5.1 Zanella 算子

我们探讨的第一种构造试图将稀疏性编码到 Stein 算子中，这可以简化其在较大或不可数集 X 上的计算。为此，在 X 上赋予无向图结构，用 $\mathrm{N}_x \subset$X 表示 $x \in$ X 的相邻顶点，并要求若 $y \in \mathrm{N}_x$，则 $x \in \mathrm{N}_y$ 。该图的边集记为 $\mathrm{E} = \{(x, y) : x \in \mathrm{X} \quad y \in \mathrm{N}_x\} \subset \mathrm{X} \times \mathrm{X}$。Zanella 过程 [Zanella 2019] 的转移矩阵具有以下形式

$$Q_{x,y} = \begin{cases} \kappa\left(\frac{p(y)}{p(x)}\right) & y \in \mathcal{N}_x, y \neq x \\ -\sum_{z \neq x} Q_{x,z} & y = x \end{cases}, \tag{3.17}$$

带有平衡函数 κ，即一个连续函数 $\kappa : (0, \infty) \to (0, \infty)$，满足平衡性质：对于所有 $t \in (0, \infty)$， $\kappa(t) = t\kappa(1/t)$。邻域结构的作用是将稀疏性引入过程，而平衡函数确保满足细致平衡，从而使过程具有 P-不变性。相关的 Stein 算子被 Hodgkinson 等人和 Shi 等人称为 Zanella Stein 算子 [2020], [2022b]:

定义 3.8 （Zanella 算子）。 固定一个平衡函数 κ : $(0, \infty) \to$ （$0, \infty$）。在假设 *3.4* 下，Zanella Stein 算子 对于 P 满足

$$(\mathcal{T}_P g)(x) = \sum_{y \in \mathcal{N}_x, y \neq x} \kappa\left(\frac{p(y)}{p(x)}\right)(g(y) - g(x))$$

f或对于所有足够正则的标量值函数 g : $X \to \mathbb{R}$ 以及所有 $x \in \mathcal{X}$.

与目前为止我们见过的连续域上的 Stein 算子相比，Zanella Stein 算子通过比值 $p(x)/p(y)$ 而非梯度来依赖于 p 。由于p 仅以比值的形式出现，因此即使 p 只在一个难以计算的归一化常数下给定，Zanella Stein 算子仍然适用；这一情况已在

本章开头讨论过。Zanella Stein 算子的特殊情形包括最小概率流*Stein*算子($\kappa(t) = \sqrt{t}$；Barp 等人，2019），该算子基于 Sohl-Dickstein 等人 [2011] 提出的最小概率流过程；以及*Barker Stein* 算子($\kappa(t) = \frac{t}{t+1}$；Hodgkinson 等人，2020年），该算子基于 Barker [1965]提出的方法。—

命题 3.6 （Zanella 算子的定义域）。 在假设 *3.4*的设定下，令

$$x \mapsto \sum_{y \in \mathcal{N}_x} \kappa\left(\frac{p(y)}{p(x)}\right) g(y)$$

属于 $L^1(P)$。则 $P(\mathcal{T}_P g) = 0$。

证明。 平衡性质蕴含，对所有 x， $y \in \mathcal{X}$，有 $p(x)\kappa\left(\frac{p(y)}{p(x)}\right) = p(y)\kappa\left(\frac{p(x)}{p(y)}\right)$，这些方程称为细致平衡方程。利用这些细致平衡方程，并利用所假设的可积性来重排绝对连续级数，

$$\begin{aligned} P(\mathcal{T}_P g) &= \sum_{x \in \mathcal{X}} p(x)(\mathcal{T}_P g)(x) \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{N}_x} p(x)\kappa\left(\frac{p(y)}{p(x)}\right)(g(y) - g(x)) \\ &= \sum_{(x,y) \in \mathcal{E}} \left[p(x)\kappa\left(\frac{p(y)}{p(x)}\right) g(y) - p(y)\kappa\left(\frac{p(x)}{p(y)}\right) g(x)\right] = 0, \end{aligned}$$

即为所求。 □

3.5.2 生灭Stein算子

我们从连续情形中获得的直觉，也可以通过为离散域$\mathcal{X}$ 赋予一个序而应用到离散情形。为简单起见，这里我们假设

$$\mathcal{X} \cong \{0, 1, \ldots, m_1 - 1\} \times \cdots \times \{0, 1, \ldots, m_d - 1\}, \tag{3.18}$$

这意味着集合 X 的基数为 $m_1 \cdots m_d$，其元素可以映射到一个笛卡尔网格上。接着定义增量算子 $\mathsf{inc}_i : \mathsf{X} \to \mathsf{X}$ 和减量算子 $\mathsf{dec}_i : \mathsf{X} \to \mathsf{X}$，它们以 $x \in \mathsf{X}$ 为输入，并分别将其第 i 个坐标替换为 $x_i + 1$ 模 m_i 和 $x_i - 1$ 模 m_i。定义出生率 $b_{i,x} = p(\mathsf{inc}_i(x))/p(x)$ 和死亡率 $d_{i,x} = 1$ 。需要注意的是，与 Zanella 过程一样，对概率质量函数 p 的依赖仅通过比率体现，从而无需显式求得归一化常数。X 上的生灭过程 [Karlin和McGregor, 1957] 是一个连续时间马尔可夫过程，其转移率矩阵为

$$Q_{x,y} = \frac{1}{d}\sum_{i=1}^{d} b_{i,x}\mathbb{I}\left[y = \mathrm{inc}_i(x)\right] + d_{i,x}\mathbb{I}\left[y = \mathrm{dec}_i(x)\right] - (b_{i,x} + d_{i,x})\mathbb{I}\left[y = x\right],$$

在此语境下，增量事件称为“出生事件”，减量事件称为“死亡事件”。生灭过程的生成元导出了生灭斯坦算子，包括以下作者在内的学者对此进行了研究：Brown 和 Xia [2001], Holmes [2004],Eichelsbacher 和 Reinert [2008], Hodgkinson 等人 [2020], Shi 等人 [2022b]:

定义 3.9（生灭算子）。 在假设 *3.4* 和 (3.18)的条件下，令出生率 $b_{i,x}$ 和死亡率 $d_{i,x}$ 如前所定义。对于 P ，生灭斯坦算子满足

$$(\mathcal{T}_P g)(x) = \frac{1}{d}\sum_{i=1}^{d} b_{i,x}(g_i(\mathrm{inc}_i(x)) - g_i(x)) - d_{i,x}(g_i(x) - g_i(\mathrm{dec}_i(x)))$$

对于所有向量值函数 $g : X \to \mathbb{R}^d$ 以及所有 $x \in X$ 。

由于状态空间有限，$\mathcal{X}$ 上的所有函数都属于生灭斯坦算子的定义域，正如下述结果所述:

命题3.7（生灭算子的定义域）。 在假设 *3.4* 和(3.18)的条件下，令 $g:\mathcal{X} \to \mathbb{R}^d$。那么 $P(\mathcal{T}_P g) = 0$。

证明。由 通过变量替换 $x \mapsto \mathrm{dec}_i(x)$，对每个 $i \in \{1,\ldots,d\}$，可得 tha t

$$\begin{aligned}\sum_{x\in\mathcal{X}} p(x)b_{i,x}(g_i(\mathrm{inc}_i(x)) - g_i(x)) &= \sum_{x\in\mathcal{X}} p(\mathrm{inc}_i(x))(g_i(\mathrm{inc}_i(x)) - g_i(x)) \\ &= \sum_{x\in\mathcal{X}} p(x)(g_i(x) - g_i(\mathrm{dec}_i(x)))\end{aligned}$$

由此得出该结果。 □

生灭斯坦算子与差分斯坦算子密切相关；对于h ： $\mathcal{X} \to \mathbb{R}$ ，其定义为

$$[(\mathcal{G}_P h)(x)]_i = h(\mathrm{dec}_i(x)) - b_{i,x}h(x),$$

Yang 等人 [2018] 对此进行了研究。事实上，若令 $h_i(x) = g_i(x) - g_i(\mathrm{inc}_i(x))$，则有关系式 ($\mathrm{T}_P g)(x) = \frac{1}{d}\sum_{i=1}^{d}[(\mathrm{G}_{\bar{P}} h_i)(x)]_i$，因此可以将生灭斯坦算子视为差分斯坦算子的逐坐标应用。Hodgkinson 等人 [2020], Matsubara et al. [2024] 讨论了将其推广至 X 可数无限的情形。

3.5.3 离散时间马尔可夫链的算子

用于构造斯坦算子的生成元方法很方便，因为非正式地说，在无穷小 $t \to 0$ 极限中计算量并不需要把积分

被计算。事实上，我们回顾定义2.14 可知，任何 P-不变马尔可夫转移核 K_t 会产生一个Stein 算子

$$(\mathcal{T}_P g)(x) = \int g(y) K_t(x, \mathrm{d}y) - g(x), \tag{3.19}$$

但这一构造并不实用，因为马尔可夫过程关于状态 y 在 $t > 0$ 个时间单位后（初始化于 x ）的积分通常无法计算。然而，对于离散时间马尔可夫过程以及呈现适当稀疏结构的离散域，此类积分可能归结为可以计算的有限和。事实上，离散空间 $\mathcal{X}$上离散时间马尔可夫过程的一步转移概率可由转移矩阵 K刻画，其元素 $K_{x、y}$ 表示转移到状态 y的概率，若该过程初始化于状态 x，其中 x、$y \in \mathcal{X}$。于是（3.19）成为本书所称的 马尔可夫链 *Stein* 算子：

定义 3.10 （马尔可夫链 Stein 算子）。 在假设 *3.4*的设定下，设 K为离散时间马尔可夫过程的转移矩阵，且 P 对其不变。与之关联的 马尔可夫链 Stein 算子 关于 P 满足

$$(\mathcal{T}_P g)(x) = \left(\sum_{y \in \mathcal{X}} g(y) K_{x,y} \right) - g(x),$$

对所有充分正则的标量值函数 g： $X \to \mathbb{R}$ 及所有 $x \in X$。

当转移矩阵每一行中的非零元素数量较少时，可以高效地计算马尔可夫链 Stein 算子。在离散时间马尔可夫过程可逆的假设下，即满足细致平衡方程时，可以得到马尔可夫链算子定义域的一个简单刻画：

命题 3.8 （马尔可夫链算子的定义域）。 在假设 *3.4* 的设定下，设离散时间马尔可夫过程具有 P-不变性，即对所有$y \in X$，$\sum_{x \in \mathcal{X}} p(x) K_{x,y} = p(y)$ ，并令 $g \in L^1(P)$ 。则 $P(T_P g) = 0$。

证明。 由 P-不变性可得，

$$\begin{aligned} P(\mathcal{T}_P g) &= \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{X}} g(y) K_{x,y} - \sum_{x \in \mathcal{X}} p(x) g(x) \\ &= \sum_{y \in \mathcal{X}} g(y) p(y) - \sum_{x \in \mathcal{X}} p(x) g(x) = 0, \end{aligned}$$

由于在我们假设的可积性条件下该级数绝对收敛，因此可以重新排列求和顺序。

□

例如，考虑随机扫描吉布斯采样器，它是在 d 维乘积空间 $\mathcal{X}$ 上的离散时间马尔可夫链，由转移矩阵定义为

$$K_{x,y} = \frac{1}{d}\sum_{i=1}^{d} p(y_i|x_{-i})\mathbb{I}\left[y_{-i} = x_{-i}\right],$$

其中， $p(y_i|x_{-i})$表示在分布 $X \sim P$下，对于所有 $j \neq i$ ，给定 $X_j = x_j$ 时 $X_i = y_i$ 的概率。这些动力学导出了*Gibbs Stein* 算子

$$(\mathcal{T}_P g)(x) = \frac{1}{d}\sum_{i=1}^{d}\left(\sum_{y\in\mathcal{X}:y_{-i}=x_{-i}} q(y_i|x_{-i})g(y)\right) - g(x) \tag{3.20}$$

相关研究见Bresler和Nagaraj [2019], Reinert和Ross [2019], Shi 等人 [2022b]。

3.5.4 一般Stein算子

我们的论述主要关注支撑在子集$\mathcal{X} \subseteq \mathbb{R}^d$上的概率分布 P，但在某些应用中，需要将其推广到有限维欧几里得空间之外。尽管我们不打算详细讨论斯坦算子的这些推广，但仍想简要提及朗之万Stein算子可以推广的两个方向。

对于 P 支撑在流形上的情形，已有多位作者对此进行研究；此时可以借助适用于该流形的、对 P不变的朗之万扩散的推广来处理 [Le等人，2024，Xu 和 Matsuda，2021年，Barp 等人，2022年，Qu和Vemuri， 2025]。例如，球面数据分析，或对满足非线性约束（如正定性）的矩阵值参数进行推断，都属于这类情形。

随机变量 $X \sim P$ 为无限维对象的情况在函数数据分析中十分常见，其中 P 支撑在适当的函数空间上。仅在我们讨论过的任何构造中直接取 $d \to \infty$ 并不可行，因为通常无法交换极限。相反，Wynne等人 [2025] 构建了适当的数学结构，以便在无限维背景下定义 Langevin Stein 算子的类比形式。该工作的关键观察是， P 相对于高斯测度绝对连续；高斯参考测度的结构被用于确保无限维背景下的可积性。

至此，我们对斯坦算子的讨论告一段落，并为下一章做好铺垫；在下一章中，我们将通过把 Stein 算子与 Stein集合相结合来构造 Stein差异。针对特定应用选择哪种 Stein 算子，部分取决于每种 Stein 算子在给定应用情境中的计算需求，同时也取决于每种 Stein 算子赋予相应 Stein 差异的性质，因此我们将对 Stein 算子选择的进一步讨论推迟到第4章。

4 Stein差异

一种*Stein* 差异是一类特殊的统计散度 $D: \mathcal{P} \times \mathcal{P} \to [0, \infty]$ （定义2.25），其形式为一种积分概率伪度量 [Müller, 1997]

$$D(P, Q) = \sup_{f \in \mathcal{F}} |P(f) - Q(f)| \tag{4.1}$$

其中，检验集 $\mathcal{F}$旨在避免显式计算关于目标分布 $P(f)$的积分 P。这是通过使用 Stein 算子 $\mathcal{T}_P$ (定义 3.1) 生成检验函数 $f = \mathcal{T}_P g$来实现的，这些函数在 P下的均值为零，因此 $P(f) = 0$可以直接计算。尽管 Stein 差异这一概念在 Gorham 和 Mackey [2015], 的工作中才相对较近地被提出，如今 Stein差异已应用于众多领域；我们将在第 6 章再讨论所有相关应用。本章旨在正式定义 Stein 差异这一概念，介绍 Stein 差异的具体例子并说明如何计算它们，同时讨论 Stein 差异的性质，尤其关注称为收敛检测 和收敛控制的概念。

首先，我们将给出 Stein 差异的严格定义。对于一个可测空间 X，令 $\mathrm{P}_{\mathcal{X}}$表示 X 上概率测度的集合；对于集合 G 上的算子 T，定义 T G $\triangleq$ {T g : $g \in$ G}。以下定义基于 Gorham 和 Mackey [2015,第 3] 节以及 Barp 等人 [2024, 定义 2] ，并采用简写 $\mathrm{L}^1_+(Q)$ 表示可测函数 f 的集合，其中其 Q-可积正部为 $(f)_+ \triangleq \max(f, 0)$。

定义*4.1*（*Stein* 差异）。设 P 是可测空间 X 上的概率分布。具有 *Stein*集合 G 和 *Stein* 算子 T_P: $G \to L^1(P)$（定义 *3.1*）的 *Stein* 差异是映射 $S(\cdot, T_P, G) : P_{\mathcal{X}} \to [0, \infty]$，其中对每个分布 $Q \in P_{\mathcal{X}}$,

$$\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}) \triangleq \sup_{f \in (\mathcal{T}_P \mathcal{G}) \cap \mathcal{L}^1_+(Q)} |Q(f)|. \tag{4.2}$$

将 (4.2) 与 (4.1) 进行比较可知，Stein 差异具有一个检验集 F $= (\mathrm{T}_P \mathrm{G}) \cap \mathrm{L}^1_+(Q)$，该检验集明确取决于目标分布 P 和所评估的候选分布 Q。按照设计，Stein 算子和 Stein集合确保每个检验函数$f \in \mathrm{T}_P \mathrm{G}$ 都是 P-可积的，且 $P(f) = 0$。然而，并非每个 $f \in \mathrm{T}_P \mathrm{G}$ 都必然在 Q 下可积，因此附加要求 $f \in \mathrm{L}^1_+(Q)$ 确保积分 $Q(f)$ 有定义。

Stein 算子$\mathcal{T}_P$ 和 Stein集合$\mathcal{G}$ 的定义4.1会影响 Stein 差异的数学性质，以及

计算 Stein 差异的难易程度。本章余下部分将讨论这一问题，并详细介绍和讨论若干 Stein 差异的具体示例。

4.1 Fisher散度

这里我们从 Fisher 散度谈起。它是一种广为人知的统计散度，接下来我们将看到，它实际上是刚刚介绍的 Stein 差异形式体系的一个特例。

定义 4.2 （Fisher散度）。设 P 和 Q 是 $\mathbb{R}^d$ 上的（*Borel*）分布，分别具有密度 p 和 q，且满足 ∇ 对数 p, ∇ 对数 $q \in L^2(Q)$。那么，Fisher散度 定义为

$$\mathrm{FD}(Q||P) \triangleq \int \|\nabla \log p - \nabla \log q\|_2^2 \, \mathrm{d}Q.$$

计算该散度时，无需知道 p 的归一化常数，此外也不需要关于 P 的期望。这使它成为估计难以处理的统计模型参数时的一种常用选择；这种方法称为 得分匹配。具体而言，假设 P 涉及参数 θ，记为 P_θ，其对应的密度为 p_θ。得分匹配是指选择参数值 θ，使 $\mathrm{FD}(Q||P_\theta)$最小。假设 Δ对数 $p_\theta \in \mathcal{L}^1(Q)$，Hyvärinen [2005] 利用分部积分证明了

$$\mathrm{FD}(Q||P_\theta) = [\text{constant in } \theta] + Q(2\Delta \log p_\theta + \|\nabla \log p_\theta\|^2). \tag{4.3}$$

如果能够获得来自 Q 的样本，那么很容易对（4.3）中出现的积分进行一致的蒙特卡洛近似。这使得得分匹配成为一种广泛适用的统计估计技术，并具备我们在第1章中作为动机介绍的许多理想特性。此外，Fisher散度还可以进行多种推广，例如推广到离散域 [Matsubara etal., 2024]。

然而，Fisher散度有两个主要局限，这促使本书随后讨论其他 Stein 差异。首先，与本书讨论的大多数 Stein 差异相比，Fisher散度在实际应用中的一个缺点是，使用（ p 或 p_θ ）需要计算二阶导数（ 4.3 ）。对于复杂的数据生成分布 P，例如通过求解物理控制方程组得到的分布，或许可以付出额外的计算成本来求得一阶导数，但求得二阶导数几乎总是难度大到无法承受。第二个潜在缺点取决于具体应用背景：Fisher散度强于大多数其他统计散度，包括 Kullback–Leibler散度（定义 2.26）以及我们将在第 4.4 节介绍的核 Stein 差异；关于这一点的进一步讨论，分别参见Ley和Swan [2013] 和Liu和Wang [2016, 定理 5。1]。使用强散度概念最明显的缺点是，其值可能为无穷大

在比较两个真正感兴趣的分布时，这会使它的用途仅限于一小部分正则性足以进行定量比较的概率分布。尽管如此，我们还是先通过说明 Fisher散度可以作为一般斯坦差异框架的一个实例来介绍 Stein 差异，不过其中的 Stein 集合依赖于 Q:

命题 *4.1*（*Fisher*散度作为 *Stein* 差异）。设 P 和 Q 是定义在 $\mathbb{R}^d$ 上的（*Borel*）分布，具有密度 p 和 q，且满足 $\nabla \log p, \nabla \log q \in L^2(Q)$ 以及 $\Delta \log p, \Delta \log q \in L^1(Q)$。那么 $\mathrm{FD}(Q||P) = S(Q, \mathcal{T}_P, \mathcal{G}_Q)$，其中 *Stein* 差异基于 *Langevin Stein* 算子 $(\mathcal{T}_P g)(x) = (\nabla \cdot g)(x) + g(x) \cdot (\nabla \log p)(x)$（见 *(3.6)*），且 *Stein*集合 $\mathcal{G}_Q = \{g : \sum_{j=1}^d |g_j|^2_{L^2(Q)} \leq 1, \mathcal{T}_P g, \mathcal{T}_Q g \in L^1(Q)\}$。

证明。 对于 Langevin Stein 算子和 $g \in \mathcal{G}$,

$$\begin{aligned} Q(\mathcal{T}_P g) &= Q(\mathcal{T}_P g) - Q(\mathcal{T}_Q g) \\ &= Q(\mathcal{T}_P g - \mathcal{T}_Q g) = Q(g \cdot (\nabla \log p - \nabla \log q)), \end{aligned}$$

其中，第一个等号成立是因为 g 属于$\mathcal{T}_Q$ 的定义域，这一点由命题 3.1得出。接着，根据 $\mathcal{L}^2(P)$ 的对偶结构,

$$\mathrm{FD}(Q||P) = \sup \left\{ \sum_{i=1}^d \int g_i (\nabla \log p - \nabla \log q)_i \, \mathrm{d}Q \ : \ \sum_{i=1}^d |g_i|^2_{\mathcal{L}^2(Q)} \leq 1 \right\}$$

其中，上确界在 $g^\star \propto \nabla \log p - \nabla \log q$ 处取得，因为我们的假设确保了$g^\star \in \mathcal{L}^2(Q)$。我们的假设进一步意味着 $g^\star \in \mathcal{G}$，因此

$$\begin{aligned} \mathrm{FD}(Q||P) &= \sup \left\{ \sum_{i=1}^d \int g_i (\nabla \log p - \nabla \log q)_i \, \mathrm{d}Q \ : \ g \in \mathcal{G} \right\} \\ &= \sup_{f \in \mathcal{T}_P(\mathcal{G})} Q(f) \end{aligned}$$

这与 Stein 差异（4.2）相同，因为根据构造 $\mathcal{T}_P(\mathcal{G}) \subset \mathcal{L}^1(Q)$，并且（4.2）中的模可以省略，因为 $\mathcal{T}_P(\mathcal{G})$ 是一个对称集。 □

Stein 差异具有非对称性，其中 P 保持固定而 Q 发生变化，这表明应采用一种比统计散度略微精细的概念，称为分离性。下面对相对于 P的任意差异 给出定义，即一个映射 $D(\cdot||P) : \mathcal{P} \to [0, \infty]$，其中$D(Q||P)$度量了 Q 作为 P近似时的不相似度。

定义 4.3（分离性）**：** 设 $P \in \mathcal{P}$ 且 $\mathcal{Q} \subset \mathcal{P}$。如果差异 $D(\cdot||P)$ 满足将 P 与 $\mathcal{Q}$ 分离，则当 $D(Q||P) = 0$ 时，对于所有 $Q \in \mathcal{Q}$，都有 $P = Q$ 。

请注意，任何有效的统计散度（参见定义2.25）都会将 P 与 $\mathcal{P}$区分开来， $P \in \mathcal{P}$可任意选取。尽管名称中有“散度”一词，Fisher散度实际上并非统计散度，因为当密度 p 不存在时，它没有明确定义。

下面的结果展示了 Fisher散度背景下的分离性；该陈述并非尽可能一般——我们将在本章后面给出更一般的结果——但我们选择在此呈现它，因为它的证明简单且具有启发性：

命题 4.2 （Fisher散度的一个分离结果）。 设 P 和 是定义在 $\mathbb{R}^d$ 上的一个（*Borel*）分布，其密度为 p，且 ∇ 对数 $p \in L^2(P)$。那么，*Fisher*散度能够将 P 与集合 Q 分离。Q 由定义在 $\mathbb{R}^d$ 上的（*Borel*）分布 Q 组成，这些分布具有密度 q，且满足 ∇ 对数 p,∇ 对数 $q \in L^2(Q)$。

证明。 若 FD$(Q||P) = 0$，那么，由于 $q > 0$ 和

$$\mathrm{FD}(Q||P) = \int q(x) \left\| \nabla \log \frac{p}{q} \right\|_2^2 \, \mathrm{d}x,$$

由此可知， $\nabla \log(p/q) = 0$ 在 $\mathbb{R}^d$上几乎处处成立。这意味着 p/q 几乎处处为常数；又由于 p 和 q 的积分都必须为 1，因此我们可以推出 P 和 Q 相等。 □

4.2 经典Stein差异

Stein方法是一种理论工具，因此自然会促使我们考虑较小的Stein集合 $\mathcal{G}$，与例如第 4.1 节中用于构造Fisher散度的Stein集合相比，这类集合在理论上更容易处理。对于支撑集位于 P上的目标分布 $\mathbb{R}^d$，经典*Stein*差异 [Gorham 和 Mackey，2015年，Gorham 等人，2019] 采用由有界函数及其有界利普希茨导数组成的Stein集合。为了明确给出这一定义，需要引入一些记号。首先介绍以下记号，它们将在本书余下部分中一直使用。对于实数集 $\|\cdot\|$ 上的范数 $\mathbb{R}^d$，回忆一下， $\|\cdot\|^*$表示第 1章中的对偶范数。

定义 4.4 (经典 Stein 集；Gorham 和 Mackey，2015年)。 给定一个范数 $\|\cdot\|$ 于 $\mathbb{R}^d$上，我们将经典 Stein 集 $\mathcal{G}_{\|\cdot\|}$ 定义为函数 g： $\mathbb{R}^d \to \mathbb{R}^d$ 满足

$$\sup_{x \neq y} \quad \max \left(\|g(x)\|^*, \|\nabla g(x)\|^*, \frac{\|\nabla g(x) - \nabla g(y)\|^*}{\|x - y\|} \right) \leq 1.$$

如第 *3.3* 节所述，该构造可以扩展到定义域 $\mathcal{X} \subseteq \mathbb{R}^d$，以及 Gorham 和 Mackey [2015], ，但这里我们关注的是 $\mathcal{X} = \mathbb{R}^d$ 的情形。这些经典 Stein 集通常与第 3.2 节中的扩散 Stein 算子配对使用。

定义 4.5 （经典扩散 Stein 差异；Gorham 等人，2019）。 在假设 *3.1*的情形下，经典扩散 Stein 差异 $\mathcal{S}(\cdot, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|})$ 由针对 $\mathcal{T}_P$ 的扩散 *Stein* 算子 P （定义 *3.5*）和经典 *Stein* 集 $\mathcal{G}_{\|\cdot\|}$ （定义 *4.4*）组合而成。

正如本章开篇所预示的，Stein 差异的实用性取决于三个因素：收敛检测、收敛控制和可计算性。前两个因素，即收敛检测和收敛控制，指的是当 Q_n 在大 n 极限下时，满足 $\mathsf{S}(Q_n, \mathsf{T}_P, \mathsf{G}) \to 0$ 的序列 $(Q_n)_{n\in\mathbb{N}}$ 的集合。只要 Stein 差异有定义，这个集合就自动包含常值序列 $Q_n = P$；不过，我们通常希望这一性质适用于在某种相当标准的意义下收敛到 P 的序列 $(Q_n)_{n\in\mathbb{N}}$。从数学角度来看，收敛检测和收敛控制分别是指用另一种统计散度对 Stein 差异作上界和下界估计；这种统计散度通常具有广为人知的数学性质。我们介绍的这类结果中的第一个由 Gorham 等人提出 [2019]，，表明只要扩散系数满足利普希茨条件，经典扩散 Stein 差异就能检测 s-Wasserstein 度量下的收敛。回想一下，s-Wasserstein 度量 $\mathsf{W}^s_{\|\cdot\|}$ 在定义 2.28 中引入。给定函数 $g : \mathsf{X} \to \mathbb{R}^d$，令 $M_0(g) \triangleq \sup_{x\in\mathcal{X}} \|g(x)\|_{\mathrm{op}}$ 并

$$M_k(g) \triangleq \sup_{x\neq y} \frac{\|(\nabla^{k-1}g)(x) - (\nabla^{k-1}g)(y)\|_{\mathrm{op}}}{\|x-y\|_2}$$

对于 $k \in \{1,2\}$。命题 4.3 和定理 4.1 中给出的上界包含以例如 $M_1(\cdot)$ 表示的常数；这里隐含地假设这些量有限，否则该界将变得平凡。

命题 *4.3*（收敛检测；*Gorham* 等人，*2019* 年的命题 *8*）。在假设 *3.1* 的条件下，考虑经典扩散 *Stein* 差异 $\mathcal{S}(\cdot, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|})$，其中 $\|\cdot\| \geq \|\cdot\|_2$。令 a 表示协方差系数，b 表示漂移，c 表示流系数。若 $a, b, c \in L^1(P)$ 且 $p(a+c) \in C^1(\mathbb{R}^d, \mathbb{R}^{d\times d})$，则，对于 $m \triangleq a + c$，

$$\begin{aligned}\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|}) \leq{}& \mathrm{W}^s_{\|\cdot\|}(Q,P)(2M_1(b) + M_1(m)) \\ &+ \mathrm{W}^s_{\|\cdot\|}(Q,P)^t\, 2^{1-t}\, \mathbb{E}_{Z\sim P}\left[(2\|b(Z)\| + \|m(Z)\|)^{\frac{s}{s-t}}\right]^{\frac{s-t}{s}}\end{aligned}$$

对于任意 $s \geq 1$ 和 $t \in (0,1]$。此外，对于 $\mu_0 \triangleq \mathbb{E}_{Z\sim P}[e^{2\|b(Z)\|+\|m(Z)\|}]$，

$$\begin{aligned}\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|}) \leq{}& \mathrm{W}^1_{\|\cdot\|}(Q,P)(2M_1(b) + M_1(m)) \\ &+ \min(\mathrm{W}^1_{\|\cdot\|}(Q,P), 2) \log\left(\frac{e\mu_0}{\min(\mathrm{W}^1_{\|\cdot\|}(Q,P), 2)}\right).\end{aligned}$$

因此，如果序列 $(Q_n)_{n\in\mathbb{N}}$ 在 s-Wasserstein 意义下收敛到 P，即 $\mathsf{W}^s_{\|\cdot\|}(Q_n, P) \to 0$（当 $n \to \infty$ 时），那么经典扩散 Stein 差异也能检测到这种收敛，具体而言，$\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|}) \to 0$ 在同一极限下成立。

决定 Stein 差异有用性的第二个因素是收敛控制，指的是构造 Stein 差异下界。下面的定理

Gorham 等人 [2019] 表明，基于快速混合扩散的经典扩散Stein差异能够控制在 1-Wasserstein 度量下的收敛。为陈述这一结果，我们采用第 2.1.4 节中引入的相同记号，并考虑一组马尔可夫过程 $(X_t^x)_{t\geq 0}$，其中每个过程都对应于 Itô扩散的解，该扩散是扩散 Stein 算子 T_P 所基于的对象（定义 3.4），但其初始分布为 $P_0 = \delta_x$，其中 $x \in \mathbb{R}^d$。令 P_t^x表示与 X_t^x相关联的概率测度。此外，若 $r\colon [0,\infty) \to \mathbb{R}$ 是一个非递增的可积函数，满足对所有$x, y \in \mathbb{R}^d$均有 $\mathsf{W}^1_{\|\cdot\|}(P_t^x, P_t^y) \leq r(t)\mathsf{W}^1_{\|\cdot\|}(\delta_x, \delta_y)$，则称 r 为与 T_P相关联的 Wasserstein衰减率。此外，对于定义在 $\mathbb{R}^d$上的矩阵值函数 g，令

$$F_k(g) \triangleq \sup_{x\in\mathbb{R}^d, \|v_1\|_2=1,\ldots,\|v_k\|_2=1} \|(\nabla^k g)(x)[v_1,\ldots,v_k]\|_F$$

其中 $\|\cdot\|_F$表示 Frobenius范数，并令

$$M_1^*(g) \triangleq \sup_{x\neq y} \frac{\|g(x)-g(y)\|_{\mathrm{op}}^*}{\|x-y\|_2}.$$

定理 *4.1*（收敛控制；*Gorham* 等人 *2019* 年的定理 *6* 和定理 *7*）。在假设 *3.1* 的设定下，考虑经典扩散 *Stein* 差异 $\mathcal{S}(\cdot, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|})$，其中 $\|\cdot\| \geq \|\cdot\|_2$。令 a 表示协方差系数，b 表示漂移，c 表示流系数，且 σ 表示扩散系数。进一步假设与 $\mathcal{T}_P$对应的 *Itô*扩散具有 *Wasserstein*衰减率 r，且 $s_r \triangleq \int_0^\infty r(t)\,dt$。

- 如果 $m \triangleq a + c$ 为常数，则

$$\mathsf{W}^1_{\|\cdot\|_2}(Q,P) \leq 3s_r \max\Big(\mathcal{S}(Q,\mathcal{T}_P,\mathcal{G}_{\|\cdot\|}), \sqrt[3]{\mathcal{S}(Q,\mathcal{T}_P,\mathcal{G}_{\|\cdot\|})\sqrt{2}\,\mathbb{E}\left[\|G\|_2\right]^2 (2M_1(b)+\tfrac{1}{s_r})^2}\Big), \tag{4.4}$$

  其中 $G \in \mathbb{R}^d$ 是标准正态向量。

- *If* σ 是利普希茨的，且 b和 σ 具有局部利普希茨二阶导数，则

$$\mathsf{W}^1_{\|\cdot\|_2}(Q,P) \leq \beta_1 \max\left(\mathcal{S}(Q,\mathcal{T}_P,\mathcal{G}_{\|\cdot\|}), \sqrt{\mathcal{S}(Q,\mathcal{T}_P,\mathcal{G}_{\|\cdot\|})\mathbb{E}\left[\|G\|_2\right]}\right),$$

  其中 β_1为依赖于 $M_0(\sigma^{-1})$ $M_1(\sigma)$ $F_2(\sigma)$ $M_1(b)$ $M_2(b)$ s_r的常数，且 $M_1^*(m)$。

- 如果 σ 是利普希茨的，且 $\nabla^3 b$ 和 $\nabla^3\sigma$ 是局部*Lipschitz*的，那么

$$\mathsf{W}^1_{\|\cdot\|_2}(Q,P) \leq \beta_2\,\mathcal{S}(Q,\mathcal{T}_P,\mathcal{G}_{\|\cdot\|})\sqrt{d}\max\left(1, \log\left(\frac{1}{\mathcal{S}(Q,\mathcal{T}_P,\mathcal{G}_{\|\cdot\|})}\right)\right)$$

  对于一个依赖于 β_1 的常数 $\beta_2 > 0$，$M_3(\sigma)$ $M_3(b)$，以及 r。

因此，若序列 $(Q_n)_{n\in\mathbb{N}}$ 根据经典扩散 Stein 差异收敛到 P，即 $\mathrm{S}(Q_n, \mathrm{T}_P, \mathrm{G}_{\|\cdot\|}) \to 0$（当 $n \to \infty$ 时），则该序列在 1-Wasserstein 度量下也收敛，即在相同极限下 $\mathrm{W}^1_{\|\cdot\|_2}(Q_n, P) \to 0$。另请参见 Mackey 和 Gorham [2016, 定理 2.1 和引理 2.2]，以及 Gorham 和 Mackey [2015, 定理 2]，其中给出了在其他前提条件下类似的收敛控制结果。

决定 Stein 差异实用性的最后一个因素，是其计算的便捷程度。如果候选分布 Q 支撑在一个大小为 n、维度为 $d = 1$的有限集合上，那么 Gorham 和 Mackey [2015, 定理 9] 以及 Gorham 等人 [2019] 证明了，经典扩散 Stein 差异可以精确地计算为一个凸二次约束二次规划的解，其线性目标函数为 $O(n)$ 个变量，并具有 $O(n)$ 个约束。然而，对于更大的 d，目前尚不清楚如何高效地计算经典扩散 Stein 差异。这一局限源于经典 Stein 集 $\mathcal{G}_{\|\cdot\|}$ 最初是作为理论工具而非计算工具提出的。幸运的是，如果我们考虑 Stein 集的替代方案，就可以避开这一计算挑战，下面将对此进行讨论。

4.3 图 Stein 差异

为了实现通常可计算的 Stein 差异，Gorham 和 Mackey [2015]引入了图 *Stein* 集，这些集合仅在属于近似分布支撑集的点对处施加有界性和光滑性约束 Q。

定义 4.6 (图 Stein 集；Gorham 和 Mackey，2015年）。对于概率分布Q 在 $\mathbb{R}^d$ 上，该分布具有离散支撑集，且给定一个图 $G =$ （V， E），其中 $V =$ 支撑集(Q)，我们定义图 Stein 集，

$$\mathcal{G}_{\|\cdot\|,Q,G} \triangleq \Bigg\{ g : \max\left(\|g(v)\|^*, \|\nabla g(v)\|^*, \tfrac{\|g(x)-g(y)\|^*}{\|x-y\|}, \tfrac{\|\nabla g(x)-\nabla g(y)\|^*}{\|x-y\|}\right) \le 1, \\ \tfrac{\|g(x)-g(y)-\nabla g(x)(x-y)\|^*}{\frac{1}{2}\|x-y\|^2} \le 1, \tfrac{\|g(x)-g(y)-\nabla g(y)(x-y)\|^*}{\frac{1}{2}\|x-y\|^2} \le 1, \\ \forall (x,y) \in E, v \in V \Bigg\}.$$

因此，图 Stein 集 $\mathcal{G}_{\|\cdot\|,\ Q,\ G}$ 是经典 Stein 集 $\mathcal{G}_{\|\cdot\|}$（定义 4.4）的超集。请注意，与经典 Stein 集不同，图 Stein 集依赖于 Q，而且还需要指定图 E 的边集 G。事实证明，只要使用适当的图 G 来构造图 Stein 集，这种依赖性在理论上是合理的，正如我们将在下文的命题4.5 中看到的那样。Gorham 和 Mackey [2015], Gorham 等人 [2019] 将这些图 Stein 集与扩散 Stein 算子（定义 3.5）相结合，得到所谓的图扩散 *Stein* 差异：

定义 *4.7*（图扩散 *Stein* 差异；*Gorham* 等人，*2019*）。在假设 *3.1* 的设定下，图扩散 *Stein* 差异 $Q \mapsto S(Q, T_P, G_{\|\cdot\|,Q,G})$ 将扩散 *Stein* 算子 T_P 用于 P（定义 *3.5*）与图 *Stein* 集 $G_{\|\cdot\|,Q,G}$（定义 *4.6*）结合起来。

值得注意的是，当 Q 具有有限支撑 $\{x_1, \cdots x_n\}$ 且范数 $\|\cdot\|$ 为 $\|\cdot\|_1$时，因此对偶范数 $\|\cdot\|^*$ 为 $\|\cdot\|_\infty$，图扩散 Stein 差异可以通过使用高效的现成求解器并行求解 d个相互独立的线性规划来计算，具体如下述结果所示：

命题 4.4 (计算图扩散 Stein 差异；Gorham 等人Gorham 等人，2019)。 假设 $Q = \sum_{i=1}^n q(x_i)\delta_{x_i}$ 对 $x_i \in \mathbb{R}^d$ 成立，且 q 是定义在 V =supp(Q)上的概率质量函数。对于任意边集 E，图 $G = (V, E)$ 上的图扩散 *Stein* 差异 $S(Q, T, G_{\|\cdot\|_1, Q, G})$ 等于

$$\sum_{j=1}^{d} \sup_{\psi_j \in \mathbb{R}^n, \Psi_j \in \mathbb{R}^{d\times n}} \Bigg\{ \sum_{i=1}^{n} q(x_i)\left(2b_j(x_i)\psi_{ji} + \sum_{k=1}^{d} m_{jk}(x_i)\Psi_{jki}\right) \tag{4.5}$$
$$s.t.\ \|\psi_j\|_\infty \le 1,\ \|\Psi_j\|_\infty \le 1,$$
$$and\ for\ all\ i \ne l, (x_i, x_l) \in E,$$
$$\max\left(\frac{|\psi_{ji}-\psi_{jl}|}{\|x_i-x_l\|_1}, \frac{\|\Psi_j(e_i-e_k)\|_\infty}{\|x_i-x_l\|_1}, \frac{|\psi_{ji}-\psi_{jl}-\langle\Psi_j e_i, x_i-x_l\rangle|}{\frac{1}{2}\|x_i-x_l\|_1^2},\right.$$
$$\left.\frac{|\psi_{ji}-\psi_{jl}-\langle\Psi_j e_i, x_l-x_i\rangle|}{\frac{1}{2}\|x_i-x_l\|_1^2}\right) \le 1\Bigg\},$$

其中 ψ_{ji} 和 $\Psi_{j\ ki}$ 分别表示定义 *4.6* 中的值 $g_j(x_i)$ 和 $\nabla_k g_j(x_i)$.

每个线性规划中的约束数量由图中的边数 $|E|$决定。粗略来说，约束越多，线性规划的数值解就越难求。另一方面，如果边集 E 过小，图 Stein 集可能会过大，难以检测收敛。考察图扩散 Stein 差异的收敛检测和控制性质，有助于了解应如何选择边集 E 。

为分析边集 E 的选择对收敛检测和控制的影响，我们需要引入一些额外术语。顶点集为 V 且嵌入在 $\mathbb{R}^d$ 中的图 $G = (V, E)$ 称为 t-伸展图 [Chew, 1986, Peleg and Schäffer, 1989] ，如果将每条边 $(x, y) \in E$ 的权重 $\|x - y\|$ 设为其在 $\mathbb{R}^d$ 中的长度，那么对于所有满足 $x \ne y$ 的 $x, y \in V$，都存在一条连接 x 和 y 的路径，其路径总权重不超过 $t\|x - y\|$。直观地说，在 $d = 1$ 的情形下，可以先按升序排列顶点，耗时 $O(n$log $n)$，再连接相邻顶点，总共需要 $n - 1$ 条边。只要图 G 是一个 t-伸展图，图扩散 Stein 差异便继承了相应经典扩散 Stein 差异的收敛检测和控制性质；事实上，二者等价：

命题 4.5 (经典扩散 Stein 差异与图扩散 Stein 差异的等价性；Gorham 等人，2019的**命题 13**）若 $G = (\mathrm{supp}(Q)\ E)$ 是 t-伸展图，且其对应于 $t \geq 1$，则

$$\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|}) \leq \mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|,Q,G}) \leq \kappa_d t^2\, \mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|})$$

其中，κ_d 与（Q、P、T_P、G）相互独立，并且仅取决于 d 和 $\|\cdot\|$。

因此，从保留经典扩散 Stein 差异所具有的理想收敛检测和控制性质的角度来看，我们应当选择 G，使其成为某个 $t \geq 1$的 t-伸展图。相应的计算成本是多少？如果我们考虑 $\mathbb{R}^d$上的范数 $\|\cdot\| = \|\cdot\|_p$，则可以在 $O(\kappa_d' n\log(n))$的期望时间内构造一个具有 $O(\kappa_d' n)$条边的 2-伸展图，其中 κ_d' 是一个依赖于 d和 p的常数 [Har-Peled 和 Mendel， 2005]。构造出 t-伸展图后，根据命题 4.4，可以通过求解一个有限维凸优化问题来计算图扩散 Stein 差异；该问题具有线性目标函数、 $O(n)$个变量和 $O(\kappa_d' n)$个凸约束。最后，我们注意到，图 Stein 差异还可用于对受约束域 X $\subset \mathbb{R}^d$ 施加边界约束，具体如 Gorham 和 Mackey [2015, 第 4.4]节所述。

4.4 核 Stein 差异

尽管图 Stein 差异克服了经典 Stein 差异所涉及的算法复杂度，但由于需要借助线性规划，其实际实现并非易事。本节将讨论另一类称为核 *Stein* 差异的 Stein 差异。其计算较为简单，但对收敛检测和控制的分析则更为复杂。我们首先给出一个定义，其中 Stein 算子和核可以是一般形式，然后详细探讨该定义对特定 Stein 算子和核所产生的影响。

定义 4.8 （核 Stein 差异）。 设 P 是可测空间 X 上的概率分布，且 $T_P : D \to L^1(P)$ 是针对 P 的 *Stein* 算子，其定义域 D 包含形如 $g : X \to \mathbb{R}^d$的函数。设 $K : X \times X \to \mathbb{R}^{d\times d}$ 是一个再生核（若 $d > 1$则为矩阵值），且 $H_K \subset D$。一个核 Stein 差异

$$\mathrm{KSD}_P(Q) \triangleq \mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_K)$$

将针对 P 的 *Stein* 算子 T_P 与 *Stein* 集合 $G_K \triangleq \{g : \|g\|_{\mathcal{H}_K} \leq 1\}$结合起来，其中后者是 H_K中的单位球。

核 Stein 差异（KSD）的一个或许最重要的性质是，对于 $Q = \sum_{i=1}^n w_i \delta_{x_i}$ 具有有限支撑的分布，KSD 可以用闭式计算。事实上，对于这类 Q，我们自动有$\mathcal{T}_P g \in \mathcal{L}^1(Q)$ 因此，根据定义4.1，KSD 可表示为

$$\mathrm{KSD}_P(Q) = \sup_{f \in \mathcal{T}_P(\mathcal{G}_K)} |Q(f)| = \sup_{\|g\|_{\mathcal{H}_K} \leq 1} Q(\mathcal{T}_P g) = \sup_{\|g\|_{\mathcal{H}_K} \leq 1} \sum_{i=1}^n w_i (\mathcal{T}_P g)(x_i) \tag{4.6}$$

这里利用$\mathcal{T}_P(\mathcal{G}_K)$是对称集这一事实去掉了模（这可由$\mathcal{G}_K$ 是对称集这一事实推出）。接下来的步骤要求满足一定的正则性；这里我们先概述论证，再于第4.4.1节给出精确结果。考虑线性泛函 $L\colon \mathcal{H}_K \to \mathbb{R}$ ，满足

$$L(g) \triangleq \sum_{i=1}^{n} w_i (\mathcal{T}_P g)(x_i) \quad \text{for each} \quad g \in \mathcal{H}_K.$$

当 L 连续时，根据定理 2.2，我们可以将 $L(g) = g, \ell_{\mathcal{H}_K}$ 写成某个 Riesz 表示元 $\ell \in \mathrm{H}_K$。利用这一 Riesz 表示，可以看出 (4.6) 中的上确界在 $g = \ell/\|\ell\|_{\mathcal{H}_K}$时达到，而上确界本身的值为 $\|\ell\|_{\mathcal{H}_K}$。为了显式确定ℓ，我们可以固定 $x \in \mathrm{X}$ 和 $u \in \mathbb{R}^d$ ，并利用来自 H_K 的再生性质（定理 2.5）以及记号 $K_x = K(\cdot\ x)$，从而看出

$$\langle \ell(x), u \rangle = \langle \ell, K_x u \rangle_{\mathcal{H}_K} = L(K_x u) = \sum_{i=1}^{n} w_i (\mathcal{T}_P K_x u)(x_i).$$

由于这对所有 $x \in \mathcal{X}$ 和 $u \in \mathbb{R}^d$均成立，因此我们有

$$\ell(\cdot) = \sum_{i=1}^{n} w_i (\mathcal{T}_P^{(1)} K_\cdot)(x_i)$$

其中 $\mathcal{T}_P^{(1)}$ 表示 Stein 算子作用于各行上 $K_\cdot$，从而 $(\mathcal{T}_{(1)\,P}\ \ K_\cdot)(x_i) \in \mathbb{R}^d$。假设 $\mathcal{T}_P$ 是一阶微分算子，我们可以利用微分再生性质 [Barp 等人，2024，引理 4] 得到

$$\begin{aligned} \|\ell\|_{\mathcal{H}_K}^2 &= \left\langle \sum_{i=1}^{n} w_i (\mathcal{T}_P K_\cdot)(x_i), \sum_{j=1}^{n} w_j (\mathcal{T}_P K_\cdot)(x_j) \right\rangle_{\mathcal{H}_K} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \underbrace{\mathcal{T}_P^{(1)} \mathcal{T}_P^{(2)} K(x_i, x_j)}_{(\star)}, \end{aligned}$$

where $\mathcal{T}_P^{(2)}$ 作用于 K 的列上。表达式 $(\star)$ 本身定义了一个核

$$k_p(\cdot, \cdot) \triangleq \mathcal{T}_P^{(1)} \mathcal{T}_P^{(2)} K(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, \tag{4.7}$$

我们称其为与 Stein 算子 $\mathcal{T}_P$ 和 基核 K 相关联的 ***Stein*** 核 。综上所述，我们已经证明

$$\mathrm{KSD}_P(Q) = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j k_p(x_i, x_j)},$$

这是一个显式公式，这与图 Stein 差异的情况形成对比，后者需要借助线性规划工具。Stein 差异的显式公式是第6章中讨论的许多算法的基础。下一节将使上述论证更加严谨。

4.4.1 Langevin KSD 详解

本节为介绍 Langevin Stein 算子$\mathcal{T}_P$（定义3.2）的一些精确结果铺垫。为此，我们将处于假设3.1的条件下，其中 P 在 p 上具有正的可微密度$\mathcal{X} = \mathbb{R}^d$，且该 Stein 算子在此条件下有良好定义。我们提出的第一个结果严谨地说明了何时可以显式计算 KSD。

定理 *4.2*（*Langevin KSD* 即最大均值差异；*Barp* 等人，*2024* 年的定理 *1*）。在假设 *3.1* 的条件下，设 $K : X \times X \to \mathbb{R}^{d\times d}$ 是一个矩阵值核，且 $T_P(H_K) \triangleq \{T_P f : f \in H_K\}$ 中的元素均有定义。那么，$T_P(H_K)$ 可以赋予再生核希尔伯特空间的结构，其再生核为

$$k_p(x,y) \triangleq \sum_{i=1}^{d}\sum_{j=1}^{d} \frac{1}{p(x)p(y)} \partial_{y_j}\partial_{x_i}\left(p(x)K_{i,j}(x,y)p(y)\right). \tag{4.8}$$

此外，若 $T_P(H_K) \subset L^1(P)$ 且 $P(T_P(H_K)) = 0$，也就是说，对所有 $g \in H_K$，$T_P g \in L^1(P)$ 和$P(T_P g) = \{0\}$ 均成立，则核 *Stein* 差异与基于 *Stein* 核 k_p 的最大均值差异相同；即对所有 $Q \in P_{\mathcal{X}}$，$KSD_P(Q) = MMD_{k_p}(P,Q)$。

最大均值差异对于分布 $Q = \sum_{i=1}^{n} w_i \delta_{x_i}$ 在有限支撑情形下的可计算性已在第2.3.3节中讨论，遵循类似的论证，我们得到

$$\mathrm{KSD}_P(Q) = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j k_p(x_i,x_j)},$$

从而展示了 Stein 核的计算作用。这将第4.4 节中的论证概要在 Langevin KSD 情形下加以形式化。

我们现在将注意力转向定理 4.2 中出现的条件 $\mathrm{T}_P(\mathrm{H}_K) \subset \mathrm{L}^1(P)$ 与 $P(\mathrm{T}_P(\mathrm{H}_K)) = 0$。为方便起见，令 $\mathrm{P}_{K,0}$ 表示 X 上使得 $\mathrm{T}_Q(\mathrm{H}_K) \subset \mathrm{L}^1(Q)$ 且 $Q(\mathrm{T}_Q(\mathrm{H}_K)) = 0$ 成立的分布 Q 的集合，从而定理 4.2 的条件可紧凑地表示为 $P \in \mathrm{P}_{K,0}$。由命题 2.2，我们有 $\mathrm{T}_P(\mathrm{H}_K) \subset \mathrm{L}^1(P)$ 当且仅当 $\mathrm{H}_{k_p} \subset \mathrm{L}^1(P)$，当且仅当 $P \in \mathrm{P}^{\mathcal{H}_{k_p}}$；而由命题 2.3，若

$$\int \sqrt{k_p(x,x)}\mathrm{d}P(x) < \infty. \tag{4.9}$$

利用（4.8）中 Stein 核的显式形式，可以推导出，只要 $\mathcal{H}_K$ 中的每个元素都有界且偏导数有界，且 $\int \|\nabla \log p\| \mathrm{d}P < \infty$，则（4.9）成立。最后，命题 3.1 给出了使 $P(\mathcal{T}_P(\mathcal{H}_K)) = \{0\}$成立的正则性条件。这些论证汇总于命题 4.6。

命题 4.6（Stein 可嵌入条件；Barp 等人的命题 3 和注 5，Barp 等人，2024）。在假设 *3.1*的设定下，设 $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d\times d}$ 是一个矩阵值核，且满足$\mathcal{T}_P(\mathcal{H}_K)$ 有定义。那么：

(a) $T_P(H_K) \subset L^1(P) \Leftrightarrow P \in P^{\mathcal{H}_{k_p}}$。*(b)* 若 $\int \sqrt{k_p(x,x)}dP(x) < \infty$，则 $P \in P^{\mathcal{H}_{k_p}}$。*(c)* 若 $\int \|\nabla log\ p\| dP < \infty$ 且 H_K 中所有 v 均有界，且偏导数有界，则 $\int \sqrt{k_p(x,x)}dP(x) < \infty$。*(d)* 若 $P \in P^{\mathcal{H}_{k_p}}$，则 $\iint k_p(x,y)dP(x)dP(y) = 0 \Leftrightarrow P(H^{k_p}) = \{0\} \Leftrightarrow P \in P_{K,0}$。*(e)* 若 $P \in P^{\mathcal{H}_{k_p}}$ 且 $H_K \subset L^1(P) \cap C^1(\mathbb{R}^d)$，则 $P \in P_{K,0}$。*(f)* 若 $x \mapsto \|K(x,x)\|$ 有界，则 H_K 中所有 v 均有界。*(g)* 若 $(x,y) \mapsto \|\partial_{x^i}\partial_{y^i}K(x,y)\|$ 存在且有界，则 H_K 中所有 v 均具有有界的 x^i-偏导数。*(h)* 若 $K \in C_b^{(1,1)}(\mathbb{R}^d)$，则 $H_K \subset C_b^1(\mathbb{R}^d)$。

特别地，当 $P \in \mathcal{P}_{K,0}$ 时，定理4.2所要求的条件即可满足，其中 $\int \|\nabla \log\ p\| \mathrm{d}P < \infty$ 且 $K \in C_b^{(1,1)}(\mathbb{R}^d)$。

最后，我们以讨论一个特殊情形下更明确的 Stein 核公式来结束本节：

注 4.1 (标量核朗之万核 Stein 差异)。在 $K(x,y) = k(x,y)I$ 的情形下，对于标量核 $k : X \times X \to \mathbb{R}$,

$$\begin{aligned} k_p(x,y) = &\nabla_x \cdot \nabla_y k(x,y) + (\nabla \log p)(x) \cdot \nabla_y k(x,y) \\ &+ (\nabla \log p)(y) \cdot \nabla_x k(x,y) + (\nabla \log p)(x) \cdot (\nabla \log p)(y) k(x,y), \end{aligned}$$

其中，例如，$\nabla_x k(x,y)$ 表示关于 x 变量的梯度，作用于 $k(x,y)$。

4.4.2 使用 Langevin KSD 区分分布

本节重点讨论使 Langevin KSD 能够将 P 与其他分布 Q 分离的充分条件；回顾定义 4.3，这意味着对于所有 $Q \in$ Q，$\mathrm{KSD}_P(Q) = 0$ 当且仅当 Q 与 P 相等。本文将介绍三个主要的分离结果。第一个结果称为基于得分的分离（定理 4.3），它要求核 K满足较弱意义上的特征性，但也要求 K 具有一定的光滑性，并限制了可与 P分离的分布集合 Q 。第二个结果称为基于 L^2的分离（定理 4.5），不要求 K 具有光滑性，适用于某些第一个结果无法涵盖的分布 Q ，但要求 Q 在 $\mathbb{R}^d$上具有密度。第三个结果或许最为实用：它只要求核具有适当的光滑性且为平移不变核，并保证 P 与所有其他分布 $Q \in$P 分离（定理 4.7）。

定义 *4.9*（$D^1_{L^1}(\mathbb{R}^d)$、嵌入与 $D^1_{L^1}(\mathbb{R}^d)$-特征核；*Barp* 等人，*2024* 的定义 *6* 和 *8*）。令 $D^1_{L^1}(\mathbb{R}^d)$ 表示 $C_0(\mathbb{R}^d)$ 上连续线性泛函的集合。若对于每个 $D \in D^1_{L^1}(\mathbb{R}^d)$，都存在 $\phi_D \in H_K$，使得对于所有 $h \in H_K$，$D(h) = \langle \phi_D, h \rangle_{\mathcal{H}_K}$，则称 $D^1_{L^1}(\mathbb{R}^d)$ 嵌入 H_K。此外，如果该嵌入 $D \mapsto \phi_D$ 是单射，则称 K 为 $D^1_{L^1}(\mathbb{R}^d)$-特征核。

这将定义2.30 中从概率分布上的积分推广得到的特征核概念，扩展到更一般的线性算子。Simon-Gabriel 和 Schölkopf [2018, 定理 12 、表 1 和推论 38] 证明了，任意 $C^1_0(\mathbb{R}^d)$-通用标量值核 k 以及任意 $C^{(1,1)}(\mathbb{R}^d)$ 平移不变 k 且谱测度支撑集为整个空间的核都是 $\mathcal{D}^1_{L^1}(\mathbb{R}^d)$-特征的，适用于维度 $d = 1$。这些结果涵盖了 KSD 中常用的所有平移不变基核，包括高斯分布核、逆多二次核和 Matérn核。接下来将介绍一组用于构造 $\mathcal{D}^1_{L^1}(\mathbb{R}^d)$-特征矩阵值核 K的方法:

命题 4.7 ($\mathcal{D}^1_{L^1}(\mathbb{R}^d)$-特征条件；Barp 等人Barp 等人，2024)。设矩阵值核 K 满足 $\mathcal{H}_K \subset C^1_b(\mathbb{R}^d)$ 且是 $\mathcal{D}^1_{L^1}(\mathbb{R}^d)$-特征的。则以下结论成立。

(a) 若 $a \in C^1_b$ 严格为正，则 $a(x)K(x,y)a(y)$ 是 $D^1_{L^1}(\mathbb{R}^d)$-特征的。

(b)若 $b : \mathbb{R}^d \to \mathbb{R}^d$ 是 *Lipschitz*（利普希茨） $C^1(\mathbb{R}^d)$-微分同胚，则复合核$K(b(x)\ b(y))$ 是 $D^1_{L^1}(\mathbb{R}^d)$- 特征的。

(c) 若 k_j 对于 $j \in \{1, \ldots, d\}$ 是 $D^1_{L^1}$-特征的，则 $\mathrm{diag}(k_1, \ldots, k_d)$ 是 $D^1_{L^1}(\mathbb{R}^d)$-特征的。

现在我们准备介绍第一个主要分离结果：基于得分的分离性。由命题 2.2 可知，$P \in \mathrm{P}^{\mathcal{H}_{k_p}}$ 当且仅当 $\mathrm{H}_{k_p} \subset \mathrm{L}^1(P)$；这里我们自然地扩展这一记号：给定 $f : \mathrm{X} \to \mathbb{R}^d$，令 P_f 表示 X 上满足 $x \mapsto \|f(x)\|$ 属于 $\mathrm{L}^1(Q)$ 的测度 Q 的集合。此外，对于密度 p，我们用 $s_p \triangleq \nabla \log p$ 作为 Stein 得分的简写。

定理 4.3 (基于得分的朗之万 KSD 分离；Barp 等人，2024)。假设矩阵值核 K 满足 $H_K \subset C^1_b(\mathbb{R}^d)$ 且具有 $D^1_{L^1}(\mathbb{R}^d)$-特征性。若 $P \in P_{K,0}$，则 *Langevin* 核 Stein 差异 可将 P 与 P_{s_P}分离。

也就是说，若 K 满足定理 4.3 的条件，且 $\int \|\nabla$对数 $p\| \, \mathrm{d}Q < \infty$，则 Langevin KSD 满足 核 Stein 差异$_P(Q) = 0$ 当且仅当 P 与 Q 相等。

接下来我们转而关注第二个主要分离结果；基于 $\mathcal{L}^2$ 的分离。为此，需要对核 K 作出以下略显严格的假设:

定义4.10（$\mathcal{L}^2$（$\mathbb{R}^d$）-积分严格正定；Liu等人，2016）。我们称矩阵值核 K 是$\mathcal{L}^2$（$\mathbb{R}^d$）-积分严格正定（*ISPD*），如果$\mathcal{H}_K \subset \mathcal{L}^2$（$\mathbb{R}^d$），并且

$$\iint g(x)^\top K(x,y) g(y) \, \mathrm{d}x \mathrm{d}y > 0$$

对于所有 $g \in L^2(\mathbb{R}^d)$，其中 $|g|_{\mathcal{L}^2(\mathbb{R}^d)} = \left(\int g(x)^2 \mathrm{d}x\right)^{1/2} > 0$。

对于连续平移不变核，我们可以使用 Bochner 定理（定理 2.4）来推导 K 是$\mathcal{L}^2(\mathbb{R}^d)$-积分严格正定的条件：

定理 4.4 （$\mathcal{L}^2(\mathbb{R}^d)$-积分严格正定条件；定理 4 出自 Barp 等人，2024）。以下断言对于矩阵值核 $K: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$。

（*a*）设 $(k_j)_{j=1}^d$ 是 $\mathbb{R}^d$ 上的连续平移不变核，且$H_{k_j} \subset L^2(\mathbb{R})$。若每个 k_j 的谱测度具有全支撑，则 $K = diag(k_j)$是$L^2(\mathbb{R}^d)$-积分严格正定。（*b*）若 K 是$L^2(\mathbb{R}^d)$-积分严格正定，且 $A: \mathbb{R}^d \to \mathbb{R}^{d\times d}$ 有界可测，并对每个 x， $A(x)$可逆，则倾斜核 $A(x)K(x,y)A(y)^\top$ 也是$L^2(\mathbb{R}^d)$-积分严格正定。（*c*）若H_K 可分，$sup_x |K_x u|_{\mathcal{L}^1(\mathbb{R}^d)} < \infty$，且对每个 x 和 $u \in \mathbb{R}^d$有 $K_x u \in L^2(\mathbb{R}^d)$，则$H_K \subset L^2(\mathbb{R}^d)$。（*d*）设对某个 $u \in \mathbb{R}^d$有 $K_x u \in L^1(\mathbb{R}^d)$。若 K 平移不变，或更一般地，若 $K_x u$ 有界，则 $K_x u \in L^2(\mathbb{R}^d)$。

现在我们可以给出第二个主要的分离结果。令 $\partial\mathcal{H}_K$ 表示如下形式的函数集：$g: \mathbb{R}^d \to \mathbb{R}^d$ $g_i(x) = \partial_{x_i} h_i(x)$，其中 $h \in \mathcal{H}_K$。

定理 4.5 （基于$\mathcal{L}^2$的 Langevin KSD 分离性；Barp 等人的定理 5，2024年）。设 $P \in \mathcal{P}_{K,0}$ 对于一个矩阵值核 $K: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$。那么

1. 若 K 是 $L^2(\mathbb{R}^d)$-积分严格正定，则 *Langevin KSD* 将 P 与 $\{Q \in P^{\mathcal{H}_{k_p}} \cap P_{K,0} : (s_p - s_q)q \in L^2(\mathbb{R}^d)\}$ 分离。*2.* 若 $Q \in P^{\mathcal{H}_{k_p}}$, $(s_p - s_q)q \in L^2(\mathbb{R}^d)$ 且 $H_K \subset L^2(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$，则 $Q \in P_{K,0}$。*3.* 若 $H_K \subset L^2(\mathbb{R}^d)$， $\partial H_K \subset L^\infty(\mathbb{R}^d)$，且 $s_q q \in L^2(\mathbb{R}^d)$，则 $Q \in P^{\mathcal{H}_{k_p}}$。

尽管 $\mathcal{L}^2$-积分严格正定要求有些严格（例如，它排除了衰减缓慢的反多二次核；参见例2.13），但它确实能对某些 Q 实现分离，而这些情况无法通过基于得分的分离来处理（例如，定理 4.5适用于柯西 Q 和高斯分布 P，而定理 4.3 则不适用，因为 $Q \notin \mathcal{P}^{s_p}$）。

人们或许会期望，在适当假设下，Langevin KSD 能够将 P 与 所有 备择分布 $Q \in \mathcal{P}$区分开来。要建立这种一般性分离结果，主要障碍在于 Langevin Stein 核 k_p无界；对于尾部足够厚的 Q，核均值嵌入 $\int k_p(\cdot\, x)\mathrm{d}Q(x)$ 将无法存在，且 $Q \notin \mathcal{H}^{k_p}$。一种直接的解决办法是修改 Stein 核，使其 有界，也就是说 $Q \in \mathcal{P}^{k_p}$ 对所有 $Q \in \mathcal{P}$都成立，但这样一来，分离性是否仍然成立便不明确了。定理4.7给出了肯定的答案，这也是我们的第三个主要分离结果。

定理 4.7 背后的想法是证明 $\mathcal{H}^{k_p}$ 包含一个由有界函数构成的子再生核希尔伯特空间，其中的函数足够丰富，可以将 P 与其他情况 $Q \in \mathcal{P}$区分开来。简而言之，如果 Stein 差异或 Stein 核能够 P 区分与 $P\,\mathcal{P}$，我们就称其具有。

定义4.11（P-区分与有界 P-区分；Barp 等人，2024 的定义3）。函数集 F 是有界 P-区分的，当且仅当 $\mathrm{L}^\infty \cap \mathrm{F}$ 是 P-区分的；也就是说，若对于所有 $h \in \mathrm{L}^\infty \cap \mathrm{F}$， $Q \in \mathrm{P}$ 且 $Q(h) = P(h)$，则 $Q = P$。

论证的第一步是获得关于基核 k 的条件，使得 Stein 核 k_p 在（4.8）中有界。

定理 *4.6*（有界 *Langevin Stein* 核；*Barp* 等人，*2024* 年的定理 *7*）。设矩阵值核 K 具有 $H_K \subset C_b^1(\mathbb{R}^d)$，且为 $D_{L^1}^1(\mathbb{R}^d)$-特征的。若 $\|s_p(x)\| \leq \theta(x)$ 对于 $\theta \in C^1(\mathbb{R}^d)$ 成立，其中 $\frac{1}{\theta} \in C_b^1(\mathbb{R}^d)$，则由倾斜后的基核诱导的（*4.8*）中的 *Stein* 核 k_p

$$\frac{K(x,y)}{\theta(x)\theta(y)} \tag{4.10}$$

有界且 P-分离。

等价地，可以将与（4.10）中的倾斜基核相关联的 Stein 差异视为扩散 Stein 算子（参见第 3.2 节），其协方差系数为 $a(x) = \theta(x)^{-1} I_{d\times d}$，并将其作用于未倾斜的基核 K。

接下来，我们将证明，基于标准平移不变核的再生核希尔伯特空间包含具有有界核、形式如（4.10）的子再生核希尔伯特空间：

命题 4.8（平移不变核具有快速衰减的子再生核希尔伯特空间；Barp 等人等人，2024年的定理 8）。 设核 k 对应的 $\mathcal{H}_k \subset C^1(\mathbb{R}^d)$ 是平移不变的，且其谱密度在紧集上有正的下界。那么，存在一个平移不变的 $\mathcal{D}_{L^1}^1$-特征核 $k_s \in C^{(1,1)}(\mathbb{R}^d)$，并且对每个 $c > 0$，存在一个正定函数 f，_满足 $\frac{1}{f} \in C^1(\mathbb{R}^d)$ 且

$$\max(|f(x)|, \|\partial f(x)\|) = O(e^{-c\sum_{i=1}^d \sqrt{|x_i|}})$$

使得 $H^{k_f} \subset H_k$ 对于 $k_f(x,y) \triangleq f(x)k_s(x,y)f(y)$。

结合定理 4.6 和命题 4.8 可得定理 4.7，这是我们的第三个主要分离结果。若函数 f 满足至多根指数增长，则

$$f(x) = O\left(\exp\left(c\sum_{i=1}^d \sqrt{|x_i|}\right)\right) \tag{4.11}$$

对于某个 $c \in \mathbb{R}$。当 $K = kI$ 时，我们还引入简写 $\mathcal{P}_{k,\,0}$ 表示 $\mathcal{P}_{K,\,0}$。

定理 *4.7*（*Langevin KSD* 下的一般 P-分离性；*Barp* 等人，*2024* 的定理 *9*）。设核 k *with* $H_k \subset C^1(\mathbb{R}^d)$ 为平移不变核，且其谱密度在紧集上有正下界。对每个严格正的 $f \in C^1(\mathbb{R}^d)$，定义倾斜核 $k_f(x,y) \triangleq f(x)k(x,y)f(y)$。

(a) 若 $P \in \mathcal{P}_{k,0}$ 且 $\|s_p\|$ 至多具有根指数增长，则由基核 k 诱导的 *Stein* 核是有界的 P-分离核。(b) 若 $P \in \mathcal{P}_{k_f,0}$ 且 f、∂f 和 $f\|s_p\|$ 至多具有根指数增长，则由 k_f 诱导的 *Stein* 核是有界的 P-分离核。

分离性是 Stein差异用于拟合优度检验时的主要要求（参见第 6.2.2 节），但对于其他应用，则需要收敛检测和控制这些更强的性质。因此，接下来我们将关注收敛检测（第 4.4.3 节）和控制（第 4.4.4 节）。

4.4.3 使用 Langevin KSD 检测收敛

在第4.2 节中，我们看到经典 Stein差异可以检测到 1-Wasserstein 度量下的收敛（命题 4.3）。Langevin 核 Stein 差异同样具有这种收敛检测能力，具体如下。在核 $K(x,y) = k(x,y)I$ 的特殊情形下，我们用 $\mathcal{G}_k$简写 $\mathcal{G}_K$。

命题 *4.9*（*Langevin KSD* 可检测收敛；*Gorham* 和 *Mackey*，*2017*年，命题 *9*）。假设 $\nabla \log p$ 是利普希茨的，且 $\nabla \log p \in L^2(P)$。那么，对于 $k \in C_b^{(2,2)}(\mathbb{R}^d)$，只要 $W_{\|\cdot\|_2}^1(Q_n, P) \to 0$，*Langevin KSD* 就满足 $KSD_P(Q_n) \to 0$。

与经典 Stein 差异（对于 Langevin Stein 算子，见命题 4.3；即 $a = I, c = 0, b = \frac{1}{2}\nabla \log p$）相比，命题 4.9 不要求 $\nabla \log p$ 的梯度有界。不过，命题 4.9 要求 $\nabla \log p \in L^2(P)$，而命题 4.3 只要求对某个 $\epsilon > 0$，$\nabla \log p \in L^{1+\epsilon}(P)$。

4.4.4 使用 Langevin KSD 控制收敛

接下来我们讨论逆向问题：Langevin 核 Stein 差异能否控制更常见的统计散度。我们主要关注控制 Dudley 度量（回顾一下，该度量可度量分布的弱收敛；参见定义 2.27），但也会讨论控制 Wasserstein 度量的可能性（与经典 Stein 差异类似；参见定理 4.1）。

作为一种实用的简写，我们称 Langevin KSD控制弱 P收敛，如果 $KSD_P(Q_n) \to 0$ 蕴含 $BL_{\|\cdot\|}(P, Q_n) \to 0$。紧性通常是证明概率测度序列弱收敛的必要条件：

定义4.12（紧性）。如果对每个 $\epsilon > 0$，都存在一个紧集 $S \subset \mathbb{R}^d$，使得 $Q_n(S^c) \leq \epsilon$ 对所有$n \in \mathbb{N}$均成立，则称概率分布序列 $(Q_n)_{n\in\mathbb{N}} \subset \mathcal{P}(\mathbb{R}^d)$ 是紧的。

直观地说，一个紧序列 $(Q_n)_{n\in\mathbb{N}}$ 不可能出现概率质量“逃逸到无穷远”的情况。

定理 4.8（使用 Langevin KSD 控制紧收敛；Barp 等人，2024 的定理 7 和 9）。在定理 *4.6* 或 *4.7* 的条件下，如果 $(Q_n)_{n\in\mathbb{N}}$ 是紧的，那么 核 Stein 差异$_P(Q_n) \to 0$意味着 有界 Lipschitz$_{\|\cdot\|}(P, Q_n) \to 0$。

因此，问题转化为理解何时 Langevin KSD 的收敛意味着$(Q_n)_{n\in\mathbb{N}}$的紧性。回顾式（4.2），核 Stein 差异$_P(Q_n)$ 可表示为最大均值差异$_{k_p}(P, \cdot)$，其中使用了式（4.8）中的 Stein 核 k_p。此外，由式（2.11）可得，最大均值差异可表示为

$$\mathrm{MMD}_{k_p}(P,Q) = \sup_{\|f\|_{\mathcal{H}_{k_p}}\leq 1} P(f) - Q(f)$$

因此，再生核希尔伯特空间 $\mathcal{H}^{k_p}$ 决定了 Langevin KSD 所诱导的拓扑。检验函数集合越丰富，对应的统计散度就越严格；若要从 Langevin KSD 的收敛推导出 $(Q_n)_{n\in\mathbb{N}}$ 的紧性，则$\mathcal{H}^{k_p}$必须满足以下条件：

定义 4.13 (P-支配示性函数；Barp 等人，2024年第4号定义。函数集合$\mathcal{F} \subset \mathcal{L}^1(P)$ 若满足以下条件，则称其为 P-支配示性函数：对于每个 $\epsilon > 0$，都存在一个紧集$S \subset \mathbb{R}^d$ 和一个函数 $h \in \mathcal{F}$ 满足

$$h(x) - P(h) \geq \mathbb{I}_{S^c}(x) - \epsilon \tag{4.12}$$

对所有 $x \in \mathbb{R}^d$。

定义 4.13 确保序列$(Q_n)_{n\in\mathbb{N}}$ 只能在以下意义下良好地逼近 P：如果它在紧集$\mathcal{F}$ 之外放置的质量一致地很小S。将定理 4.8 与定义 4.13 结合起来，可为弱P-收敛控制提供充分条件：

推论 4.9（利用 P-收敛控制；Barp 等人的推论 3，Barp 等人，2024年）。在定理 *4.3*、定理 *4.6* 或 *4.7* 的条件下，如果 $\mathcal{H}^{k_p}$ 也 P- 支配示性函数，那么 ***Langevin*** KSD 控制弱 P- 收敛。

剩下的工作是将 P-支配示性函数这一性质与 Langevin KSD 中核 K 的选择联系起来。作为一个有用的铺垫，我们回顾一下强制的 函数的定义：

定义 4.14（强制函数；Hodgkinson 等人，2020年）。函数 $h: \mathbb{R}^d \to \mathbb{R}$ 是强制的，如果对于任意 $M > 0$，都存在一个紧集 $S \subset \mathbb{R}^d$，使得 $\inf_{x\in S^c} h(x) > M$。

注意，任何连续的强制函数也都有下界，因为连续函数在任何紧集上都是有界的。

引理 4.10 (强制函数支配指标函数；Barp 等人，2024的引理1）。若存在一个元素 $h \in \mathcal{H}^{k_p}$，它是强制的且有下界，并且 $P \in \mathcal{P}^{\mathcal{H}_{k_p}}$，那么$\mathcal{H}^{k_p}P$支配指标函数。

回顾一下，$\mathcal{H}^{k_p}$中的元素具有如下形式：$\mathcal{T}_P h = \nabla \cdot h + s_p \cdot h$，其中 $h \in \mathcal{H}_K$。对于大多数常见的 K 取值，第一项 $\nabla \cdot h$将是有界的。另一方面，对于 K尾部足够厚且适当递增的 s_p，第二项 $s_p \cdot h$ 可以是强制的且有下界。事实证明，逆多二次核（例2.13）满足这一要求：

定理 4.11 （IMQ KSD 控制弱 P收敛；Barp 等人定理 11，Barp 等人，2024）。考虑目标测度 $P \in \mathcal{P}$，其得分函数为 $s_p \in C(\mathbb{R}^d) \cap \mathcal{L}^1(P)$，并假设对于某个耗散率 $u > 1/2$ 以及 c， c_1，$c_2 > 0$， P 满足广义耗散性条件

$$-\langle s_p(x), x\rangle - c\left\|s_p(x)\right\|_1 \geq c_1 \left\|x\right\|_2^{2u} - c_2 \quad for\ all \quad x \in \mathbb{R}^d. \tag{4.13}$$

若 k 是逆多二次核（例 2.13），其指数为 $\beta \in (0,\ 2u-1)$，则$\mathcal{H}_{k_p}$ P支配指示函数。若此外， $\|s_p\|$ 按照 （4.11）的定义至多呈根指数增长，则 *Langevin* 核 Stein 差异 控制弱 P收敛。

定理 4.11 利用一个可能无界的 Stein 核 k_p 实现了弱收敛控制。在核方法的分析和应用中，使用有界核往往很有帮助；事实证明，也可以构造有界的 Stein 核。此外，我们在定理 4.12 中构造的有界 Stein 核能够精确度量到 P的弱收敛，这意味着 核 Stein 差异$_P$（Q_n）$\to 0$当且仅当 有界 Lipschitz$_{\|\cdot\|}$（P，Q_n）$\to 0$。为获得这一更强的结果，我们付出的代价是基核 K 不再具有平移不变性，这意味着需要选择一个合理的原点：

定理 4.12 （用有界 Stein 核度量弱 P-收敛；Barp 等人，2024）。考虑目标测度 $P \in \boldsymbol{P}$，其得分函数为 s_p，对于某个耗散率 $u > 1/2$ 以及 c、c_1、$c_2 > 0$，满足广义耗散条件 (4.13)。定义具有对角基核 K 的 *Stein* 核，使得 $K_{i,i}(x, y) = a(\|x\|_2)(x_i y_i + k(x, y))a(\|y\|_2)$，即，

$$k_p(x, y) = \sum_{1\leq i\leq d} \frac{\mathrm{d}x\ \mathrm{d}y(p(x)a(\|x\|_2)(x_i y_i + k(x, y))a(\|y\|_2)p(y))}{p(x)p(y)},$$

对于 k 在 $\boldsymbol{D}^1_{L^1}$ 上具有特征性，且 $\boldsymbol{H}_k \subset C_0^1(\mathbb{R}^d)$ 并且 $a(\|x\|_2) \triangleq (\sigma^2 + \|x\|_2^2)^{-\gamma}$ 为倾斜函数，满足 $\sigma > 0$ 且 $\gamma \leq u$。以下陈述成立：

(a) 若 $P \in \boldsymbol{P}_{K,0}$，则 $\boldsymbol{H}_{k_p}$ P-控制指标函数。

(b) 若 $P \in \boldsymbol{P}_{K,\ 0}$、$\gamma \geq 0$ 和 $\|s_p(x)\|_2 \leq (\sigma^2 + \|x\|_2^2)^{\gamma}$，则 k_p 是有界的 P-分离函数，并控制弱 P-收敛。

(c) 若 $\|s_p(x)\|_2 \cdot \|x\|_2 \leq (\sigma^2 + \|x\|_2^2)^{\gamma}$ 且 $s_p \in C(\mathbb{R}^d)$，则 $\boldsymbol{H}_{k_p} \subset C(\mathbb{R}^d)$ 和 k_p 可度量弱 P-收敛；也就是说，核 Stein 差异$_P(Q_n) \to 0$ 当且仅当有界 Lipschitz$_{\|\cdot\|}(P, Q_n) \to 0$。

这里讨论的是如何利用 Langevin 核 Stein 差异控制弱收敛，但也存在更强的收敛概念，人们可能希望了解何时可以利用核 Stein 差异对这些概念进行控制。 Kanagawa 等人 [2025] 的贡献在于，为 P 和 K 建立了充分条件，使得 Langevin 核 Stein 差异能够按照 Wasserstein 度量的意义控制收敛（参见定义 2.28）；也就是说，核 Stein 差异$_P($ $Q_n) \to 0$意味着 $\mathrm{W}^s_{\|\cdot\|}(P, Q_n) \to 0$。在 s-Wasserstein 度量下收敛，等价于弱收敛加上直到 s 阶矩的收敛（当 $\mathcal{X} = \mathbb{R}^d$时），因此显然能够得到更强的结论。

4.4.5 约束核Stein差异

现在考虑一个受约束域 $\mathcal{X} \subset \mathbb{R}^d$ 其边界为 $\partial\mathcal{X}$ 且 $n(x)$ 是 $\partial\mathcal{X}$ 在 $x \in \partial\mathcal{X}$处的外向单位法向量。选择一个在边界上消失的核，即$K(\cdot\ x)n(x) = 0$ 对所有 $x \in \partial\mathcal{X}$成立，也能确保核 Stein 集 g中的每个元素 $\mathcal{G}_K$都以式（3.12）所述的意义在边界上消失。这为确保

$$\oint_{\partial\mathcal{X}} p(x)g(x) \cdot \mathrm{n}(x)\ \mathrm{d}x$$

从而有 $P(\mathcal{T}_P g) = 0$根据散度定理 3.2成立。

4.4.6 一般情形下的核 Stein 差异

核 Stein 集在计算上十分便利，无需对 (4.2) 中的上确界进行数值计算。前几节重点讨论了 Langevin Stein 算子与核 Stein 集的结合，但还有其他几种 Stein 算子可供选择（参见第 3 章），它们也可以与核 Stein 集结合。例如，扩散核斯坦差异将扩散 Stein 算子（定义 3.5）与核 Stein 集结合，由 Barp 等人 [2019], 提出，并在其中给出了分离性的充分条件；Kanagawa 等人 [2025], 进一步对其进行了分析，并给出了 Wasserstein 收敛控制的条件。平方扩散核斯坦差异也应用并分析于因果模型学习的场景中，称为核偏离平稳性 [Lorch et al., 2024, Bleile et al., 2026]。Shi 等人 [2022a], 研究了镜像Stein算子（定义 3.6）与核 Stein 集的结合，并给出了弱收敛控制的充分条件。Fisher 和 Oates [2023], 将无梯度 Stein 算子（定义 3.7）与核 Stein 集配对，并确立了收敛检测和控制的条件。Yang 等人 [2018], 考察了离散 Stein 算子（第 3.5 节）与核 Stein 集的结合，并给出了分离性的充分条件。此外，研究还将 KSD 扩展到了更复杂的定义域，例如高维欧几里得空间 [Gong 等人， 2021], 序列空间 [Baum 等人， 2023],黎曼流形 [Xu 和 Matsuda, 2020年, Barp 等人，2022], 以及无限维函数空间 [Wynne 和 Duncan，2022] 。

4.5 随机特征 Stein 差异

图 Stein 差异和核 Stein 差异的一个潜在缺点是，它们的计算成本至少会随样本量呈二次增长 [Gorham 和 Mackey，2017]。为了解决这一局限，Huggins 和 Mackey [2018] 开发了一类 Stein 差异，可通过随机抽样以较低成本准确近似。其思路是考虑一种所谓的特征Stein 集：

定义 *4.15*（特征 *Stein* 集；*Huggins* 和 *Mackey*，*2018*年）。考虑一个特征函数Φ：$\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$，它对某个 $r \in [1, \infty)$ 以及所有 $x, z \in \mathbb{R}^d$，满足 $\Phi(x, \cdot) \in L^r(\mathbb{R}^d)$ 和$\Phi(\cdot\ z) \in C^1$。我们将相应的特征 *Stein* 集定义为

$$\mathcal{G}_{\Phi,r} \triangleq \left\{ g : \mathbb{R}^d \to \mathbb{R} \middle| g_i(x) = \int \Phi(x,z)\overline{f_i(z)}\,\mathrm{d}z \quad with \right.$$
$$\left. \sum_{i=1}^{d} \|f_i\|^2_{\mathcal{L}^s(\mathbb{R}^d)} \leq 1 \ for\ s = \frac{r}{r-1} \right\}.$$

自然地，与特征 Stein 集相关联的 Stein 差异称为特征 Stein 差异：

定义 4.16（特征 Stein 差异；Huggins 和 Mackey，2018年）。设$\mathcal{X}$ 是 $\mathbb{R}^d$ 的凸子集，且 P 在 $p > 0$ 上具有密度$\mathcal{X}$。若$\mathcal{G}_{\Phi,\ r}$ 是一个特征 *Stein* 集（定义*4.15*），且$\mathcal{T}_P$ 是 P的 *Langevin Stein* 算子（定义*3.2*），则称

$$\Phi\mathrm{SD}_{\Phi,r}(\cdot, P) \triangleq \mathcal{S}(\cdot, \mathcal{T}_P, \mathcal{G}_{\Phi,r})$$

一个特征 Stein 差异。

在本节的表述中，将 Langevin Stein 算子分解为一个和，较为方便

$$(\mathcal{T}_P g)(x) = \sum_{i=1}^{d} \partial_{x_i} g(x) + g_i(x)\partial_{x_i}(\log p)(x) = \sum_{i=1}^{d} (\mathcal{T}_i g)(x) \tag{4.14}$$

其中 $\mathcal{T}_i$ 用于表示 $\mathcal{T}_P$ 的第 i 个坐标，对（固定的） P 的依赖则保持隐含。对于二元函数 $\Phi(\cdot,\cdot)$，我们令 $\mathcal{T}_i\Phi$ 表示 $\mathcal{T}_i$ 作用于 Φ 的第一个自变量的动作。

命题 4.10 (ΦSD 显式形式；Huggins 和 Mackey，2018年)。在定义 *4.16*的设定下，

$$\Phi\mathrm{SD}^2_{\Phi,r}(Q, P) = \sum_{i=1}^{d} \|Q(\mathcal{T}_i\Phi)\|^2_{\mathcal{L}^r(\mathbb{R}^d)} \tag{4.15}$$

对所有 $Q \in P(X)$。

证明。 利用（4.14）中的记号，以及$\mathcal{G}_{\Phi\text{的定义},\ r}$ 在定义4.15中，

$$\begin{aligned}\Phi\mathrm{SD}^2_{\Phi,r}(Q,P) &\triangleq \sup_{g\in\mathcal{G}_{\Phi,r}} |Q(\mathcal{T}_P g)|^2 = \sup_{g\in\mathcal{G}_{\Phi,r}} \left|\sum_{i=1}^d Q(\mathcal{T}_i g_i)\right|^2 \\ &= \sup_{f:v_i=\|f_i\|_{\mathcal{L}^s(\mathbb{R}^d)},\|v\|_2\le 1} \left|\sum_{i=1}^d \int Q(\mathcal{T}_i\Phi)(z)\overline{f_i(z)}\,\mathrm{d}z\right|^2 \\ &= \sup_{v:\|v\|_2\le 1} \left|\sum_{i=1}^d \|Q(\mathcal{T}_i\Phi)\|_{\mathcal{L}^r(\mathbb{R}^d)}\, v_i\right|^2 = \sum_{i=1}^d \|Q(\mathcal{T}_i\Phi)\|^2_{\mathcal{L}^r(\mathbb{R}^d)},\end{aligned}$$

如所要求。 □

此时尚未实现计算成本的降低；例如当$r=2$时， ΦSD 只不过是以 $K(x,y)=k(x,y)I_{d\times d}$ 为基核的 Langevin KSD，其中$k(x,y)=\int\Phi(x,z)\Phi(y,z)\,\mathrm{d}z$，因此它具有二次计算成本。事实上，

$$\begin{aligned}\mathrm{KSD}^2_P(Q) &= \sum_{i=1}^d (Q\times Q)((\mathcal{T}_i\otimes\mathcal{T}_i)k) \\ &= \sum_{i=1}^d \|Q(\mathcal{T}_i\Phi)\|^2_{\mathcal{L}^2(\mathbb{R}^d)} = \Phi\mathrm{SD}_{\Phi,2}(Q,P)^2.\end{aligned}$$

为降低计算成本，需要某种形式的近似。使用重要性采样来近似范数项 $\|Q(\mathcal{T}_i\Phi)\|_{\mathcal{L}^r(\mathbb{R}^d)}$，可得到特征 Stein 差异的随机化近似：

定义 4.17 （随机特征 Stein 差异；Huggins 和 Mackey，2018年）。 对于目标特征 *Stein* 差异 $\Phi\mathrm{SD}_{\Phi,\ r}$ 以及在 $\mathbb{R}^d$ 上具有（勒贝格）密度 ν 的重要性采样分布，我们定义随机特征 Stein 差异

$$\mathrm{R}\Phi\mathrm{SD}^2_{\Phi,r,\nu,M}(Q,P) \triangleq \sum_{i=1}^d \left(\frac{1}{m}\sum_{j=1}^M \frac{|Q(\mathcal{T}_i\Phi)(Z_j)|^r}{\nu(Z_j)}\right)^{2/r}$$

其中 $Z_1,\ldots,Z_M \overset{\text{独立同分布}}{\sim} \nu$ 且 $Q\in P(X)$。

对于 $Q=\sum_{i=1}^n w_i\delta_{x_i}$ 有限支撑的情形，RΦSD 的计算可简化为计算 mnd缩放后的随机特征，$(\mathcal{T}_i\Phi)(x_i,\ Z_j)/\nu(Z_j)^{1/r}$，而且可以并行执行这项计算。

4.5.1 RΦSD 的特殊情形

可以直接使用定义 4.17 中的 RΦSD 框架，构造标准 KSD 的低成本近似。例如，当 KSD 基核 $k(x,y) = \Psi(x-y)$ 对某个 $\Psi \in L^2(\mathbb{R}^d)$ 成立时，可以选择 $r=2$, $\nu \propto F(\Psi)$，以及 $\Phi(x,z) = e^{-i\langle z,x\rangle} F(\Psi)(z)^{1/2}$，设计一种随机傅里叶特征 [Rahimi 和 Recht, 2007, RFF]近似 $R\Phi SD_{\Phi,2,\nu,M}$，以近似 KSD_k [Huggins 和 Mackey, 2018]。尽管这种 RFF-KSD 在实践中非常实用，但它确实存在一个已知局限：对于不可数个分布 $Q \neq P$，以高概率 [，该 $R\Phi SD_{\Phi,2,\nu,M}(Q,P) = 0$ [Chwialkowski 等人, 2015, 命题 1]。

为避免这一不理想的特性，Jitkrittum 等人 [2017] 开发了一种替代性的 RΦSD，称为随机有限集 Stein 差异（FSSD-rand）。按照定义 4.17 中的记号，FSSD-rand 的构造方式是选取 $r=2$、任意特征数量 M 以及重要性采样分布 ν，并令 $\Phi(x,z) = f(x,z)\nu(z)^{1/2}$，其中 f 是实解析且 C_0-通用的 [再生核（Carmeli 等人，2010，定义 4.1] ）。特别是，实解析性可确保对每个 $Q \neq P$，都有 $\mathbb{P}(R\Phi SD_{\Phi,2,\nu,M}(Q,P) > 0) = 1$ [（Jitkrittum 等人，2017年，定理 1]）。在第 4.5.2 节和第 4.5.3 节中，我们将看到，形式不同的 RΦSD 特征还具有强收敛确定性质；此外，选取 $r \neq 2$能够在显著缩小采样预算的情况下，获得可证明准确的近似 M。

4.5.2 选择特征函数 Φ

选择特征函数时， Φ 主要关注的是，不要“丢失”有助于将近似 Q_n 与目标 P进行比较的信息。命题 4.11 旨在确立 ΦSD 控制 Langevin KSD 的条件；Langevin KSD 的性质现已得到充分理解（参见第4.4.1节）。令$\mathcal{F}(f)$ 表示 f 的广义傅里叶变换： $\mathbb{R}^d \to \mathbb{C}$。由广义 Hölder 不等式和 Babenko–Beckner 不等式可得：

命题 4.11 （用 ΦSD控制 Langevin KSD；Huggins 和 Mackey，2018年）。 若 $r \in [1,2]$, $\rho \in \mathcal{L}^t(\mathbb{R}^d)$ 对于 $t = r/(2-r)$，且 $K(x,x') = k(x,x')I_{d\times d}$ 满足

$$k(x,y) = \int \mathcal{F}(\Phi(x,\cdot))(\omega)\overline{\mathcal{F}(\Phi(y,\cdot))(\omega)}\rho(\omega)\,\mathrm{d}\omega,$$

那么 $KSD_P^2(Q_n) \leq \|\rho\|_{\mathcal{L}^t(\mathbb{R}^d)} \Phi SD_{\Phi,r}^2(Q_n, P)$。

也就是说，对于上述形式的核 k， ΦSD 的收敛意味着 Langevin KSD 的收敛。选定特征函数后，我们现在可以着手使用所谓的 随机特征 [Rahimi 和 Recht， 2007]。在介绍使用随机特征进行收敛检测与控制的结果之前，我们先描述对基核的假设。

假设 *4.1*。基核的形式为 $k(x,y) = A_n(x)\Psi(x-y)A_n(y)$，其中 $\Psi \in C^2(\mathbb{R}^d)$、 $A \in C^1(\mathbb{R}^d)$ 和 $A_n(x) \triangleq A(x-m_n)$，且 $A > 0$, $\nabla \log A$ 有界且满足利普希茨条件，并且 $m_n \triangleq \int x\, dQ_n(x)$。

请注意，假设 4.1 允许基核 k 依赖于（通过序列索引n）正在评估其近似质量的分布 Q_n 。

假设 4.2。 假设 *4.1*成立，且 $\Phi(x,z) = A_n(x)F(x-z)$，其中 $F \in C^1(\mathbb{R}^d)$ 为正数，并且存在一个范数 $\|\cdot\|$ 和常数 $s, C_1 > 0$ 使得

$$|\partial_{x_d} \log F(x)| \leq C_1(1 + \|x\|^s), \quad \lim_{\|x\|\to\infty} (1 + \|x\|^s)F(x) = 0,$$

and

$$F(x-z) \leq C_1 \frac{F(z)}{F(x)}.$$

此外，存在一个常数 $C_2 \in (0,1]$ 以及连续非递增函数 f，使得 $C_2\, f(\|x\|) \leq F(x) \leq f(\|x\|)$。

下面的结果表明，在适当的倾斜的 1-Wasserstein距离（参见第 2.3.2节）中的收敛可由适当的 实数集ΦSD检测到：

定理 4.13 （用实数集ΦSD检测收敛；Huggins 和 Mackey，2018年）。 假设假设 *4.2*成立，且 $F \in \mathcal{L}^r(\mathbb{R}^d)$， $1/A$ 有界， $x \mapsto x/A(x)$ 利普希茨，且 $\mathbb{E}_P[A(Z)\, \|Z\|_2^2] < \infty$。则

$$\mathrm{W}^{1,A_n}_{\|\cdot\|_2}(Q_n, P) \to 0 \implies \begin{cases} \Phi\mathrm{SD}_{\Phi,r}(Q_n, P) \to 0 \\ \mathrm{R}\Phi\mathrm{SD}_{\Phi,r,\nu_n,M_n}(Q_n, P) \xrightarrow{P} 0 \end{cases}$$

对于 $r \in [1,2]$、 ν_n 和 $M_n \geq 1$ 的任意取值。

4.5.3 选择重要性采样分布 ν

选择重要性采样分布 ν 的动机在于使实数集ΦSD 接近其参考 ΦSD ，即使重要性样本量 M 很小。Huggins 和 Mackey所采用的策略是选择 ν ，使每个随机特征的二阶矩 $|Q$（$\mathcal{T}_i\Phi$）（Z）$|^r/\nu$（Z）都由其均值的某个幂所界定：

定义 *4.18*（（C， γ）二阶矩）。固定一个目标分布 P。对于 $Z \sim \nu$、 $i \in$*{1, . . . , d}* 和 $n \geq 1$，令 Y_n、$_i \triangleq w_i$（Z、Q_n） $\triangleq |$（$Q_n T_i\Phi$） （Z）$|^r/\nu$（Z）。如果对于某些 $C > 0$ 和 $\gamma \in [0,2]$，对于所有 $i \in$*{1, . . . , d}* 和 $n \geq 1$，都有 $\mathbb{E}[Y_{n,i}^2] \leq C\mathbb{E}[Y_{n,i}]^{2-\gamma}$，那么我们称（$\Phi$、 r、 ν）为 P 和 Q_n 产生（C、 γ）二阶矩。

定理 *4.14*（利用 *RΦSD* 控制 *Langevin KSD*；*Huggins* 和 *Mackey*，*2018*年的命题 *3.6*）。假设 *(Φ, r, ν)* 为 P *and* Q_n产生 *(C, γ)* 阶矩。如果对于所有 $i \in$*{1, . . . , d}*， $M \geq 2C\mathbb{E}[Y_{n,i}]^{-\gamma}$ *log*$(d/\delta)/\epsilon^2$，那么，以至少 $1-\delta$ 的概率，

$$\mathrm{R}\Phi\mathrm{SD}_{\Phi,r,\nu,M}(Q_n, P) \geq (1-\epsilon)^{1/r}\, \Phi\mathrm{SD}_{\Phi,r}(Q_n, P).$$

在进一步假设命题 *4.11*成立的情况下，如果参考 核 Stein 差异$_P(Q_n) \geq c_1 n^{-1/2}$ 对于某个 c_1，那么存在 $c_2 > 0$ 使得样本量

$$M \geq \frac{c_2 C n^{\gamma r/2} \|\rho\|^{\gamma r/2}_{\mathcal{L}^t(\mathbb{R}^d)} \log(d/\delta)}{\epsilon^2}$$

只需以至少 $1-\delta$ 的概率满足，

$$\|\rho\|^{1/2}_{\mathcal{L}^t(\mathbb{R}^d)} \mathrm{R\Phi SD}_{\Phi,r,\nu,M}(Q_n, P) \geq (1-\epsilon)^{1/r} \mathrm{KSD}_P(Q_n).$$

在若干情形下，核 Stein 差异$_P(Q_n) \gtrsim n^{-1/2}$ 成立；最值得注意的是状态$\{x_i\}_{i=1}^n$ 相互独立且同分布时。这使得实数集ΦSD 适用于拟合优度检验（参见第 6.2.2节），例如，在此情境下，人们关注独立同分布原假设下 Stein 差异检验统计量的性能。较小的 r 会显著降低样本复杂度 $M =\Omega(n^{\gamma r/2})$。例如，如果权函数有界（即 $\gamma = 1$），则只需选择 $r = 1$ 和 $M =\Omega(n^{1/2})$。接下来，我们将说明如何选择 ν，使其 γ 任意接近 0。

为了利用定理 4.14，还需确定何时 (Φ, r, ν) 可为 (C, γ) 提供二阶矩 P 和 Q_n。

假设 4.3 (远距耗散性； Eberle，2015年，Gorham 等人，2019)。 分布 P具有 ∇ 对数 p 利普希茨性，并且存在 $\kappa > 0$ 和 $r \geq 0$，使得 $(\kappa,\ r)$-远距耗散性 条件

$$\langle \nabla \log p(x) - \nabla \log p(y), x - y \rangle \leq -\kappa \|x - y\|_2^2 + r, \qquad \forall x, y \in \mathbb{R}^d$$

得到满足。

下面给出简单的条件以及 ν 的一种选择，在这些条件下可保证 $(C,\ 1)$ 阶矩：

命题 *4.12*（*Huggins* 和 *Mackey*，*2018* 年的命题*3.7*）。假设假设 *4.1* 和 *4.2* 在 $s = 0$ 下成立，假设 *4.3* 成立，并且存在常数 $C' > 0$，使得对所有 $n \geq 1$、$Q_n([1 + \|\cdot\|]A_n) \leq C'$。如果 $\nu(z) \propto Q_n([1 + \|\cdot\|]\Phi(\cdot - z))$，那么对于任意 $r \geq 1$，(Φ, r, ν) 可为 P 和 Q_n 产生 $(C, 1)$ 阶矩。

为了获得 $\gamma < 1$ 的 (C, γ) 阶矩，我们将选择 ν，使得 $w_i(z, Q_n)$ 随着 $\|z\| \to \infty$ 增大而足够快地衰减。为此，需要满足两个涉及 Ψ 和 F 的傅里叶变换的可积性条件：

假设*4.4*。假设*4.1*和*4.2*成立，$\omega_1^2 F(\Psi)^{1/2}(\omega) \in L^1(\mathbb{R}^d)$，并且对于 $t = r/(2-r)$，$F(\Psi)/F(F)^2 \in L^t(\mathbb{R}^d)$。

该$\mathcal{L}^1$ 条件较弱，而$\mathcal{L}^t$ 条件确保命题 4.11 适用于我们选定的 ΦSD。定理 4.15 表明，通过提高 γ 的过度离散参数RΦSD 的光滑性 λ，以及降低 F 的过度离散参数 ξ，可以提高 ν 的重要性样本增长率。

定理 4.15 (Huggins 和 Mackey的，2018年年定理 3.8)。 设假设 *4.1* 至 *4.4* 成立，并假设存在 $\mathcal{C} > 0$ 使得，

$$Q_n\left([1 + \|\cdot\| + \|\cdot - m_n\|^s]A_n/F(\cdot - m_n)\right) \leq \mathcal{C} \quad for\ all \quad n \geq 1.$$

那么，存在一个常数 $b \in [0, 1)$，使得以下结论成立。对于任意 $\xi \in (0、 1-b)、 c > 0$ 和 $\alpha > 2($ $1-\lambda)$，若 $\nu(z) \overline{\geq} c\,\Psi(z - m_n)^{\xi r}$，则存在一个常数 $C_\alpha > 0$，使得$(\Phi、 r、 \nu)$ 对 P 和 Q_n具有$(C_\alpha、$ $\gamma_\alpha)$ 二阶矩，其中 $\gamma_\alpha \triangleq \alpha + (2-\alpha)\xi/(2-b-\xi)$。

4.6 随机 Stein 差异

我们考虑的最后一类 Stein 差异，是为处理以下情形而提出的：目标分布 P 的密度是

$$p(x) = \prod_{l=1}^{L} p_l(x)$$

大量因子 $p_l(x)$ 的乘积，而这些因子各自不必积分为 1。这种结构出现在高维数据情境中，其中 x 表示统计模型的参数；给定该参数后，数据被视为相互独立。在这种情况下，使用 Langevin Stein 差异的主要挑战在于，计算梯度 $\nabla \log p$ 需要对 L 个独立项求和；对于离散的 $Q_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$，这会产生 $O(nL)$ 级的计算成本，计算开销可能相当可观。然而，这与机器学习中的经验风险最小化所遇到的问题相同，而针对该问题，人们已提出基于子采样的有效计算方法。随机 *Stein* 差异也旨在利用子采样，以较低的计算成本提供有意义的差异。其主要思路是考虑形如$\mathcal{T}_P = \sum_{l=1}^{L} \mathcal{T}_l$ 的可分解 Stein 算子，其中$\mathcal{T}_l$ 是以 p_l为目标的 Stein 算子，不过一般而言， p_l 不必是归一化概率分布。接着，对于子集 $\sigma \subseteq [L]$, 我们将子集算子记作$\mathcal{T}_\sigma \triangleq \sum_{l\in\sigma} \mathcal{T}_l$，其计算成本与子集大小 $|\sigma|$ 呈线性关系。

定义4.19 （随机 Stein 差异；Gorham 等人，2020年）。 考虑形式如下的可分解 *Stein* 算子 $\mathcal{T}_P = \sum_{l=1}^{L} \mathcal{T}_l$，并且 $Q_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$。固定批量大小 m，对于每个 $i \in [n]$，独立地从 σ_i 中均匀随机选取大小为 m 的子集 $[L]$。那么，对于定义域包含于各个 $_i$的任意 *Stein*集合 $\mathcal{G}$， 我们将随机 Stein 差异定义为如下随机量

$$\mathcal{SS}(Q_n, \mathcal{T}_P, \mathcal{G}) \triangleq \sup_{g\in\mathcal{G}} \left| \frac{1}{n}\sum_{i=1}^{n} \frac{L}{m} (\mathcal{T}_{\sigma_i} g)(x_i) \right|. \tag{4.16}$$

虽然随机 Stein 差异（SSD）在批次大小$\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G})$ 等于 m 时，与标准 Stein 差异 L 一致，但我们并不将 SSD 视为对 Stein 差异的近似。相反，SSD 本身就是一种差异，并附带（概率意义上的）收敛检测（第 4.6.1 节）和收敛控制（第 4.6.2 节）保证。

4.6.1 使用 SSD 检测收敛

以下定义允许 Stein 集合序列$\mathcal{G}_n$ 随时间演化，以适配 Gorham 和 Mackey [2015], Gorham 等人 [2019]的图 Stein 集。令 $\binom{[L]}{m} \triangleq \{\sigma \subseteq [L] : |\sigma| = m\}$表示所有大小为 m 的 $[L]$子集。

定理 *4.16*（*SSD* 检测收敛；*Gorham* 等，*2020* 年的定理 *2*）。设 X 是 $\mathbb{R}^d$ 的凸子集。设 $G = \cup_{n\geq 1} G_n$。考虑如下形式的可分解 *Stein* 算子 $T_P = \sum_{l=1}^{L} T_l$，其中 $P(T_P G) = \{0\}$。假设对于某个 a，$c > 0$ 且对每个 $\sigma \in \binom{[L]}{m}$，$T_\sigma G \subset C(X)$，

$$\sup_{g\in\mathcal{G}} |(\mathcal{T}_\sigma g)(x)| \leq c(1 + \|x\|_2^a)$$
$$\sup_{g\in\mathcal{G}} \sup_{x,y\in K} \frac{|(\mathcal{T}_\sigma g)(x) - (\mathcal{T}_\sigma g)(y)|}{\|x - y\|_2} < \infty$$

对于每个紧集 K。若 $\mathrm{W}^a_{\|\cdot\|_2}(Q_n, P) \to 0$，则 $SS(Q_n, T_P, G_n) \overset{a.s.}{\to} 0$.

定理 4.16 表明，如果算子 $\mathcal{T}_\sigma$ 生成的连续函数增长速度不超过多项式，且其导数在局部有界，那么 SSD 几乎必然能够检测 Wasserstein 收敛。

4.6.2 使用 SSD 控制收敛

首先，对每个 Stein 差异$\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_n)$，一个基于修正的 Stein 集的有界 *Stein* 差异

$$\mathcal{G}_{b,n} \triangleq \left\{ g \in \mathcal{G}_n : \|\mathcal{T}_\sigma g\|_\infty \leq 1 \forall \sigma \in \binom{[L]}{m} \right\} \tag{4.17}$$

其中每个 Stein 函数都被约束为在每个子集算子$\mathcal{T}_\sigma$下有界。

定理 *4.17*（有界 *Stein* 差异控制紧收敛；*Gorham* 等，*2020* 的定理 *3*）。设 $X = \mathbb{R}^d$。考虑满足假设 *4.3* 的 *Langevin Stein* 算子 T_P（参见定义 *3.2*）。假设对每个 $\sigma \in \binom{[L]}{m}$ 有 $\sup_{x\in\mathcal{X}} \|\nabla \log p_\sigma(x)\|_2/(1 + \|x\|_2) < \infty$，固定一个概率测度序列 $(Q_n)_{n=1}^\infty$，并考虑 *(4.17)* 中针对以下任一集合 G_n 的有界 *Stein* 集 $G_{b,n}$：

（***A.1***）$\mathcal{G}_n = \mathcal{G}_{\|\cdot\|}$，具有任意向量范数的经典 *Stein* 集 $\|\cdot\|$（定义 *4.4*）。

*(**A.2**)* $\mathcal{G}_n = \mathcal{G}_{\|\cdot\|,Q,G}$，带有任意向量范数 $\|\cdot\|$ 的图 *Stein* 集，以及一个有限图$G = (V, E)$，其顶点为 $V \subset \mathcal{X}$*(*定义 *4.6)*。

*(**A.3**)*$\mathcal{G}_n = \mathcal{G}_K$，具有 $K(x, y) = \Phi(x - y)I_{d\times d}$的核 *Stein* 集，适用于 $\Phi \in C^2(\mathbb{R}^d, \mathbb{R})$ 且其傅里叶变换不消失（定义 *4.8*）。

如果 $Q_n \not\Rightarrow P$，那么 $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_{b,n}) \not\to 0$ 或 $(Q_n)_{n=1}^{\infty}$ 不紧（参见定义 *4.12*）。

定理 4.18（核化 Stein discrepancy 控制有界 Stein 差异收敛；Gorham 等人 的2020年年论文中的定理 4）。若 $\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{b,n}) \not\to 0$，则以概率 1，$\mathcal{SS}(Q_n, \mathcal{T}, \mathcal{G}_n) \not\to 0$。

命题 4.13（强制型核化 Stein discrepancy 保证紧性；Gorham 等人 ，2020年）。若 $(Q_n)_{n=1}^{\infty}$ 不紧，且 $\frac{L}{m}\mathcal{T}_{\sigma}g$ 对某个 $g \in \bigcap_{n=1}^{\infty} \mathcal{G}_n$ 及所有$\sigma \in \binom{[L]}{m}$而言，均为强制的且有下界，则必有 $\mathcal{SS}(Q_n, \mathcal{T}, \mathcal{G}_n) \not\to 0$。

综合这些结果可知，在次采样算子满足标准耗散性和增长条件的情况下，配备 Langevin 算子的核化 Stein discrepancy 以及定理4.17 中的任一 Stein 集合都能以概率 1控制收敛：

定理 *4.19*（核化 *Stein discrepancy* 控制收敛；*Gorham* 等，*2020* 的定理 *6*）。在定理 *4.17* 的记号下，假设 $\nabla \log p$ 是利普希茨的，$\sup_{x\in\mathbb{R}^d} \|\nabla \log p_\sigma(x)\|_2 / (1+\|x\|_2) < \infty$对所有 $\sigma \in \binom{[L]}{m}$均成立，并且对于某些 $\kappa > 0$和 $r \geq 0$，对所有满足 $\sigma \in \binom{[L]}{m}$的 p_σ，均一致满足一个 (κ, r)-远距耗散条件（参见假设 *4.3*）。考虑径向函数$\Phi(x) \triangleq (1 + \|\Gamma x\|_2^2)^{\beta}$，其中 $\beta \in (-1, 0)$，且 Γ为任意正定矩阵。令 G_n 为定理 *4.17* 中任意一个 *Stein*集合。若 $Q_n \not\Rightarrow P$，则以概率 *1*，$SS(Q_n, T_P, G_n) \not\to 0$。

5 Stein动力学

前一章讨论了 Stein 差异这一具有出色计算特性的统计散度。本章将探讨通过最小化库尔贝克-莱布勒散度进行推断与学习，并借助质量传输阐明其与 Stein 方法之间的重要联系。

5.1 动态质量传输

假设我们希望将同一定义域 $\mathcal{X} \subseteq \mathbb{R}^d$ 上的一个分布 Q 转换为另一个分布 P。实现这种转换有两种不同的方法。

一种方法是从 Q 中进行模拟，并根据密度比 $r = \mathrm{d}P\,\mathrm{d}Q$ 为每个样本赋予权重，这与重要性采样、稀疏化和生灭过程等技术有关。然而，当 Q 与 P 之间的密度比 $\mathrm{d}P\,\mathrm{d}Q$ 的方差较大时，这种重加权方法的性能往往会下降。

另一种方法是基于传输的方法，它通过应用变换，将服从 Q 分布的随机变量转换为服从 P 分布的随机变量。这种方法能够利用 P 密度函数的梯度信息，因此有望避免重要性采样及相关方法所面临的性能下降问题。

具体而言，基于传输的方法旨在设计一个 传输映射 $T\colon \mathcal{X} \to \mathcal{X}$，使得只要 $Z \sim Q$，$T(Z)$ 就服从分布 P，记作 $P = T_\sharp Q$。假设 T 是可微双射，那么根据密度的变量替换公式，

$$p(x) = q(T^{-1}(x)) \left|\det(\nabla T^{-1}(x))\right|, \tag{5.1}$$

其中 q 和 p 分别是 Q 和 P的密度函数。不幸的是，以复杂的非线性方式依赖于 T（5.1）而基于（5.1）的实用方法在高维情况下可能计算成本高昂，因为其中涉及雅可比行列式。

这一困难可以通过采用动态视角加以规避，即将 T 视为连续时间动力系统的流映射。具体而言，我们假设从 Q 中抽取的随机变量Z_0 按照常微分方程连续演化

$$\frac{\mathrm{d}}{\mathrm{d}t} Z_t = v_t(Z_t), \qquad t \geq 0, \tag{5.2}$$

其中 $v\colon\ [0,\infty)\ \times \mathbb{R}^d \to \mathbb{R}^d$ 是一个待指定的连续可微速度场。从 $Z_0 \sim Q$出发，我们的目标是设计随时间变化的速度场 v_t ，使得 Z_t的分布（记为 Q_t）随着 $t \to \infty$收敛到目标分布 P。

假设常微分方程 (5.2) 对于每个 $\mathbb{R}^d$中的初始化都有唯一解，并且 q_t 的密度 Q_t 存在，且在 (t,x)上连续可微，同时 v_t 在 (t,x) 上也连续可微。在这些经典正则性条件下， Q_t的演化由连续性方程控制

$$\partial_t q_t(x) = -(\nabla \cdot (q_t v_t))(x). \tag{5.3}$$

注意，连续性方程可以用与 $\mathcal{T}_{Q_t}$ 相关联的 Langevin Stein 算子表示 q_t (参见定义 3.2)，即 $\partial_t \log q_t(x) = -(\mathcal{T}_{Q_t} v_t)(x)$。

(5.3) 的另一种形式是 刘维尔公式，它给出了$q_t(Z_t)$ 随 Z_t 按照 (5.2) 演化时的全时间导数：

$$\frac{\mathrm{d}}{\mathrm{d}t} \log q_t(Z_t) = -(\nabla \cdot v_t)(Z_t). \tag{5.4}$$

这可由链式法则推出：$\mathrm{d}\,\mathrm{d}t\, q_t(Z_t) = \partial_t q_t(Z_t) + v_t(Z_t) \cdot \nabla q_t(Z_t)$，再结合连续性方程 (5.3)。

5.2 Kullback–Leibler散度的梯度流

设计速度场 v_t 的一种方法，是沿着当前分布 Q_t 与目标 P之间某一合适散度的梯度。散度的选择至关重要，因为它必须导出一个在计算上易于求值或近似的系统。事实证明，Kullback–Leibler散度 $\mathrm{KL}(Q_t \| P)$ 是一个特别合适的候选，这要归功于下述结果，该结果刻画了它在 (5.3) 下与 Langevin Stein 算子相关的下降速率。

定理5.1. 假设微分方程 $Z_t = v_t(Z_t)$ 对 $\mathbb{R}^d$ 中的每一个初始化在$[0,T]$ 上都存在唯一解。设 Q_t 表示 Z_t 的分布，其密度为 q_t。假设v_t 与 $\log\, q_t$ 在 $[0,T] \times \mathbb{R}^d$ 上连续可微。设 P 是一个概率测度，具有正的且连续可微的密度 p，使得 $\mathrm{KL}(Q_0 \| P) < \infty$，并且$\int_0^T Q_t(|(\mathcal{T}_P v_t)(Z_t)|)\mathrm{d}t < \infty$，其中 $\mathcal{T}_P$是 P 的 ***Langevin Stein*** 算子。则

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathrm{KL}(Q_t \| P) = -\, Q_t(\mathcal{T}_P v_t). \tag{5.5}$$

证明。 由于 Z_t 服从分布 Q_t，我们有

$$\mathrm{KL}(Q_t \| P) = \mathbb{E}\left[\log q_t(Z_t) - \log p(Z_t)\right].$$

定义 $R_t = \log q_t(Z_t) - \log p(Z_t)$。其时间导数为

$$\begin{aligned}\dot{R}_t &= \frac{\mathrm{d}}{\mathrm{d}t}\log q_t(Z_t) - \frac{\mathrm{d}}{\mathrm{d}t}\log p(Z_t) \\ &= -\nabla\cdot v_t(Z_t) - v_t(Z_t)\cdot\nabla\log p(Z_t) \\ &= -\mathcal{T}_P v_t(Z_t),\end{aligned}$$

其中，第一项使用了刘维尔公式（5.4），第二项使用了链式法则，

$$\frac{\mathrm{d}}{\mathrm{d}t}\log p(Z_t) = \nabla\log p(Z_t)^\top \dot{Z}_t = \nabla\log p(Z_t)^\top v_t(Z_t).$$

因此，

$$\begin{aligned}\mathrm{KL}(Q_t\|P) - \mathrm{KL}(Q_0\|P) &= \mathbb{E}\left[R_t - R_0\right] \\ &= \mathbb{E}\left[\int_0^t \dot{R}_\tau\,\mathrm{d}\tau\right] \\ &= \int_0^t \mathbb{E}\left[\dot{R}_\tau\right]\,\mathrm{d}\tau \\ &= -\int_0^t \mathbb{E}\left[(\mathcal{T}_P v_\tau)(Z_\tau)\right]\,\mathrm{d}\tau,\end{aligned}$$

其中，假设$\int_0^t \mathbb{E}[|(\mathcal{T}_P v_\tau)(Z_\tau)|]\,\mathrm{d}\tau < \infty$，我们使用富比尼–托内利定理交换积分次序。对所得恒等式求导，并利用$t \mapsto \mathbb{E}\left[\mathcal{T}_P v_t(Z_t)\right]$的连续性，可得 $\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(Q_t\|P) = -\mathbb{E}\left[(\mathcal{T}_P v_t)(Z_t)\right],$，从而完成证明。 □

注 5.1. 我们可以通过泰勒近似，在离散时间下理解这个结果。对于测度 Q，令 $Q' \triangleq (\mathrm{Id}+\epsilon v)_\sharp Q$ 为给定 $Z\sim Q$时 $Z' \triangleq Z + \epsilon v(Z)$ 的分布。那么我们有

$$\begin{aligned}\frac{1}{\epsilon}(\mathrm{KL}((\mathrm{Id}+\epsilon v)_\sharp Q \parallel P) - \mathrm{KL}(Q\|P)) &\approx \frac{\mathrm{d}}{\mathrm{d}\epsilon}\mathrm{KL}((\mathrm{Id}+\epsilon v)_\sharp Q \parallel P)\Big|_{\epsilon=0} \\ &= -Q(\mathcal{T}_P v).\end{aligned}$$

因此，*Stein* 算子 $Q(\mathcal{T}_P v)$ 的期望刻画了沿向量场 v对 Q施加微小位移时，***Kullback–Leibler***散度 $\mathrm{KL}(Q\|P)$ 的下降速率。

5.3 梯度流与 Stein 差异

既然我们已经理解了 Kullback–Leibler散度与 Langevin Stein 算子之间的关系，接下来便要寻找一个速度场 v^*，使其能够高效地最小化 Kullback–Leibler 散度，从而实现计算高效的测度传输。这可以表述为以下优化问题

$$v_t^* \in \operatorname*{argmax}_{v_t\in\mathcal{G}_t}\left\{-\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(Q_t\|P)\right\}, \qquad \forall t\geq 0,$$

其中$\mathcal{G}_t$ 是候选速度场集合。利用定理5.1中的 Kullback–Leibler 散度导数公式,

$$v_t^* \in \operatorname*{argmax}_{v_t \in \mathcal{G}_t} Q_t(\mathcal{T}_P v_t), \qquad \forall t \geq 0. \tag{5.6}$$

因此，Kullback–Leibler散度的最大下降速率为 Q_t（$\mathcal{T}_P v_t^*$）在适当的正则性假设下，该速率与 Stein 差异的定义一致$\mathcal{S}$（Q_t，$\mathcal{T}_P$，$\mathcal{G}_t$）（参见定义 4.1）。因此，Stein 差异可以被视为质量传输下 Kullback–Leibler散度的最大下降速率，而用于质量传输的最优向量场 v_t^* 也是实现 Stein 差异的 Stein集合中的最优元素。

正如我们将在下文中展示的那样，如果将候选空间 G_t 定义为概率测度集合上某种几何结构的切空间，则具有 v_t^* 的动力学可以解释为 Kullback–Leibler散度 $\mathrm{KL}(Q_t \| P)$ 的一种梯度流。对候选速度集合 G_t 的不同选择，使我们能够推导出不同的质量传输动力学。其中包括朗之万扩散，它利用候选空间 $\mathsf{G}_t \subset \mathsf{L}^2(Q_t)$，使得 $\mathsf{S}(Q_t, \mathsf{T}_P, \mathsf{G}_t)$ 与 Fisher散度一致（参见命题 4.1）；以及 Stein变分梯度下降，其中 G_t 取为再生核希尔伯特空间（RKHS）中的一个球，且 $\mathsf{S}(Q_t, \mathsf{T}_P, \mathsf{G}_t)$ 与核 Stein 差异（KSD）一致（参见定义 4.8）。

注 5.2. 假设 Q_t 已达到平衡态，并收敛至 P。*Stein*恒等式 (1.1)现在等价于

$$P(\mathcal{T}_P v_t) = -\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(Q_t||P) = 0,$$

这反映了这样一个事实： P 是 *Kullback–Leibler*散度 $Q \mapsto \mathrm{KL}(Q||P)$的一个平稳点。

几何解释

可以将上述优化问题置于一个几何框架中，将其视为定义在概率测度空间上的 Kullback–Leibler散度泛函的梯度流，并为该空间配备适当的输运度量。这一观点与分布空间中的梯度流理论相呼应，例如 Ambrosio等人 [2005], 所发展的理论，不过这里我们仅在形式和启发式层面进行讨论。

假设允许的速度场构成某个赋范空间的单位球,

$$\mathcal{G}_t \triangleq \{v : \|v\|_t \leq 1\},$$

其中 $\|\cdot\|_t$ 是一个（ （可能随时间变化的）范数将在后文中指定。对于两个概率

分布 Q 和 Q'，我们定义一种广义运输成本

$$\mathcal{W}(Q,Q') \triangleq \inf_{(q_t,v_t)} \left\{ \left(\int_0^1 \|v_t\|_t^2 \,\mathrm{d}t \right)^{1/2} \ \text{s.t.} \right. \tag{5.7}$$
$$\left. \partial_t q_t = -\nabla \cdot (v_t q_t), \qquad Q_0 = Q, \quad Q_1 = Q' \right\}.$$

这是 Benamou–Brenier公式，通过将欧几里得动能替换为范数 $\|\cdot\|_t$，推广了 2-Wasserstein度量。

现在，我们考虑 KL($Q\|P$) 在度量 $\mathcal{W}$ 下的梯度流。形式上， Q 处的最速下降方向可由下式刻画:

$$\operatorname*{argmin}_{Q'} \left\{ \mathrm{KL}(Q'\|P) \ : \ \mathcal{W}(Q',Q) \le \epsilon \right\}, \tag{5.8}$$

或通过其惩罚形式

$$\operatorname*{argmin}_{Q'} \left\{ \mathrm{KL}(Q'\|P) + \frac{1}{2\epsilon^2}\mathcal{W}^2(Q',Q) \right\},$$

这在度量梯度流理论中是标准做法。代入$\mathcal{W}$的定义后，优化问题可化简为

$$\min_{\{v_t\}} \left\{ \mathrm{KL}(Q^v\|P) + \frac{1}{2\epsilon^2} \int_0^1 \|v_t\|_t^2 \,\mathrm{d}t \right\},$$

为了得到易于处理的一阶刻画，考虑如下形式的无穷小扰动

$$Q' = (\mathrm{Id} + \epsilon v)_\sharp Q,$$

其中， v 是一个光滑向量场， $\epsilon > 0$ 是一个小量。这种扰动对应于一条满足连续性方程且速度恒定为 $v_t \equiv v$的路径，因此满足

$$\mathcal{W}(Q',Q) \le \epsilon \|v\|_t.$$

将（5.8）限制于这些扰动，可得到如下近似

$$\operatorname*{argmin}_{v} \left\{ \mathrm{KL}((\mathrm{Id} + \epsilon v)_\sharp Q\|P) - \mathrm{KL}(Q\|P) \ : \ \|v\|_t \le 1 \right\}. \tag{5.9}$$

接下来，沿前推映射对 Kullback–Leibler散度进行一阶泰勒展开，可得

$$\mathrm{KL}((\mathrm{Id} + \epsilon v)_\sharp Q\|P) = \mathrm{KL}(Q\|P) - \epsilon\, Q(\mathcal{T}_P v) + o(\epsilon),$$

其中 $\mathcal{T}_P$ 表示与 P相关的 Langevin Stein 算子。将此展开代入（5.9）并忽略高阶项，可得

$$\operatorname*{argmax}_{v\in\mathcal{G}_t} Q(\mathcal{T}_P v),$$

这就将（5.6）恢复为输运度量 $\mathcal{W}$ 下 KL散度的最速下降方向。

从这一几何视角来看，斯坦变分更新可视为 Kullback–Leibler散度的一种依赖度量的梯度流。对范数$\|\cdot\|_t$的不同选择，以及由此对可容许集合 $\mathcal{G}_t$的不同选择，会在概率测度空间上诱导出不同的几何结构，并产生不同的动力学。

5.4 朗之万扩散作为梯度流

假设 Q_t 具有足够正则的密度 q_t，根据 Stein 恒等式

$$Q_t(\mathcal{T}_P v_t) = \int \nabla\cdot v_t + v_t\cdot\nabla\log p \,\mathrm{d}Q_t = \langle \nabla\log p - \nabla\log q_t,\ v_t\rangle_{L^2(Q_t)}$$

其中， $L^2(Q_t)$ 表示勒贝格空间（例 2.11）。令 $\mathcal{G}_t = \{f : \|f\|_{L^2(Q_t)} \le 1\}$ 为配备范数 $\|f\|_{L^2(Q_t)} = Q_t(\|f\|_2^2)^{1/2}$的 $L^2(Q_t)$ 单位球。最优的 $v_t \in L^2(Q_t)$ 应满足

$$\operatorname*{argmax}_{v_t} \langle \nabla\log p - \nabla\log q_t,\ v_t\rangle_{L^2(Q_t)} \quad \text{s.t.} \quad \|v_t\|_{L^2(Q_t)} \le 1,$$

其解在 Q_t-几乎处处意义下等于

$$v_t^* = \frac{\nabla\log p - \nabla\log q_t}{\|\nabla\log p - \nabla\log q_t\|_{L^2(Q_t)}}.$$

将 v_t^* 代入连续性方程（5.3）得到

$$\dot{q}_t = -\nabla\cdot((\nabla\log p - \nabla\log q_t)q_t) = -\nabla\cdot((\nabla\log p)q_t) + \Delta q_t, \tag{5.10}$$

其中 $\Delta q_t \triangleq \nabla\cdot\nabla q_t$ 为拉普拉斯算子。（5.10）与过阻尼朗之万扩散的福克-普朗克方程一致（参见例2.5）

$$\mathrm{d}X_t = \nabla\log p(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t.$$

这表明，过阻尼朗之万动力学是Kullback–Leibler散度在度量（5.7）下的梯度流。与此相关，沿该流Kullback–Leibler散度的下降速率为

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(Q_t\|P) &= -Q_t(\mathcal{T}_P v_t^*)\\ &= -\|\nabla\log p - \nabla\log q_t\|^2_{L^2(Q_t)} = -\mathrm{FD}(Q_t||P),\end{aligned}$$

其中最终表达式为当前分布 Q_t 与目标 P 之间的Fisher散度（定义4.2）。这表明，当 Q_t 远离 P （在$\mathrm{FD}(Q_t||P)$的意义下）时，Kullback-Leibler的下降更为迅速。

5.5 Stein 变分梯度

朗之万扩散中的速度有一个问题：它取决于演化密度的对数密度，即 $\log q_t$，而后者无法通过计算获得。利用 Stein方法，我们可以改为通过将候选空间 $\mathcal{G}_t$ 设为再生核希尔伯特空间中的单位球，得到计算上可处理的最优速度。

令 $\mathcal{G}_t = \{v \in \mathcal{H}_K\colon\ ||v||_{\mathcal{H}_K} \leq 1\}$ 为一个向量值再生核希尔伯特空间中的单位球，其核具有如下形式 $K(x,y) = k(x,y)I$ 其中标量值核 $k\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ （参见第 2.2.3 节）。对于 $v_t \in \mathcal{G}_t$，根据 Langevin Stein 算子的线性性质，

$$Q_t(\mathcal{T}_P v_t) = \langle v_t, v_t^* \rangle_{\mathcal{H}_K}$$

其中

$$v_t^*(\cdot) = \int k(x,\cdot)\nabla_x \log p(x) + \nabla_x k(x,\cdot)\, \mathrm{d}Q_t(x). \tag{5.11}$$

因此，式（5.6）中的优化问题可化简为

$$\max_{v_t \in \mathcal{G}_t} \langle v_t, v_t^* \rangle_{\mathcal{H}_K}$$

其最优解为

$$v_t = \frac{v_t^*}{\|v_t^*\|_{\mathcal{H}_K}}.$$

由此我们证明，相应的流$(Q_t)_{t\geq 0}$ 由（5.3）给出，它会单调降低 Kullback–Leibler散度，且降低速率等于核 Stein 差异（KSD）（参见定义4.8）：

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(Q_t \mid\mid P) = -\,\mathrm{KSD}_P(Q_t). \tag{5.12}$$

速度场 v_t 刻画了所谓的*Stein* 变分梯度，它构成了多种强大数值算法的基础，详见第6.3.2节。

注 5.3。 该结果表明，*Kullback–Leibler*散度存在一个路径积分公式；$\mathrm{KL}(Q_0||P) = \int_0^\infty$ 核 Stein 差异$_P(Q_t)\,\mathrm{d}t$，这对于估计 *Kullback–Leibler*散度或归一化常数可能很有用 *[Han 和 Liu, 2017]*。

6 应用

本章简要概述了 Stein差异近年来及新近出现的应用，这些应用涉及概率推断和学习中一系列重要的实际任务。

6.1 概述

第4 章引入的 Stein 差异具有一个独特特征：它们能够计算分布 P 与经验测度之间的差异

$$Q_n = \sum_{i=1}^{n} w_i \delta_{x_i}, \tag{6.1}$$

即使无法直接获得 P 的概率质量函数或密度函数，也能对其进行显式计算；这里的不可处理性源于归一化常数难以计算。此处， P 和 Q_n 是同一可测空间$\mathcal{X}$上的概率分布，而 Q_n 由有限个点 x_1、……、 $x_n \in \mathcal{X}$构成，并分别以 w_1、……、 $w_n \in \mathbb{R}$ 为权重，其中 $\sum_i w_i = 1$且 $w_i \geq 0$。回顾一下，Stein 差异由 Stein 算子$\mathcal{T}_P$ 和 Stein集合$\mathcal{G}$定义。对于接下来要讨论的所有应用，$\mathcal{T}_P$ 和$\mathcal{G}$的选择都有灵活性，因此可以根据具体任务定制 Stein 差异。不过，为了简化本章的表述，我们将这些依赖关系隐去，并采用以下简写

$$\mathcal{S}_P(Q_n) \triangleq \mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}).$$

Stein差异及其相关的Stein动力学来自第5 章，它们提供了用途广泛的工具，自然适用于一系列重要的统计应用，大致可分为四类，现将予以介绍。

衡量近似质量 （第6.2节）给定样本 $\{x_i\}_{i=1}^n$ 和一个密度为 $p(x) \propto \exp(-V(x)$、仅精确到一个归一化常数的分布P，$\mathcal{S}_P(Q_n)$的大小可用于量化样本$\{x_i\}_{i=1}^n$ （赋予相同权重）所形成的经验分布 Q_n 与 P之间的不相似度。因此，Stein差异已成为一种广受欢迎且灵活的拟合优度检验方法，可用于检验原假设 $x_i \overset{\text{独立同分布}}{\sim} P$。在原假设下，$\mathcal{S}_P(Q_n) \to 0$当 $n \to \infty$时，因此

若 $\mathcal{S}_P(Q_n) \geq \tau$，则拒绝原假设，其中 τ是恰当选取的阈值。这种方法极大地扩展了传统的拟合优度检验，例如卡方检验；后者仅适用于简单且低维的分布。与此同时，Stein差异也已成为衡量旨在从后验分布 P中抽样的算法所生成近似质量的常用工具，可用于调整采样方法，并在贝叶斯统计背景下比较不同的方法。

基于粒子的近似算法（第6.3节）给定一个概率分布 P，其概率质量函数或密度函数 $p(x) \propto \exp(-V(x))$ 仅在一个难以处理的归一化常数意义下确定，可以将数值近似 P 的问题转化为一个优化问题：寻找状态 x_1、……、$x_n \in \mathsf{X}$ 及相应的权重 w_1、……、$w_n \in \mathbb{R}$，使得相关的经验分布 (6.1) 能最小化 Stein 差异 $\mathsf{S}_P(Q_n)$。固定均匀权重 $w_i = \frac{1}{n}$后，可以通过梯度下降或其他适当的数值优化方法对粒子 $\{x_i\}_{i=1}^n$ 进行优化，从而得到新一类基于粒子的算法。另一方面，第5章讨论的 Stein动力学可用于对 Kullback–Leibler散度执行一种梯度下降，由此得到一种称为 Stein变分梯度下降（SVGD）的算法。固定粒子 $\{x_i\}_{i=1}^n$，转而优化权重 $\{w_i\}_{i=1}^n$，则可得到与重要性采样理念相近的算法；如果 $\{x_i\}_{i=1}^n$ 并非通过从 P中抽样而产生，这些算法还可以校正偏差。此外，对权重 $\{w_i\}$施加稀疏性约束，可以得到稀疏化算法，从 $\{x_i\}_{i=1}^n$中选取一个较小的子集，以准确近似 P。最后，给定一个特定函数 $f : \mathsf{X} \to \mathbb{R}$，且关注其相对于 P 的期望，可以推导出求积近似 $\sum_{i=1}^n w_i f(x_i)$ 中权重 w_i的最优值；当 x_i 是随机抽样得到时，这是一种有效的方差缩减技术。

生成模型训练 （第 6.4 节）给定数据集 $\{x_i\}_{i=1}^n$，可以从集合 $\{P_\theta\}_{\theta\in\Theta}$ 中寻找合适的生成模型，即在指标集 θ 上搜索参数 Θ的适当取值。对于易于处理的生成模型，可以使用最大似然估计等标准统计技术。然而，许多重要的生成模型实例并不易于处理；例如，P_θ 可以通过密度函数 $p_\theta(x) \propto \exp(-V_\theta(x))$ 来指定，但归一化常数难以求得。在此情境下，Stein 差异和 Stein动力学提供了多种有用的替代参数估计方法。

梯度估计（第6.5节）机器学习中常见的一项技术挑战是，无法通过解析方法获得相对于分布 $\mathbb{E}_{X\sim P_\theta}[f(X)]$的参数 θ的梯度P_θ，因此必须对其进行近似。这一挑战出现在许多学习问题中，包括训练用于变分推断的潜变量模型，以及运用策略梯度进行强化学习。通常可以使用蒙特卡洛方法，但其相关误差可能相当大，从而对下游

机器学习任务的性能产生负面影响。Stein方法催生了若干方差缩减策略，可用于应对这一问题。

6.2 衡量近似质量

我们首先详细考察的一个应用，是衡量针对未归一化分布目标的经验近似质量这一问题。具体而言，设 P 是定义在概率空间（$\mathcal{X}$，$\mathcal{S}_{\mathcal{X}}$，$\lambda_{\mathcal{X}}$）上的一个感兴趣分布，其相对于 $\lambda_{\mathcal{X}}$ 的概率密度函数 $p(x)$ 已知，但归一化常数难以计算，即：

$$p(x) \propto \frac{\tilde{p}(x)}{Z}$$

其中 $\tilde{p}$（x）可以显式求得，但归一化常数 $Z = \int \tilde{p}$（x）$\mathrm{d}\lambda_{\mathcal{X}}$（$x$）是一个难解积分。下文将讨论两种不同情形（第6.2.1节讨论量化采样方法的性能，第6.2.2节讨论拟合优度检验）；两种情形下的任务都是确定经验近似$Q_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$与 P之间的质量，其中 x_1、……、$x_n \in \mathcal{X}$是 Q_n的支撑集中的状态。

6.2.1 量化采样方法的性能

量化采样方法性能的问题常见于贝叶斯统计中，其中 P 是后验分布，其密度 $p(x) \propto \pi(x)$ L(x) 可以未归一化的形式表示为先验 $\pi(x)$ 与似然 L(x) 的乘积（参见第 1.2 节）。这里，$\tilde{p}(x)$ 可以表示为乘积 $\pi(x)$L(x)，而归一化常数

$$Z = \int \pi(x)\mathcal{L}(x)\, \mathrm{d}\lambda_{\mathcal{X}}(x)$$

称为边际似然。衡量样本质量的任务会出现在以下情形中：监测可扩展的[1] 采样算法的收敛情况、比较不同（精确和/或近似）采样算法所产生的近似结果，以及调整特定采样算法的超参数。在每种情形下，Stein差异都被用作定量标准，以便明确衡量性能。

例 6.1（调节未校正朗之万算法）。 用于从目标分布近似抽样的未校正朗之万算法（ULA）是对 P 进行的欧拉-丸山离散化（X_t）$_{t\in\mathbb{N}}$，它对应于 P-不变的过阻尼朗之万扩散（参见例 *2.5*）：

$$X_{t+1} = X_t + \frac{\epsilon}{2}\nabla \log p(X_t) + \sqrt{\epsilon} Z_t$$

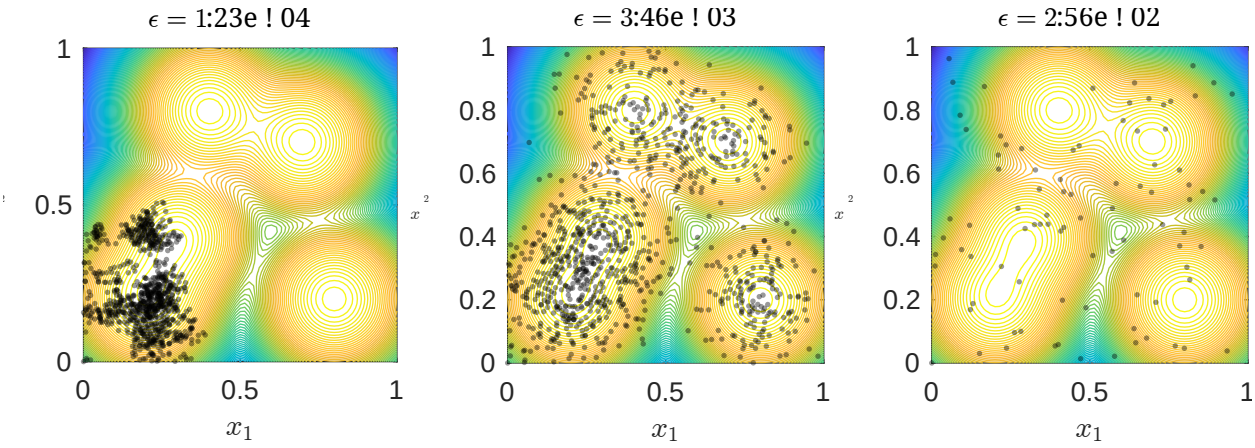

图 6.1： 使用未校正朗之万算法（ULA）生成的样本。这里，目标 P 是一个高斯混合模型，其密度为 p，ϵ 表示 ULA 的步长参数。图中展示了 $\log p$ 的等高线。

其中，Z_t 是一个独立于 X_1、……、X_t生成的标准正态随机变量。这里，$\epsilon > 0$ 是一个需要选定的步长参数。假设我们的计算预算允许计算 $(X_t)_{t=1}^n$ 次，其中 $n = 1$为 *000*；我们希望选择 ϵ，使经验分布 $Q_n = \frac{1}{n}\sum_{t=1}^n \delta_{X_t}$ 很可能为 P提供最佳近似。具体而言，我们假设要使用 *ULA* 的经验平均值，近似图 *6.1* 中分布 P 各个分量的均值和方差。若 ϵ过小，随机过程就无法有效探索 P 的高概率区域（图 *6.1*；左图）。另一方面，若 ϵ 过大，则由于对随机微分方程进行了一阶离散化，会在 X_t 的分布中引入偏差，并最终可能导致 *ULA* 出现爆炸行为（图 *6.1*；右图）。遗憾的是，由于目标分布难以处理，通常无法预先知道合适的 ϵ取值（图 *6.1*；中图）。*Stein* 差异提供了一种解决方案，使得能够明确衡量近似 Q_n 的质量，即 $S_P(Q_n)$。事实上，与 P 均值和方差的经验估计相关的均方误差（通常无法计算）在 $\epsilon \approx 10^{-2}$ 附近达到最小值（图 *6.2*；左图），而核 *Stein* 差异（*KSD*）（可以计算）也在 $\epsilon \approx 10^{-2}$附近达到最小值（图 *6.2*；右图）。

另一个相关情形涉及所谓的难以处理的概率模型，这类模型在统计物理中颇受欢迎，并以相对概率的形式 $p(x)/p(y)$ 来定义，以避开难以计算的归一化常数。由于从 P 中进行精确模拟通常不可行，因此人们开发了各种采样算法 [例如参见Newman andBarkema, 1999] 并且可以按照刚才所述，使用 Stein 差异来评估这些算法。

[1]如果一个采样算法原则上能够生成无限的状态序列 $(x_n)_{n\in\mathbb{N}}$，那么它就是可扩展的；在实际操作中，一旦满足适当的停止准则，便可以截断该序列并终止算法，从而获得Q_n 的近似 P。

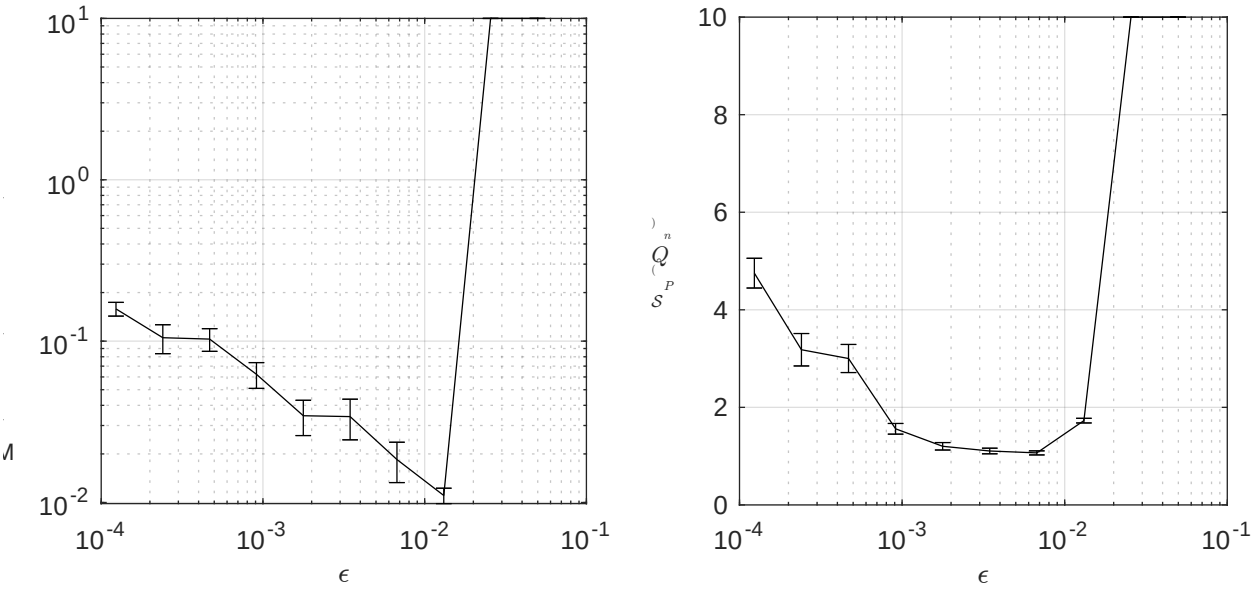

图 6.2： 调整未校正朗之万算法（ULA）。左图展示了与 P 的均值和方差经验估计相关的均方误差（这些量通常无法计算），右图展示了核 Stein 差异（KSD）（可以计算）。此处使用了逆多二次核（例2.13），其指数为 $\beta = 1$，带宽为 $\ell = 0.01$。实验重复进行 10 次，并报告均值和标准误。

6.2.2 拟合优度检验

在针对仅由难以计算的归一化常数所限定的统计模式进行拟合优度检验时，也会遇到衡量样本质量的问题。也就是说，给定一个数据集 $\{x_i\}_{i=1}^n$ 和一个概率分布 P，其概率密度函数的形式为 $p(x) \propto \exp(-V(x))$，其中势函数 $V(x)$ 已明确给出，我们希望检验如下原假设：$x_i \overset{\text{独立同分布}}{\sim} P$。任何能够区分分布的 Stein 差异，也就是说，当且仅当 $Q = P$ 时，$\mathcal{S}_P(Q) = 0$，都能提供一个可计算的统计量 $\mathcal{S}_P(Q_n)$，用于进行此类检验。如果原假设成立，我们可能会预期 $\mathcal{S}_P(Q_n) \to 0$，当 $n \to \infty$ 时亦然，因此，如果 $\mathcal{S}_P(Q_n)$ 不够小，我们就希望拒绝原假设。实际上，原假设下 Stein 差异 $\mathcal{S}_P(Q_n)$ 的抽样分布通常未知，但可以使用自助法近似得到。在此处选择 Stein 差异时，主要考虑的是它能否快速计算（以便进行自助重抽样步骤），以及在原假设不成立时，它是否有足够的检验力来区分 P 与产生该数据集的真实机制。

为便于可计算性，我们采用一种核 Stein 差异（KSD）（第 4.4 节），其 Stein 核为 k_p：$\mathcal{X} \times \mathcal{X} \to \mathbb{R}$。假设 $\int k_p(x, x) \mathrm{d}P(x) < \infty$，则在零假设下

假设

$$\begin{aligned}\mathbb{E}\left[\mathcal{S}_P(Q_n)^2\right] &= \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k_p(x_i,x_j)\right]\\ &= \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}k_p(x_i,x_i)+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}k_p(x_i,x_j)\right]\\ &= \frac{1}{n}\int k_p(x,x)\mathrm{d}P(x)\to 0,\end{aligned}$$

其中最后一行成立，是因为 $x_i \sim P$在零假设下相互独立，且 $\int k_p(\cdot\ x)\ \mathrm{d}P(x)\ = 0$，因为 k_p 是Stein核。因此，我们可以利用统计量$\mathrm{S}_P(Q_n)$构造一个水平为 α 的拟合优度检验，并在 $\mathrm{S}_P(Q_n)\ > \tau$时拒绝原假设，其中阈值 τ的选取使得 $\mathbb{P}(\mathrm{S}_P(Q_n)\ \leq \tau)\ = 1-\alpha$。由于通常无法从 P 进行模拟（归一化常数难以处理），在原假设下数值计算$\mathrm{S}_P(Q_n)$抽样分布的 $1-\alpha$ 分位数并不直接。文献中一种已变得流行的方法 [Liu et al., 2016, Chwialkowski等人, 2016]是使用野生自助法 [Shao, 2010, Fromont等人, 2012年, Leucht和Neumann, 2013], , 该方法引入额外的独立拉德马赫随机变量$\{\epsilon_i\}_{i=1}^n$（即每个 ϵ_i 在$\{-1, 1\}$上服从均匀分布），并指出 $n\hat{}D_P(Q_n)^2$的分布在零假设下趋近于 $n\mathrm{S}_P(Q_n)^2$ 的分布，其中

$$\hat{D}_P(Q_n)^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\epsilon_i\epsilon_j k_p(x_i,x_j) \tag{6.2}$$

并且我们考虑其 $n\to\infty$ 极限。生成多个拉德马赫随机变量的实例并重新计算（6.2），即可获得 $\mathcal{S}_P(Q)$ 在零假设下抽样分布的经验近似，据此可以提取合适的阈值 τ。

例 6.2。考虑检验高斯–伯努利受限玻尔兹曼机对数据集 $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$ 的拟合优度。该模型具有潜变量 $h\in\{-1,1\}^{d'}$，其联合密度形式如下

$$p(x,h) = \frac{1}{Z}\exp\left(\frac{1}{2}x^\top Bh + b^\top x + c^\top h + \frac{1}{2}x^\top x\right)$$

其中，Z 是适当的归一化常数。可观测量 x 的边缘分布 P 的密度为

$$p(x) = \sum_{h\in\{-1,1\}^{d'}} p(x,h),$$

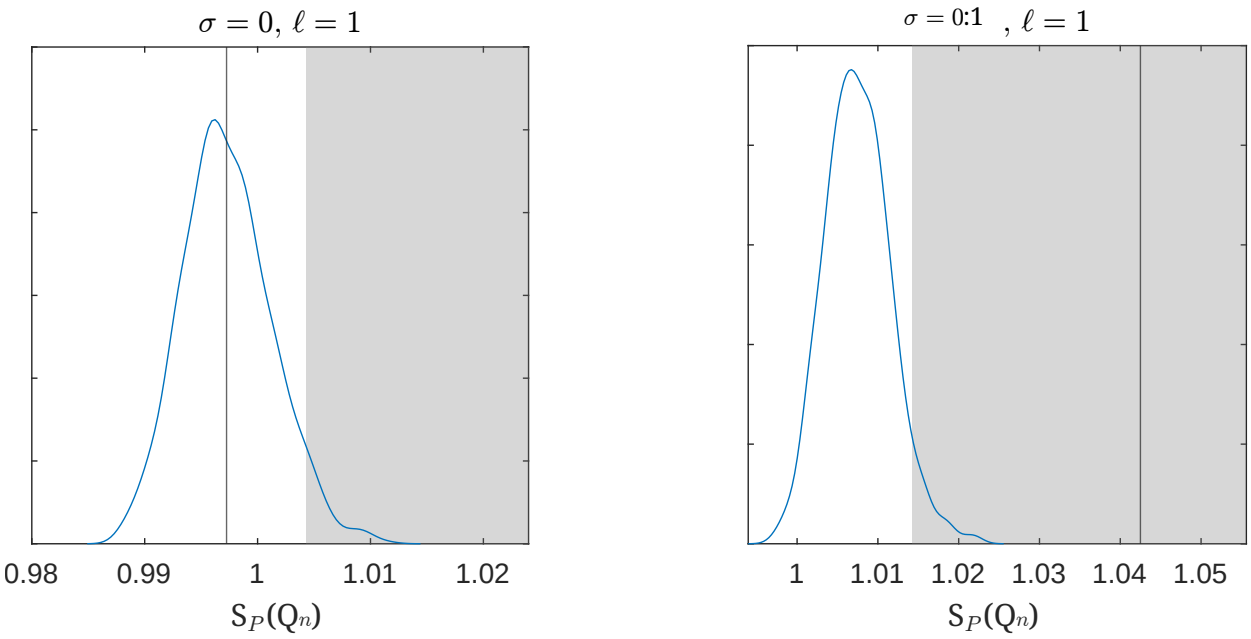

图 6.3： 基于核 Stein 差异的拟合优度检验。图中展示了通过野生自助法近似得到的 KSD 检验统计量的零分布（蓝色）、KSD 检验统计量的实际值（黑色），以及零假设会被拒绝的取值范围（阴影部分）。模型设定错误的程度用 σ 表示，因此当 $\sigma = 0$ 时，模型正确。使用长度尺度为 ℓ 的逆多二次核。

但由于依赖于 Z，这可能难以处理。另一方面，得分函数具有闭式表达

$$\nabla \log p(x) = b - x + \frac{1}{2} B \tanh\left(\frac{1}{2} B^\top x + c\right),$$

因此，我们可以利用 *Stein* 差异构造拟合优度检验。

为作说明，我们沿用 *Liu* 等人 *[2016]*，的设定：取 $d = 50$、$d' = 10$、$n = 100$，从标准高斯分布中对 b 和 c 的元素进行抽样，并从 $\{-1, 1\}$ 中均匀抽取 B 的元素。我们的实验使用逆多二次核 k（例 *2.13*），并采用长度尺度 $\ell > 0$ 来构造 *Stein* 核 k_P。数据要么来自真实模型，要么来自真实模型的扰动版本；在后者中，向 B 的元素添加方差为 σ^2 的高斯噪声。图 *6.3* 展示了通过野生自助法近似得到的 *KSD* 检验统计量的零分布，以及 *KSD* 检验统计量的实际值和对应于 $\alpha = 0.05$ 的拒绝阈值 τ。在 $\sigma = 0$ 的情况下，模型正确，实际 *KSD* 值落在零分布的中心区域内，因此检验不会拒绝原假设；而在 $\sigma = 0.1$ 的情况下，实际 *KSD* 值远落在零分布的尾部，因此原假设被拒绝。

一个自然的问题是，如何选择 *Stein* 核 k_p，使检验功效（即检测偏离零假设的概率）最大化。对于我们的简单设定，

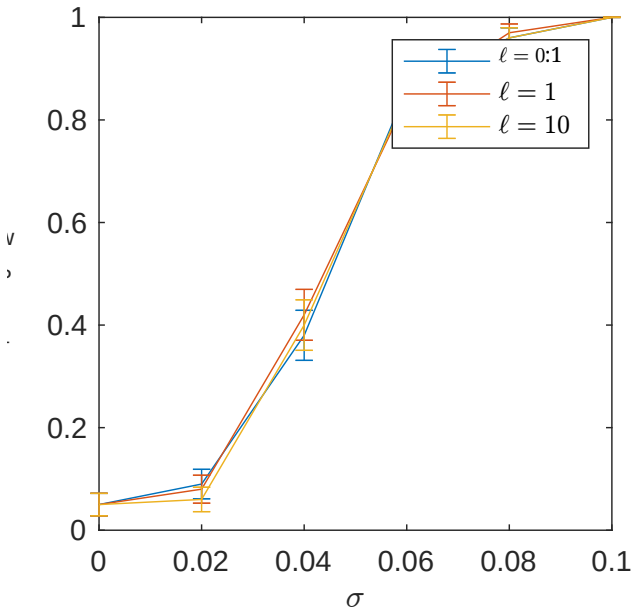

图 6.4： 使用核 Stein 差异进行拟合优度检验。此处绘出了检验功效（即拒绝零假设的概率）随模型设定错误程度 σ 和核长度尺度 ℓ变化的情况。报告的是 100 次实验的均值和标准误。

图 *6.4* 绘制了三种不同核长度尺度 ℓ下的检验功效，其中每种情况下，检验功效都是模型设定错误程度 σ的函数。有趣的是，这项检验的检验功效似乎几乎不受 ℓ的影响；各种选择的表现都差不多。

通常，选择合适的 Stein 核可能需要采用数据拆分等技术，并对检验功效进行易于计算的近似，以便在数据集的训练子集上进行优化。作为起始点，读者可参考 Jitkrittum 等人 [2017], ，其中给出了检验功效的显式近似。还有一些作者提出了提高检验功效的其他策略，例如汇总多项检验 [Schrab 等人， 2022], 谨慎选择 Stein 算子 [Liu 等人， 2023], 以及进行谱正则化 [Hagrass 等人， 2026]。与此同时，检验的计算成本也在降低 [Jitkrittum 等人，2017年，Huggins 和 Mackey， 2018], 并且正在开发稳健检验 [Liu 和 Briol， 2025], 在线检验 [Martinez-Taboada 和 Ramdas， 2025], 以及针对潜变量模型的相对拟合优度检验 [Kanagawa 等人， 2023]等其他功能。

6.3 基于粒子的近似算法

在上一节中，我们扮演被动观察者的角色，评估已提供样本的质量。现在，我们将注意力转向如何主动构造 分布目标 P的离散近似这一问题。我们再次受到

贝叶斯统计等应用的启发。在这些应用中，我们只能通过未归一化的密度函数获取 P。为了演示，我们将反复考虑二维$Rosenbrock$ 目标

$$p(x,y) \propto \exp\left(-x^2 - 3(y-x^2)^2\right). \tag{6.3}$$

首先，在第 6.3.1 节和 6.3.2 节中，我们介绍若干算法，旨在选取状态 $\{x_i\}_{i=1}^n$，使其对应的（等权重）经验测度能够准确近似目标。在第 6.3.3节中，我们探讨如何为给定样本赋予权重，从而得到加权经验测度，以更忠实地近似目标。第 6.3.4节进一步探讨稀疏近似方案，这类方案既能实现压缩，又能以较低的计算成本提高样本质量。随后，在第6.3.5节中，我们将讨论范围缩小到对单个期望感兴趣量的近似。

6.3.1 序贯算法

首先，我们考虑序贯算法，或称可扩展算法，也就是说，原则上会生成一个无限序列 $(x_i)_{i\in\mathbb{N}}$。在实践中，这类算法会在有限次迭代n 后终止，并取前 n 个粒子$\{x_i\}_{i=1}^n$ 构成对分布目标 P的经验近似$Q_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$。可扩展算法的主要优势在于，可以根据需要持续运行，直至获得对P 足够准确的近似。

Stein 点一种值得考虑的自然算法是对 Stein 差异进行序贯贪心最小化。该算法以取决于先前所选粒子 x_1、……、x_{n-1} 的方式选择第 n 个粒子 x_n，具体如下：

$$x_n \in \operatorname*{argmin}_{x\in\mathcal{X}} \mathcal{S}_P\left(\frac{1}{n}\delta_x + \frac{1}{n}\sum_{i=1}^{n-1}\delta_{x_i}\right) \tag{6.4}$$

其中，按惯例，P 是定义在 X 上的分布目标。以这种方式选出的粒子 $(x_i)_{i\in\mathbb{N}}$ 称为 Stein 点。实际中，几乎无法通过解析方法求解（6.4）中的优化问题，因此需要采用数值方法。Chen 等人 [2018b] 分析了每次迭代所需的数值优化精度，以确保在核 Stein 差异的情形下收敛至 $\mathrm{S}_P(Q_n) \to 0$。对于能够控制收敛的 Stein 差异，我们由此可以得出结论：Q_n 收敛至 P。遗憾的是，使用通用数值优化技术会在$d =$ 维数 dim(X) 上引入维数灾难。为缓解维数灾难，Chen 等人 [2019b] 提出了 Stein 点马尔可夫链蒙特卡罗，其求解的是

$$x_n \in \operatorname*{argmin}_{x\in\{y_1^{(n)},\ldots,y_{m_n}^{(n)}\}} \mathcal{S}_P\left(\frac{1}{n}\delta_x + \frac{1}{n}\sum_{i=1}^{n-1}\delta_{x_i}\right) \tag{6.5}$$

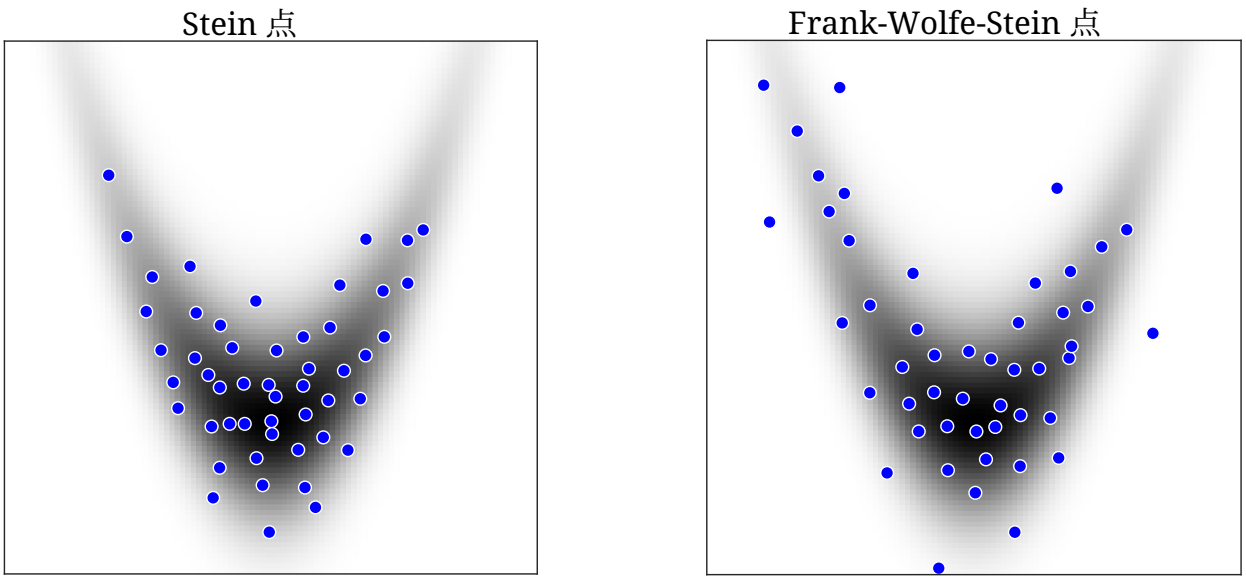

图 6.5： Stein 点（左）和 Frank–Wolfe–Stein 点（右）被用于近似 Rosenbrock 目标分布（6.3）（黑色阴影）。

其中，$(y_i^{(n)})_{i=1}^{m_n}$ 是一个 P-不变的马尔可夫链，长度为 m_n，初始状态为 $y_1^{(n)} \in \mathcal{X}$。参见图6.5左图。只要马尔可夫链的混合足够快，且 $m_n \to \infty$ 当 $n \to \infty$时，便可在相当一般的条件下证明（6.5）的一致性。例如，可以令初始状态 $y_1^{(n)}$ 依赖于已选点的历史 x_1、……、x_{n-1}；一种性能出色的选择是所谓的最具影响力的点x_i，其 Stein 差异中去掉 x_i，即

$$\mathcal{S}_P\left(\frac{1}{n-2}\sum_{j\in\{1,\ldots,n-1\}\setminus\{i\}}\delta_{x_j}\right),$$

被最大化。选择这一做法的直观理由是，x_i 对 Q_{n-1} 的近似质量至关重要，因此在 x_i 的邻域中再添加一个状态 x_n 可能是有益的。

Frank–Wolfe–Stein 点 现在考虑核 Stein 差异这一特定情形下的 Stein 点算法，从而更新（6.4）可以借助 Stein 再生核 k_p 显式写出（例如，对于 Langevin KSD，如（4.8）所示），即

$$x_n \in \operatorname*{argmin}_{x\in\mathcal{X}} \frac{k_p(x,x)}{2} + \sum_{i=1}^{n} k_p(x_i,x) \tag{6.6}$$

其计算成本可见为 $O(n)$。式（6.6）中的第一项可解释为正则化项，实际上也可以考虑求解这一问题的未正则化版本，

从而得到一组不同但相关的点序列

$$\tilde{x}_n \in \operatorname*{argmin}_{x \in \mathcal{X}} \sum_{i=1}^{n} k_p(\tilde{x}_i, x). \tag{6.7}$$

Chen 等人 [2018b]注意到，序列 $(\tilde{x}_i)_{i=1}^n$ 通过求解（6.7）中的无正则化问题而产生，它恰好是将*Frank–Wolfe*方法应用于 Stein 差异最小化所得到的序列，正如核方法文献中所研究的那样 [例如 Chen 等人，2010b，Bach 等人， 2012] 其中名称 *herding*算法 也被使用。[2] 参见图 6.5 的右图。这一视角表明，可以采用 Frank–Wolfe 文献中的多种技术来提高近似精度，例如额外允许移除先前已选定的次优状态，或使用加权近似（我们将在第6.3.3节讨论加权近似）；参见 Bomze 等人 [2024] 以了解背景。

在此情境下选择 Stein 差异时，或许最主要的考虑是：Stein 差异的收敛控制性质应当反映采样算法的输出将被用于的具体应用。例如，若样本将用于近似后验期望，则应使用能够控制此类期望收敛的 Stein 差异（参见第4.4.4节）。]

6.3.2 批处理算法

可扩展算法无法对 n的所有可终止取值都产生高度精确的近似 [Owen， 2016]。这一点可以通过对称性破缺直观地看出；假设希望近似一个高斯分布 $P = N(0, 1)$，并使用 n 个粒子。对于 $n = 1$，自然的做法是把 $x_1 = 0$ 选在中心点，但我们随即被迫做出一个困难的选择 x_2；要么也选择 $x_2 = 0$，这样就不能准确捕捉 P的散布，要么被迫打破近似的对称性（尽管 P 本身是对称的）。批量算法试图通过一开始就指定数目 n 个粒子来避免这一困境。这使其能够在数目 n 与可扩展算法相当的粒子上获得更好的近似，但也意味着，若希望进一步提高精度，则（至少在原则上）必须在 n 增大之后从头重新运行该算法。下面将讨论两种说明性的批量算法：

Stein 差异下降 一种自然的批量算法构造方法，是从 Stein 差异 $\mathcal{S}_P(Q)$ 的所谓 *Wasserstein* 梯度流 出发。它实际上是关于 Q 参数进行连续时间梯度下降的一种形式，产生一条测度值轨迹 $(Q^{(t)})_{t\geq 0}$，其中每个 $Q^{(t)}$ 都是 $\mathcal{X}$ 上的概率分布。直观而言，

[2]另一种相关技术，*MMD-FW* [Futami 等人， 2019], 使用 Frank–Wolfe方法和 LangevinStein 算子来最小化相关的核 MMD，而不是直接最小化 Stein 差异。

至少，梯度流 $Q^{(t)}$ 应在适当的正则性条件下收敛到 P，当$t \to \infty$。出于数值计算的目的，测度 $Q^{(t)}$ 用一个（权重均匀的）经验分布 $Q_n^{(t)} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i^{(t)}}$ 来近似，其支撑点 $x_i^{(t)}$ 此时依赖于 t。Korba 等人 Korba 等人 [2021], 考虑了这种方法，并基于 KSD 的平方进行梯度下降，推导出以下数值方案：

$$x_i^{(t+1)} = x_i^{(t)} - \frac{\epsilon}{n^2}\sum_{j=1}^n \nabla_2 k_p(x_j^{(t)}, x_i^{(t)})$$

其中 $\epsilon > 0$ 是一个 学习率，具体取值将在后文说明。也可以考虑使用更高阶的数值优化技术，例如 L-BFGS；读者可参阅Korba 等人 [2021]。Stein 差异下降能够生成目标 P的紧凑表示，但代价是需要计算 P的对数密度的二阶导数。接下来将介绍一种只需要一阶导数的替代策略。

Stein 变分梯度下降 将梯度下降应用于 Stein 差异，会导致需要计算 P 的对数密度的二阶导数，因为 Stein 差异本身涉及 P的一阶导数。一个解决方案是对不同的目标函数进行梯度下降；具体而言，自然的选择是考虑 Kullback–Leibler散度（定义 2.26）。

设 $\mathcal{X} = \mathbb{R}^d$，并考虑一个映射 $T\colon \mathbb{R}^d \to \mathbb{R}^d$，其形式为 $T(x) = x + \epsilon g(x)$ 其中 $\epsilon \ll 1$且 $g\colon \mathbb{R}^d \to \mathbb{R}^d$。由第 5.1 定理可知，在第 5 章中我们有

$$-\frac{\mathrm{d}}{\mathrm{d}\epsilon}\mathrm{KL}(T_{\#}Q\|P)\Big|_{\epsilon=0} = Q(\mathcal{T}_P g) \tag{6.8}$$

其中 $\mathcal{T}_P$ 是 Langevin Stein 算子（定义 3.2），因此我们假设 g 足够规则，属于该 Stein 算子的定义域（参见命题 3.1）。给定函数 g的一组候选项 $\mathcal{G}$，最速下降方向对应于选择

$$g^* \in \operatorname*{argmax}_{g\in\mathcal{G}} Q(\mathcal{T}_P g). \tag{6.9}$$

一种计算上颇具吸引力的选择是取集合 $\mathcal{G}$ 为向量值再生核希尔伯特空间 $\mathcal{H}(K)$的单位球，其中（6.9）中的目标函数即为核 Stein 差异（定义4.8）。事实上，在这种情况下

$$g^*_{P,Q}(y) = \int \mathcal{T}_{(1)P} K(x, y)\,\mathrm{d}Q(x) \tag{6.10}$$

可以显式计算。这些计算表明，可以采用一种实用算法：将 Q 替换为离散分布 $Q_n^{(t)} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i^{(t)}}$，并按照最速下降方向，以依赖于时间 t的方式更新支撑点$x_i^{(t)}$ 的位置：

$$x_i^{(t+1)} = x_i^{(t)} + \epsilon g^*_{P,Q_n^{(t)}}(x_i^{(t)}) \tag{6.11}$$

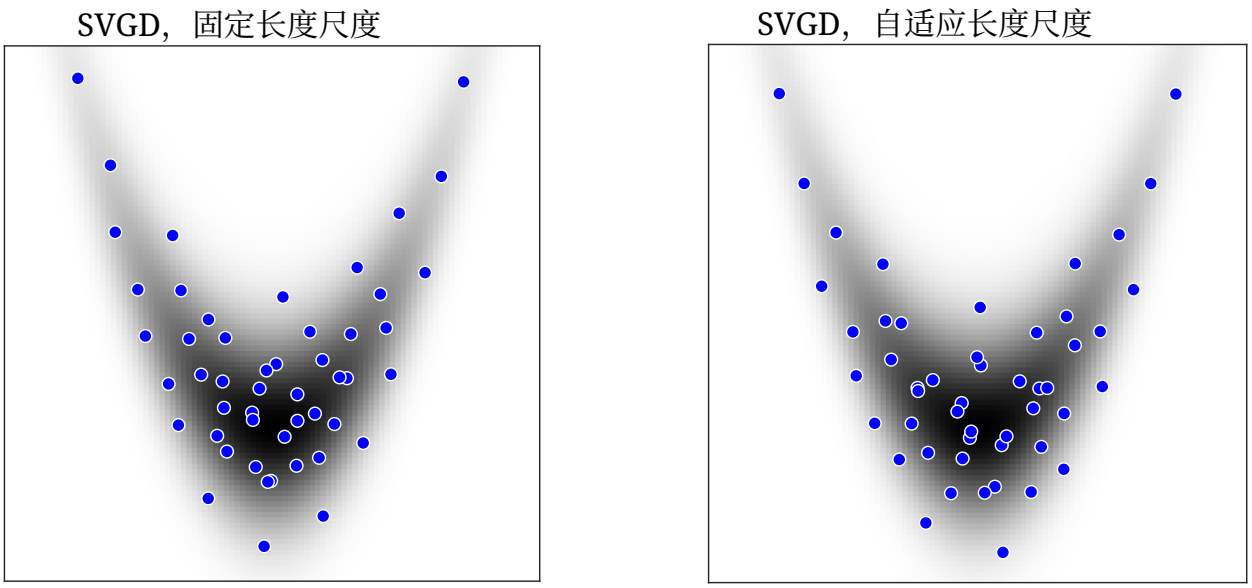

图6.6： 采用固定核长度尺度的Stein变分梯度下降 ℓ （左）和自适应选择的核长度尺度 ℓ （右）来近似Rosenbrock目标分布（6.3）（黑色阴影部分）。

对于矩阵值核 K 取如下形式 $K(x,\ y)\ = k(x,\ y)I$这一特殊情形，最速下降方向的显式形式为

$$g^*_{P,Q_n^{(t)}}(x_i^{(t)}) = \frac{1}{n}\sum_{j=1}^{n}(\nabla \log p)(x_j^{(t)})k(x_j^{(t)}, x_i^{(t)}) + \nabla_1 k(x_j^{(t)}, x_i^{(t)}) \tag{6.12}$$

该算法称为Stein变分梯度下降（SVGD） [Liu和Wang，2016]。（一般矩阵值核的情形见Zhuo 等人 [2018], Wang等人 [2017]。）

对于任何合理的核，例如长度尺度为 ℓ的高斯核，离散分布序列 $Q_n^{(t)}$会收敛到对 P的近似，如图6.6（左）所示。事实上，在选择合适的初始化、步长和步数 t_n后，SVGD近似 $Q_n^{(t_n)}$ 已知会以 $O(1/\sqrt{n})$速率在核 Stein 差异 [下收敛至 P 2023] Gorham 等，2020；Shi和Mackey，Banerjee等人，2025]。实践中，核长度尺度 ℓ 通常设为依赖于 t，一种默认设置是令 ℓ 等于$\{x_i^{(t)}\}_{i=1}^n$之间成对距离的中位数；见图6.6（右）。

(6.12) 中的两项各有直观作用：含梯度 $(\nabla\log\ p)(x)$的项会将粒子推向 P 的高概率区域，而含 $\nabla_1 k(x, x_i)$的项则提供排斥力，以确保粒子之间保持一定程度的多样性。如果没有排斥力，或者只有一个粒子（且核满足 $\nabla_1 k(x, x') = 0$对于 $x = x'$），解就会收敛到 P 的密度局部最优点，从而转化为最大后验（ MAP ）问题。因此，通过使用不同数量的粒子，SVGD

在最大后验与完整的基于粒子的近似之间提供了一种插值，从而能够在准确性与计算成本之间实现高效权衡。

备注 6.1（SVGD 对经验测度的良定义性）。SVGD 可以视为通过迭代更新经验粒子测度 $Q_n^{(t)}$，以最小化$\mathrm{KL}(Q_n^{(t)}\|P)$。这里一个关键但微妙的要点是，$\mathrm{KL}(Q_n^{(t)}\|P)$在技术上是无穷大（或未定义）的，因为经验测度 $Q_n^{(t)}$相对于 P并非绝对连续；更具体地说，

$$\mathrm{KL}(Q\|P) = \mathbb{E}_{X\sim Q}[\log q(X) - \log p(X)]$$

取决于测度 Q的密度 q，当 Q 是经验测度时，无法妥善定义。

一个关键的观察是，在可逆传输下，***KL***散度的差值 仍然有限，并且可以通过关于 Q 的期望计算，即使 Q 是经验测度。令$T:\ \mathbb{R}^d \to \mathbb{R}^d$ 为一个 C^1 微分同胚。那么

$$\begin{aligned}
&\mathrm{KL}(T_{\#}Q\|P) - \mathrm{KL}(Q\|P) \\
&\quad = \mathrm{KL}(Q\|T_{\#}^{-1}P) - \mathrm{KL}(Q\|P) &(6.13)\\
&\quad = -\,\mathbb{E}_{X\sim Q}\Big[\log p(T(X)) + \log\Big|\det \nabla T(X)\Big| - \log p(X)\Big]\,, &(6.14)
\end{aligned}$$

其中 $T^{-1}{}_{\#}P$表示 P经 T^{-1}前推所得的测度，其密度为 $x \mapsto p\left(T\left(x\right)\right)\left|\det \nabla T\left(x\right)\right|$.在 (6.13) 中，我们用到了同时进行可逆变换时 ***KL***散度的不变性，即 $\mathrm{KL}\left(T_{\#}Q\|T_{\#}P\right) = \mathrm{KL}\left(Q\|P\right)$。

注意，(6.14) 的右侧 Q 仅通过期望 $\mathbb{E}_{X\sim Q}[\cdot]$,依赖于 Q，因此对于经验 Q仍有良好定义。

此外，当 $T(x) = x + \epsilon g(x)$ 且 ϵ较小时，(6.14) 的一阶展开可SVGD 目标下降公式（例如(6.8)），该公式的右侧也仅通过 Q依赖于 $\mathbb{E}_{X\sim Q}$。因此，可以通过将SVGD 粒子更新中的 Q替换为经验粒子测度来推导该更新。

注 6.2 (与 Stein 差异下降的比较**)**。***Stein*** 差异下降与SVGD 密切相关。打个比方，如果我们要在欧几里得空间 $f(x)$上最小化一个凸函数 $x \in \mathbb{R}^d$，那么可以尝试最小化 $f(x)$或 $\|\nabla f(x)\|$；最小化任一目标都会得到相同结果。因此，***Stein*** 差异下降旨在最小化 ***Stein*** 差异，这类似于最小化 ***Kullback–Leibler***散度的梯度（定理*5.1*）；而在SVGD 中，最小化的是 ***Kullback–Leibler***散度本身。

自Liu 和 Wang [2016], 提出 SVGD 以来，它已通过多种方式得到扩展和改进。以下列举一些例子，但并非穷尽：摊销SVGD [Feng 等人，2017年、Wang 和 Liu，2016、Liu 和 Wang，2017] 该方法学习神经采样器，而非使用粒子近似；无梯度SVGD [Han 和 Liu，2018] 该方法无需梯度

目标分布的信息 P；图结构 SVGD [Wang等人，2017年，Zhuo 等人， 2018]以及矩阵核 SVGD [Wang 和 Liu， 2019] 这两种方法将结构化信息融入核中，以提升高维情形下的性能；黎曼流形上的 SVGD [Liu 和 Zhu， 2018]以及 SVGD 的牛顿变体 [Detommaso 等人，2018 年，Chen 等人，2019a]；通过对粒子进行子采样来加速的随机变体 [Li等人，2020年，Gorham 等人， 2020]；通过最小化分位数损失来提高稳健性的分位数 SVGD [Gong 等人， 2019]；最小化 KL 散度之外更一般的非线性损失函数的非线性 SVGD [Wang 和 Liu， 2019]；Stein 变分重要性采样 [Han 和 Liu，2017]；以及通用粒子优化框架 [Chen 等人，2018a]。

6.3.3 Stein重要性采样

到目前为止，我们一直考虑使用权重均匀的粒子进行近似。Stein差异的一个吸引人之处在于，它们使最优加权近似 问题能够得到易于处理的解：给定固定 状态 $x_1, \ldots, x_n \in \mathcal{X}$（这些状态可能通过本节前述的任一算法获得），我们寻求一个加权测度

$$Q_n^\star = \sum_{i=1}^n w_i^\star \delta(x_i), \qquad w^\star \in \operatorname*{argmin}_{w \geq 0,\, 1^\top w = 1} \mathcal{S}_P \left(\sum_{i=1}^n w_i \delta(x_i) \right)$$

使得 $Q_n^\star$ 与 P 之间的 Stein 差异最小化。具体而言，考虑第 4.4 节中使用 Stein 核 k_p 的核 Stein 差异： $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$。由于

$$\mathcal{S}_P \left(\sum_{i=1}^n w_i \delta(x_i) \right)^2 = \sum_{i=1}^n \sum_{j=1}^n w_i w_j k_p(x_i, x_j),$$

最优权重 $w^\star = (w_1^\star, \ldots, w_n^\star)^\top$ 即为以下线性约束二次规划的解

$$\operatorname*{argmin}_{w \in \mathbb{R}^d} \; w^\top K_p w \qquad \text{s.t.} \qquad w \geq 0,\; 1^\top w = 1 \tag{6.15}$$

其中 $[K_p]_{i,\,j} = k_p(x_i,\ x_j)$。式（6.15）中的约束确保 $Q_n^\star$ 是 $\mathcal{X}$上的概率分布。该规划没有闭式解，但可以通过数值方法求解。

这种方法最初由 Liu 和 Lee [2017]提出，当时被称为 黑箱重要性采样，因为与传统重要性采样（后者会对$x_i \sim \pi$ 进行抽样，并赋予 π相关的权重 $\tilde{w}_i \propto p(x_i)/\pi(x_i)$）不同，式（ x_i6.15）并不要求了解这些状态 x_i是如何生成的。该方法在贝叶斯统计中取得了成功，其中 P是通过某种（可能存在偏差的）MCMC 方法抽样得到的后验分布；在这种情况下，Stein 重要性采样可以减轻输入样本中的系统偏差，例如由预热期、退火、近似 MCMC 或其他偏离目标的采样造成的偏差。事实上，在

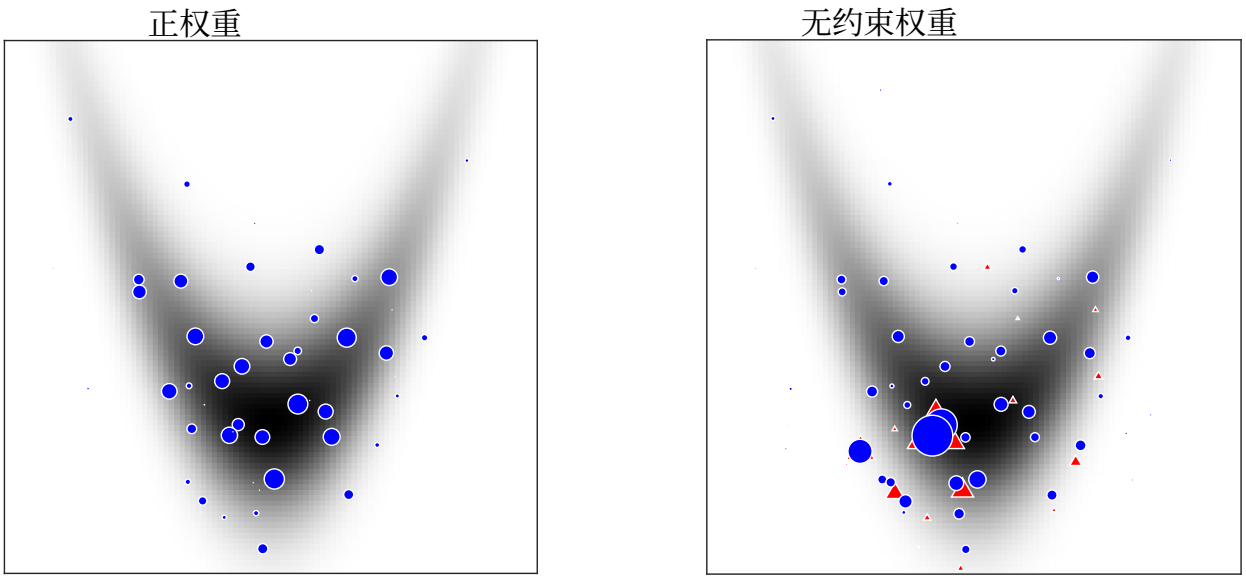

图 6.7： 使用正权重（左）和无约束权重（右）的 Stein重要性采样对 Rosenbrock 目标分布（6.3）（黑色阴影部分）进行了近似。每个标记的大小与对应状态所分配的权重成正比；蓝色圆圈表示正权重，红色三角形表示负权重。

在输入样本满足适当条件时，Stein 重要性采样近似$Q_n^\star$ 是一致的，具体而言，$\mathrm{S}_P(Q_n^\star) \to 0$ 几乎必然成立 [Riabiz等，2022年，定理 3]；$\mathrm{S}_P(Q_n^\star) = O(1/\sqrt{n})$ 依概率收敛 [Hodgkinson 等人，2020年，Li等，2024年，定理 1]。事实上，在更严格的条件下，Stein 重要性采样收敛到 P 的速度快于从 P 获取独立同分布样本的速度：$\mathbb{E}[\mathrm{S}_P(Q_n^\star)^2] = o(1/n)$ [Li等，2024年，定理 2]。

算法示于图 6.7 （左），我们首先从标准二元正态分布中独立抽取状态 x_i，然后为每个状态分配权重，以近似（6.3）中的 Rosenbrock 目标分布。与标准重要性采样相比，Stein重要性采样中分配的权重彼此相关；若抽样 x_i 和x_j 因偶然而过于接近，这种情况在 Stein重要性采样中会得到缓解，而在标准重要性采样中则不会。图 6.7 右侧面板展示了移除正性约束后 Stein重要性采样的行为；值得注意的是，负权重有助于提升核 Stein 差异意义下的近似效果。这一观察将在第6.3.5节中发挥作用。

受图 6.7 中 Stein重要性采样表现的鼓舞，我们可以进一步探讨：哪些状态 $\{x_1, \ldots, x_n\}$ 能使近似误差 $\mathrm{S}_P(Q_n^\star)$ 最小？这是一个具有挑战性且尚未解决的数学问题。即使假设这些状态是从分布 $\pi(x)$ 中独立抽取的，π 的最优选择也很可能依赖于 n。作为一种启发式方法，Wang等人 [2023] 基于对自归一化重要性采样的分析，提出取 $\pi(x) \propto p(x)\sqrt{k_p(x,x)}$。对于采用基核 $k(x,y) = \phi(x-y)$ 的 Langevin KSD，我们

有 $k_p(x,x) = -\Delta\phi(0) + \phi(0)\|\nabla \log p(x)\|^2$，该值通常会在 $\|x\| \to \infty$ 尾部增大。因此，这种 π 的选择可能会相对于 P 引入过度离散，并更加重视 P 中梯度幅值最高的区域。

6.3.4 稀疏近似

尽管 Stein 重要性采样优雅简洁，但当样本数量6.15较大时，求解（ n ）的计算负担可能相当沉重；粗略来说，计算成本为 $O(n^3)$。遗憾的是，对于基于大数定律的算法（例如 MCMC）， n 通常很大。我们可以从 MCMC输出中按固定间隔抽取较少数量的状态 $m \ll n$ （这通常称为稀疏化），但一般而言，这些子样本的近似质量可能较差。因此，自然的做法是转而从 MCMC输出中寻找大小为 m 的最优子集。

这个问题与第6.3.3节中的最优加权近似密切相关，因为我们可以寻找至多有 $m \ll n$ 个非零元素的权重向量。这一观察促成了多种稀疏近似技术，旨在迭代地构造形如 $Q_{n,m} = \frac{1}{m}\sum_{i=1}^{m} \delta(y_i)$ 的近似，其中每个 y_i 都是候选状态列表 $\{x_1, \ldots, x_n\}$中的一个元素，而这些候选状态可以是 MCMC输出。一个典型例子是贪心算法，它在第 j次迭代时选择一个状态

$$y_j \in \operatorname*{argmin}_{y \in \{x_1,\ldots,x_n\}} \mathcal{S}_P\left(\frac{1}{j}\delta(y) + \frac{1}{j}\sum_{i=1}^{j-1}\delta(y_i)\right) \tag{6.16}$$

使 Stein 差异最小化。在核 Stein 差异的背景下，贪心算法（6.16）称为 ***Stein***稀疏化 ，其计算成本为 $O(mn)$ 若采用适当的充分统计量实现 [Li等人，2024，附录 D。1]。此外，在适当的假设下，稀疏近似收敛到最优加权近似$Q_n^\star$ 当 $m \to \infty$ 且 n 固定时，因为

$$\mathcal{S}_P(Q_{n,m})^2 \le \mathcal{S}_P(Q_n^\star)^2 + \frac{\max_i k_p(x_i,x_i)(1+\log(m))}{m}. \tag{6.17}$$

因此，Stein稀疏化同时执行偏差校正和压缩，产生几乎与最佳单纯形加权输入同样精确的稀疏输出。详见 Riabiz等人 [2022] 以获取完整细节，以及 Teymur等人 [2021] ，其中还考虑了贪心算法的非短视和小批量扩展。

稀疏近似在预测建模中尤其有价值，因为为了进行概率预测，需要对模型参数的后验不确定性进行边缘化。对于复杂的预测模型，例如基于物理过程详细机理描述的模型，从模型进行前向模拟可能会产生相当高的计算成本。因此，能够生成后验分布的稀疏近似 P 意味着可以控制总体模拟成本。

保证（6.17）的一个缺点是，它要求 $m = \Omega(n)$ 个样本点，才能达到 $\mathsf{S}_P(Q_n^\star)$ 通常的 $O(1/\sqrt{n})$ 收敛速率。幸运的是，人们已经开发出许多扩展方法，只需显著更少的样本点即可恢复标准的 $O(1/\sqrt{n})$ 收敛速率。例如，Li等人的 Stein核稀疏化算法 [2024]能够以 $m = \sqrt{n}\,\mathrm{polylog}(n)$ 个点达到 $\mathsf{S}_P(Q_{n,m}) = O(1/\sqrt{n})$-误差；而通过使用加权近似，Li等人的 Stein重组和 Stein Cholesky 各自只需 $m = \mathrm{polylog}(n)$ 个点即可达到这一误差。

6.3.5 积分近似

通常，人们感兴趣的是将后验 P 近似作为计算感兴趣的后验期望的一个过渡步骤。事实上，给定 P的一个（可能带权的）经验近似$Q_n = \sum_{i=1}^n w_i \delta_{x_i}$，便可以构造相应的近似

$$\int f \,\mathrm{d}P \approx \sum_{i=1}^{n} w_i f(x_i) \tag{6.18}$$

来估计感兴趣的后验期望。然而，在这种两步方法中，权重 w_i 和状态 x_i 的选取并未考虑函数 $f\colon \mathcal{X} \to \mathbb{R}$。因此，自然会想知道，允许权重、状态或两者都依赖于 f，是否能得到更好的结果。

通过求解 Stein 方程进行积分

一种改进方法，已在早期工作中概述Stein [1986], ，是选取一个 Stein 算子 $\mathcal{T}_P$ 和一个 Stein集合 $\mathcal{G}$，并尝试求解*Stein* 方程

$$f(x) = c + (\mathcal{T}_P g)(x) \tag{6.19}$$

对于常数 $c \in \mathbb{R}$、元素 $g \in \mathcal{G}$，以及 P-几乎处处的 $x \in \mathcal{X}$。事实上，根据 Stein 算子的定义，我们便有 $c = \int f \,\mathrm{d}P$ 恰好就是所关注的积分。通过对 (6.19) 应用数值方法，可以得到依赖于 f的权重 w_i 和状态 x_i，用于如 (6.18) 所示的近似，下面将对此进行演示。

为便于说明，取$\mathcal{G}$ 为向量值再生核希尔伯特空间$\mathcal{H}(K)$中的单位球，并考虑最小范数插值

$$\operatorname*{argmin}_{c' \in \mathbb{R},\, g' \in \mathcal{G}} \|g'\|_{\mathcal{H}(K)} \quad \text{such that} \quad f(x_i) = c' + (\mathcal{T}_P g')(x_i), \quad \forall i \tag{6.20}$$

在固定的配点$\{x_i\}_{i=1}^n$处。假设配点互不相同，且 K 为正定，则满足（ $c_n(f)$）的 c' （6.20）的值可通过解析方法计算：

$$c_n(f) = \frac{1_n^\top K_p^{-1} f_n}{1_n^\top K_p^{-1} 1_n} \tag{6.21}$$

其中$[K_p]_{i,\ j} = k_p(x_i,\ \ x_j)$ 且 k_p 为 Stein 核（4.7），1_n 表示一个长度为 n的全 1 列向量，且 $[f_n]_i = f(x_i)$。

该控制泛函估计量（6.21）对应于如下形式的加权近似（6.18），其中权重为 $w \propto K_p^{-1}1_n$，并经归一化使得 $1_n^\top w = 1$。重要的是，这种非均匀加权能够对积分 $\int f \,\mathrm{d}P$ 进行更高保真度的估计。例如，对于适当的 f，$\mathcal{T}_P$，以及$\{x_i\}_{i=1}^n$从 P中独立抽取时，估计量（6.21具有如下性质：

$$\mathbb{E}\left[\left(\int f \,\mathrm{d}P - c_n(f)\right)^2\right] = O(n^{-7/6}) \tag{6.22}$$

均方误差 [Oates等人，2017年，定理 2], 严格优于 $\Theta(n^{-1})$ 标准蒙特卡罗积分的均方误差。权重 w 也可以理解为在移除正性约束后，对（6.15）中 Stein 重要性采样问题的一个松弛解；参见图6.7右图。

我们刚刚推导出的权重与 f无关，状态也只是固定的；如何将感兴趣的特定函数 f 纳入考虑？答案是通过依赖于f来选择核 K。核选择有许多实用方法，例如交叉验证；但为了帮助理解，我们在此考察误差界

$$\left|\int f \,\mathrm{d}P - c_n(f)\right| \leq \underbrace{\left(\inf_{\substack{f=c+\mathcal{T}_P g \\ c\in\mathbb{R},\ g\in\mathcal{G}}} \|g\|_{\mathcal{H}(K)}\right)}_{\text{(i)}} \underbrace{\mathcal{S}_P\left(\sum_{i=1}^n w_i\delta_{x_i}\right)}_{\text{(ii)}}$$

该方法将积分误差分解为两项的乘积：(i) 一项取决于真实函数 f 和核 K，但不取决于状态 $\{x_i\}_{i=1}^n$；(ii) 与经验分布 $\sum_{i=1}^n w_i\delta_{x_i}$作为对 P的近似相关联的 Stein 差异，该项与 f 无关。与 Stein 重要性采样仅在核固定时考虑第 (ii) 项不同，我们可以转而选择 K，同时平衡第 (i) 项和第 (ii) 项的乘积，从而使总体误差界最小化。第 (i) 项对 f的依赖性，进而使核 K的选择取决于 f；这可以显著提升所用权重在近似目标积分时的适用性。

注 6.3（Stein 控制变量）。*Stein 等人 [2004]* 采用了与此相关的视角来求解 *Stein* 方程，其形式是一种用于蒙特卡洛的方差缩减技术。事实上，标准的蒙特卡洛估计量

$$\frac{1}{m}\sum_{i=1}^m f(X_i), \qquad X_1,\ldots,X_m \overset{i.i.d.}{\sim} P$$

可以替换为另一种估计量

$$\frac{1}{m}\sum_{i=1}^m f(X_i) - (\mathcal{T}_P g)(X_i) \tag{6.23}$$

使用任意一个元素 g （来自 *Stein* 集合$\mathcal{G}$）。一个最优的 f或零方差选择（*Assaraf* 和 *Caffarel*, *2003*; *Mira* 等人, *2013*] 选择 g 对应于 *Stein* 方程(6.19)的一个解；而在实际应用中，可以寻求 g 以最小化对方差的估计(6.23)。对于涉及贝叶斯后验的应用，通常无法获得来自 P的精确样本，因此，直接求解 *Stein* 方程可能比尝试降低基于马尔可夫链蒙特卡洛的估计量方差更可取。另一方面，*Stein* 控制变量已成功用于梯度估计，正如我们将在第*6.5.1*节看到的那样。

已有多项研究推动了 Stein 方程数值求解方法的发展，包括 Oates 等人 [2017, 2019], Barp 等人 [2022], Si 等人 [2020],Belomestny 等人 [2020], South 等人 [2022a,b, 2023], Sun 等人 [2023], Leluc 等人 [2025],Belomestny 等人 [2024]。

通过SVGD**选择配点**

给定一个（可能依赖于 f 的）核 K，SVGD（第 6.3.2 节）提供了一种颇具吸引力的配点选择方法 $\{x_i\}_{i=1}^n$，用于积分近似；现在我们假设这些配点是 SVGD 的不动点，并讨论它们的性质。为简化讨论，我们假设 $K(x,y) = k(x,y)I$。根据更新规则（6.11），在 SVGD 的不动点处

$$\frac{1}{n}\sum_{j=1}^{n} \mathcal{T}_P^{(1)} k(x_j, x_i) = 0, \qquad \forall i = 1, \dots, n. \tag{6.24}$$

另一方面，根据 Stein 恒等式，$\mathbb{E}_{X\sim P}[\mathcal{T}_{(1)\,P}\ k(X, x_i)] = 0$。这意味着，这些粒子恰好估计了形如 $f(x) = \mathcal{T}_{(1)\,P}\ k(x, x_i)$ 的函数的期望，其中每个 $i = 1, \dots, n$ 均成立。根据期望的线性性，对于该集合的线性张成中的所有函数，同样的结论也成立。也就是说，

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) = P(f)$$

对于所有 $f \in \mathcal{F}^*$，即 *Stein* 匹配集

$$\mathcal{F}^* \triangleq \left\{ \sum_{i=1}^{n} a_i^\top \mathcal{T}_P^{(1)} k(\cdot, x_i) + b \colon \forall a_i \in \mathbb{R}^d,\ b \in \mathbb{R} \right\}.$$

注意，当配点$\{x_i\}_{i=1}^n$ 是 SVGD 的不动点集时，精确性成立，但不动点集未必唯一。进一步地，与$\mathcal{F}^*$接近的函数，其期望比与$\mathcal{F}^*$相距较远的函数的期望估计得更好。具体而言，令$\mathcal{F}_\xi^*$为$\mathcal{F}^*$的 ξ邻域，即$\mathcal{F}_\xi^* = \{f \colon \inf_{f' \in \mathcal{F}} \|f - f\|_\infty \le \xi\}$，则很容易证明

$$\left| \frac{1}{n}\sum_{i=1}^{n} f(x_i^*) - P(f) \right| \le 2\xi, \qquad \forall f \in \mathcal{F}_\xi^*,$$

参见Liu和Wang [2018]。

一个有趣的特殊情形是， P 服从多元高斯分布，且 k 是次数为 p 的多项式核（例2.15），此时 SVGD 产生的数值近似与 p 的前 P阶矩完全匹配。例如，若使用线性核$k(x, x') = 1 + x^\top x'$，且独立粒子的数量不少于 d （更准确地说，矩阵 $[x_i; 1]_{i=1}^n$ 的秩不少于 $d+1$），那么匹配集$\mathcal{F}^*$ 包含所有线性函数和二次函数，因此 SVGD 的任意不动点都能精确估计 P的均值和协方差矩阵。由于许多实际分布都接近高斯分布，得益于中心极限定理和伯恩斯坦–冯·米泽斯定理，这一观察可用于设计更优算法，使其在类高斯分布上表现尤为出色。

6.4 生成模型训练

本节聚焦于训练生成模型所面临的挑战。一类重要的基于能量的生成模型具有难以处理的归一化常数。对于这一类模型，Stein 差异和 Stein动力学可作为最大似然估计的有效替代方法，具体形式包括斯坦对比散度(第6.4.1节）和最小斯坦差异估计（第6.4.2节）。要生成超逼真的数据，通常会采用对抗方法；在此，Stein动力学为训练生成对抗网络（第6.4.3节）提供了一种互补的方法。在简单的共轭情形之外，贝叶斯方法往往在计算上具有挑战性；在此，Stein 差异可为变分贝叶斯方法（第6.4.4节）提供有用的目标函数；类似地，在变分自编码器（第6.4.5节）的设定中，Stein动力学能够支持表达能力更强的建模选择。

6.4.1 斯坦对比散度

在 $\mathbb{R}^d$ 上构建灵活概率模型的一种通用方法是，采用灵活的函数逼近器（例如深度神经网络），并对其应用正性约束变换。这会得到一种所谓的基于能量的 模型 P_θ，其密度形式为

$$p_\theta(x) = \frac{1}{Z_\theta} \exp(f_\theta(x)), \quad Z_\theta = \int \exp(f_\theta(x)) \mathrm{d}x, \tag{6.25}$$

其中 $f_\theta : \mathbb{R}^d \to \mathbb{R}$ 是一个灵活的（负）能量 函数，其参数为 $\theta \in \Theta$。给定一个数据集 $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$，一项重要任务是为参数$\theta$ 选择一个合适的值，使基于能量的模型能够生成在统计上与数据集中样本相似的样本（其确切含义取决于具体应用）。遗憾的是， P_θ 缺少归一化密度，这对使用最大似然等经典统计程序构成了重大障碍，因为其隐含的

（6.25）中的归一化常数通常依赖于 θ。事实上，若考察对数似然的梯度

$$\nabla_\theta \sum_{i=1}^{n} \log p_\theta(x_i) = \left[\sum_{i=1}^{n} f_\theta(x_i)\right] - \nabla_\theta \log Z_\theta$$

那么，第一项可以计算，但第二项

$$\begin{aligned}\nabla_\theta \log Z_\theta &= \frac{1}{Z_\theta}\nabla_\theta \int \exp(f_\theta(x))\mathrm{d}x \\ &= \frac{1}{Z_\theta}\int \exp(f_\theta(x))\nabla_\theta f_\theta(x)\mathrm{d}x = \mathbb{E}_{X\sim P_\theta}[\nabla_\theta f_\theta(X)]\end{aligned}$$

这需要计算关于 P_θ 的难解期望。令 $P_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ 表示数据集的经验分布，梯度上升的一步可写为

$$\theta \leftarrow \theta + \epsilon\{\mathbb{E}_{X\sim P_n}[\nabla_\theta f_\theta(X)] - \mathbb{E}_{X\sim P_\theta}[\nabla_\theta f_\theta(X)]\}$$

对于某个学习率 $\epsilon > 0$。直观而言，这条更新规则会迭代地提高观测数据中 $f_\theta(x)$ 的值，这些数据来自 P_n （或正样本）；同时降低从假设模型 P_θ （又称负样本）中抽取的数据对应的 $f_\theta(x)$ 的值。当算法收敛时，经验分布 P_n 和模型 P_θ 下 $\nabla_\theta f_\theta(x)$ 的期望应当相等。

为继续进行，实用算法需要近似关于P_θ的期望，例如可以使用马尔可夫链蒙特卡洛 [，例如 Geyer、 1991], 对比散度 [辛顿， 2002], 或变分推断 [，例如 Wainwright 等人，2008]。特别是，对比散度（CD）是一种简单的方法，它通过运行 m 步 P_θ以数据分布 P_n 为初始分布的 -不变马尔可夫转移核来近似模型期望项:

$$\theta \leftarrow \theta + \epsilon\{\underbrace{\mathbb{E}_{X\sim P_n}[\nabla_\theta f_\theta(X)] - \mathbb{E}_{X\sim K_m^\theta(P_n,\cdot)}[\nabla_\theta f_\theta(X)]}_{(\star)}\}, \tag{6.26}$$

其中， K_θ 是一个 P_θ-不变的马尔可夫转移核，而 $K_m^\theta(P_n, \cdot)$ 表示从 P_n 中随机抽取一个样本作为初始值后，经过 m 次马尔可夫链迭代所得到的分布。对于遍历马尔可夫链，$K_m^\theta(P_n, \cdot)$ 在 $m \to \infty$ 时收敛到 P_θ ，因此对比散度可恢复对数似然的梯度上升。不过在实践中，为节省计算，建议使用较小的 m。在这种情况下，增量（$\star$）可解释为（相差一个符号）$\mathbb{E}_{X\sim P_n}[\mathcal{T}_{P_\theta} g(X)]$, ，其中

$$(\mathcal{T}_{P_\theta} g)(x) = \int g(y) K_m^\theta(x, \mathrm{d}y) - g(x)$$

是根据离散时间马尔可夫链构造的 Stein 算子（参见第3.5.3节），并且$g(x) = \nabla_\theta f_\theta(x)$。

我们可以不使用马尔可夫链，而将 SVGD 更新用作对比散度中的扰动。也就是说，我们可以像在 SVGD 中那样，用确定性变换$x \mapsto x + \epsilon g_\theta(x)$扰动观测数据$\{x_i\}_{i=1}^n$，其中速度场

$$g_\theta(x) = g^*_{P_\theta, P_n}(x) = \frac{1}{n}\sum_{i=1}^n \mathcal{T}^{(1)}_{P_\theta} K(x_i, x) \tag{6.27}$$

与最速下降相关联，如（6.10）。注意，（6.27）中的 Stein 算子是 SVGD 中使用的 Langevin Stein 算子，它为对比散度中的马尔可夫转移提供了确定性对应，并自然地与 Stein 差异相联系，类似于对比散度与 Fisher散度相联系的方式。则 θ 的梯度更新为

$$\theta \leftarrow \theta + \epsilon\{\mathbb{E}_{X\sim P_n}[\nabla_\theta f_\theta(X) - \nabla_\theta f_\theta(X + \epsilon g_\theta(X))]\}, \tag{6.28}$$

其中 g_θ（x）表示 SVGD 的最优速度场。斯坦对比散度 [Liu和Wang, 2017] 使用了该扰动的一种对称变体

$$\theta \leftarrow \theta + \epsilon\{\mathbb{E}_{X\sim P_n}[\nabla_\theta f_\theta(X - \epsilon g^*_\theta(X)) - \nabla_\theta f_\theta(X + \epsilon g^*_\theta(X))]\}, \tag{6.29}$$

其中，对比是相对于扰动 $x \pm \epsilon g^*_\theta$（$x$）从两个相反方向进行的；其理由在第6.4.2节中说明。

6.4.2 最小Stein差异估计量

基于能量的模型（6.25）中，参数估计的另一种方法是最小化数据的经验分布 P_n 与参数模型 P_θ 之间的 Stein 差异。回顾第 4.1 节，Hyvärinen[2005]提出的（经典）得分匹配方法选择 θ，以最小化 Fisher散度

$$\mathrm{FD}(P_n||P_\theta) = [\text{constant in } \theta] + \frac{1}{n}\sum_{i=1}^n 2(\Delta_x f_\theta)(x_i) + \|(\nabla_x f_\theta)(x_i)\|^2,$$

该散度通过在（4.3）中用经验分布 P_n 替代 Q 得到。对 θ 进行梯度下降，得到 $\theta \leftarrow \theta - \epsilon g_{\mathrm{FD}}$，其中

$$g_{\mathrm{FD}} \triangleq \mathbb{E}_{X\sim P_n}[\nabla_\theta(\Delta_x f_\theta)(X) + (\nabla_x f_\theta)(X)\cdot\nabla_\theta(\nabla_x f_\theta)(X)]. \tag{6.30}$$

事实证明，这个更新规则是对比散度的极限，其中扰动定义为朗之万动力学的一步更新，且步长 ϵ 趋近于零：

$$\mathbb{E}_{X\sim P_n}[\nabla_\theta f_\theta(\hat{X}) - \nabla_\theta f_\theta(X)] = \epsilon g_{\mathrm{FD}} + o(\epsilon), \tag{6.31}$$

其中，$\hat{} X = X + \epsilon \nabla_x f_\theta(X) + \sqrt{2\epsilon} Z$是朗之万更新，$Z$服从标准正态分布。通过二阶泰勒展开可以看出这一点。令 $\delta = \epsilon \nabla_x f_\theta(X) + \sqrt{2\epsilon}\, Z$，使得$\hat{} X = X + \delta$。则

$$\nabla_\theta f_\theta(X + \delta) = \nabla_\theta f_\theta(X) + \nabla_x \nabla_\theta f_\theta(X)\, \delta + \tfrac{1}{2}\, \delta^\top \nabla_x^2 \nabla_\theta f_\theta(X)\, \delta + o(\|\delta\|^2).$$

对 Z 取期望，并利用 $\mathbb{E}[Z] = 0$ 和 $\mathbb{E}[ZZ^\top] = I$，我们得到

$$\begin{aligned} &\mathbb{E}_Z[\nabla_\theta f_\theta(\hat{X})] - \nabla_\theta f_\theta(X) \\ &\qquad = \epsilon(\nabla_x f_\theta(X) \cdot \nabla_\theta \nabla_x f_\theta(X) + \nabla_\theta \Delta_x f_\theta(X)) + o(\epsilon). \end{aligned}$$

将其代入（6.31），并对 $X \sim P_n$ 取期望，即可得到所述展开式。

Fisher散度只是 Stein 差异的一种；更一般地，最小斯坦差异估计量 定义为

$$\min_\theta \ \mathcal{S}_{P_\theta}(P_n),$$

对于某个Stein 差异 $\mathcal{S}_{P_\theta}$ [Liu和Wang，2017年，Liu 等人，2019，Barp 等人，2019]。例如，Langevin KSD（第4.4.1节）已被提出，可通过适当选择核，便捷地构造对数据集中的离群值具有鲁棒性的基于能量的模型估计量 [Barp 等人，2019]。对（平方）核 Stein 差异进行梯度下降，相当于使用以下梯度

$$\nabla_\theta \mathrm{KSD}^2_{P_\theta}(P_n) = 2\mathbb{E}_{X \sim P_n}[\nabla_\theta (\nabla_x f_\theta)(X) g_\theta(X)], \tag{6.32}$$

其中 g_θ 定义见（6.27）。

与经典分数匹配更新（6.30）相比，后者需要计算三阶导数 $\nabla_\theta \Delta_x f$，对（平方）核 Stein 差异进行梯度下降只涉及二阶导数，因此更容易实现。

此外，对（6.32）进行一阶泰勒展开，可得（6.28）中的斯坦对比散度更新：

$$\begin{aligned} &\mathbb{E}_{X \sim P_n}[\nabla_\theta f_\theta(\hat{X}_{\mathrm{SVGD}}) - \nabla_\theta f_\theta(X)] \\ &\qquad = \epsilon \mathbb{E}_{X \sim P_n}[\nabla_\theta (\nabla_x f_\theta)(X) g_\theta(X)] + o(\epsilon), \end{aligned}$$

其中 $\hat{} X_{\mathrm{SVGD}} = X + \epsilon g_\theta(X)$。因此，Stein 对比散度提供了一种对核 Stein 差异进行梯度下降的有限差分近似，只需计算一阶导数即可。这种有限差分解释说明了（6.29）中采用对称扰动的原因，因为中心有限差分近似与核 Stein 差异的梯度更加吻合。

最小 Stein 差异估计量的渐近统计性质可以通过 M-估计量 [Barp 等人，2019，Matsubara et al.，2022年，Oates，2022]这一框架来理解。基于有限斯坦集的最小 Stein 差异估计量这一具体情形由 Ebner et al. [2025]进行了研究。采用扩散 Stein 算子（见第 4.4.6 节）的最小核斯坦差异估计量也已应用于因果模型学习，并在该背景下进行了分析 [Lorch et al.，2024，Bleile et al.，2026]。

6.4.3 Stein GAN

对比散度已被广泛用于学习基于能量的模型 P_θ （6.25），并且通常能够训练出测试似然良好的模型。然而，通过对比散度训练的模型往往难以生成真正逼真的图像。这是因为对比散度根据观测数据邻域内的局部扰动来学习模型参数，并未明确训练模型从零开始生成图像。生成对抗网络（GAN）绕开了这一难题 [Goodfellow et al., 2014], 它们通过一个生成元 （一种接收随机噪声并输出图像的深度神经网络），在 判别器的帮助下，使生成数据与观测数据相匹配；该判别器以对抗方式区分生成数据与训练集中的数据。受 GAN 启发，我们可以改进对比散度的思路，在训练过程中明确引入生成元。

其思路是通过生成器对 P_θ 的抽样过程进行摊销，并将模拟生成的样本作为负样本来更新 θ。具体而言，设 G_β 为将随机噪声 ξ （例如标准高斯噪声）输入参数为 β的神经网络 $N_\beta(\cdot)$后得到的分布。我们首先将自适应地调整参数 β，使 G_β 尽可能接近模型 P_θ，然后使用生成器 G_β 代替模型 P_θ，通过 CD 更新 θ ，即:

$$\theta \leftarrow \theta + \epsilon\{\mathbb{E}_{X\sim P_n}[\nabla_\theta f_\theta(X)] - \mathbb{E}_{X\sim G_\beta}[\nabla_\theta f_\theta(X)]\}. \tag{6.33}$$

这里的关键问题是如何更新 β，使分布 G_β 能够紧密近似模型 P_θ。给定当前生成器 $\{\tilde{x}_i\}_{i=1}^m$ 生成的样本 G_β，即按 $\tilde{x}_i = N_\beta(\xi_i)$生成的样本，应用 SVGD 更新后得到新样本 $\{\tilde{x}_i'\}_{i=1}^m$，其中 $\tilde{x}_i' = \tilde{x}_i + \epsilon g_\theta(\tilde{x}_i)$应当能更准确地近似 P_θ。Feng 等人Feng 等人提出的摊销 SVGD 2017]的思路是，通过对最小二乘目标 β 执行梯度下降步骤来更新

$$\sum_{i=1}^m \|N_\beta(\xi_i) - \tilde{x}_i'\|^2,$$

从而鼓励生成元 G_β 逼近新的样本 $\{\tilde{x}_i\}_{i=1}^m$，这些样本能够更好地表示 P_θ。经过显式计算可得

$$\beta \leftarrow \beta + \epsilon \sum_{i=1}^m (\nabla_\beta N_\beta)(\xi_i) g_\theta(N_\beta(\xi_i)). \tag{6.34}$$

这称为 $SteinGAN$ [Liu和Wang, 2017]。形式上，这可以看作是基于 Kullback–Leibler散度（定义 2.26）近似求解以下极小极大目标函数:

$$\min_\theta \max_\beta \Big\{\mathrm{KL}(P_n\|P_\theta) - \mathrm{KL}(G_\beta\|P_\theta)\Big\}$$

此处，能量模型 P_θ作为判别器，试图在Kullback–Leibler散度意义下更接近观测数据 P_n，并远离“虚假”数据分布 G_β；而生成元 G_β 则试图利用摊销SVGD更接近能量模型 P_θ。

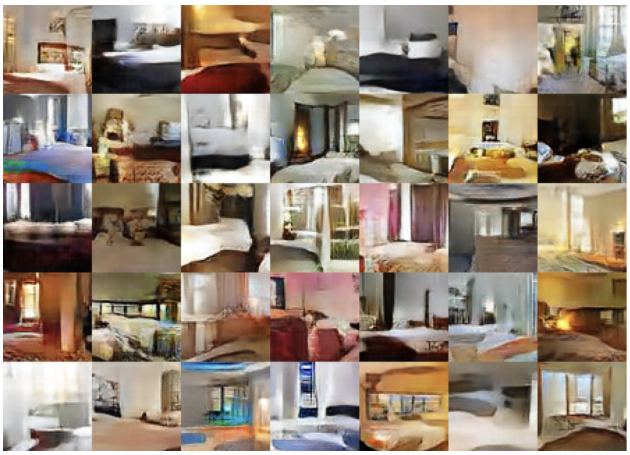

图6.8： 使用 LSUN 数据集训练的 SteinGAN 生成的图像 [Yu 等人， 2015], 该数据集包含近 300 万张卧室图像。经许可转载自Liu 和 Wang [2017]。

SteinGAN 可视为GAN式对抗博弈的一种Kullback–Leibler散度变体 [Goodfellow等人， 2014], 并明确关联到基于能量的模型的最大似然训练，其中难以处理的抽样（负相）通过经训练以逼近SVGD动力学的神经生成器进行摊销。SteinGAN所生成的逼真图像示例见图6.8。

6.4.4 后验近似的变分方法

用于贝叶斯分析的一类常用数值方法属于变分贝叶斯范畴；简言之，其思想是使用生成模型近似后验分布。这种方法有多项优势，其中最显著的是，由于后验的密度具有显式形式，因此可以轻松地将其用作后续分析的先验。主要的技术难题在于如何使用生成模型近似后验。

变分推断的经典方法试图通过从一个易于处理的分布族中进行选择，来近似目标后验 P，并从中选出一个分布 $\mathcal{Q}$，使 Kullback–Leibler散度（定义 2.26）

$$\underset{Q\in\mathcal{Q}}{\operatorname{argmin}}\ \mathrm{KL}(Q\|P), \tag{6.35}$$

达到最小值。令 $p(x,y)$ 表示未观测到的 x 与观测数据 y的联合密度，并类似地令 $p(x|y)$ 表示条件分布， $p(y)$ 表示边缘分布；

$q(x)$ 表示 Q的密度，则 Kullback–Leibler散度为

$$\begin{aligned}\mathrm{KL}(Q\|P) &= \int q(x)\log\frac{q(x)}{p(x|y)}\,\mathrm{d}x \\ &= \int q(x)\log\frac{q(x)}{p(x,y)}\,\mathrm{d}x + \log p(y)\end{aligned}$$

因此，最小化 Kullback–Leibler散度等价于最小化证据下界（ELBO）

$$\mathcal{L}_Q(y) = \int q(x)\log\frac{p(x,y)}{q(x)}\,\mathrm{d}x.$$

对于先验、似然和变分族 $\mathcal{Q}$的某些组合，ELBO 可以通过解析方式计算，因此可以将优化问题（6.35）作为数值优化任务来处理。然而，这种对“共轭”组合的限制，限制了变分近似在一般情况下准确反映后验的程度。对于非共轭情形，可以使用蒙特卡洛方法近似 ELBO [及其梯度；例如，参见 Ruiz 等人， 2016], 但这些方法需要显式访问 Q 的密度（及其梯度），由此推动了对归一化流 [Rezende 和 Mohamed，2015年，Kingma 等人，2016，Dinh 等人， 2017]的研究。

Stein 差异提供了另一种解决方案，无需访问 Q 的密度。也就是说，我们可以将其表述为

$$\operatorname*{argmin}_{Q\in\mathcal{Q}} \mathcal{S}_P(Q) \tag{6.36}$$

并选择一种 Stein 差异，使得 $\mathcal{S}_P(Q)$ 可以精确计算，也可以通过一致近似得到（例如，利用来自 Q 的样本）。这使得将（6.36）作为数值优化任务来求解成为可能。特别是，这消除了对 Q 具有显式密度的限制，并支持灵活的近似形式，例如 $Q_\theta = {T^\theta}_{\#}Q_0$，其中 T^θ 是由 θ 参数化的通用神经网络，而 Q_0 是某个参考分布。这种构造与 Langevin KSD 结合使用的研究见 Fisher 等人 [2021], ；他们发现，在正则情形下，通过最小化核 Stein 差异得到的近似，其准确度不低于通过最小化 Kullback–Leibler散度得到的近似；参见图 6.9。

一种相关方法称为 算子变分推断，由 Ranganath 等人 [2016],提出，它考虑了（6.36）中的同一个极小极大问题，即

$$\min_\theta \max_{g\in\mathcal{G}} \mathbb{E}_{X\sim Q_\theta}[\mathcal{T}_P g(X)]$$

其中 $\mathcal{T}_P$ 和 $\mathcal{G}$ 分别是与 Stein 差异 $\mathcal{S}_P$相关联的 Stein 算子和 Stein 集合。然而，作者转而提出使用神经网络对其进行参数化 g ，随后交替执行随机梯度步：先更新 θ，再更新该神经网络的参数 g。原则上，这使 Stein 集合 $\mathcal{G}$具有更大的灵活性，从而能够更好地检测 Q_θ 与 P 之间的差异，但代价是需要调试更复杂的数值优化方法。

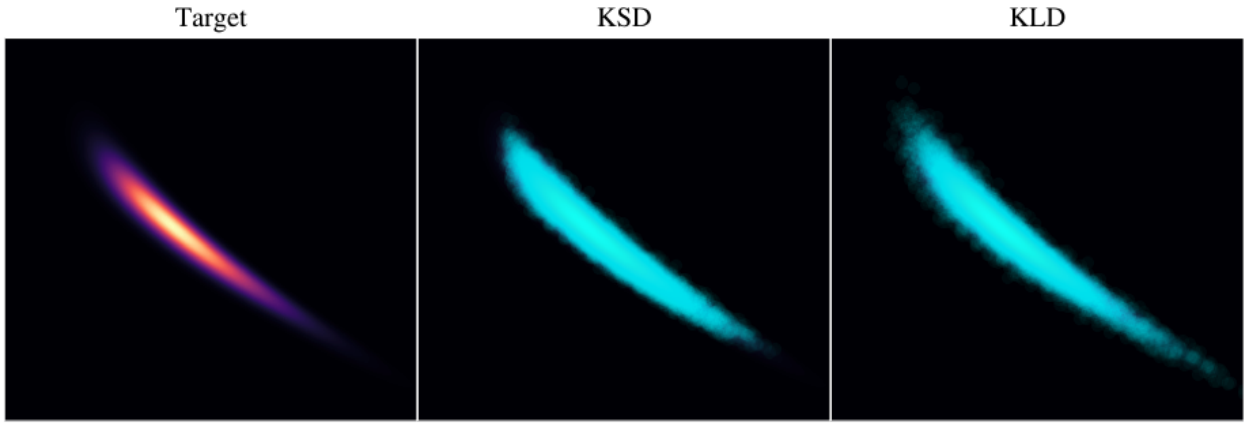

图 6.9： 对目标的变分近似比较 P（左），分别通过最小化核 Stein 差异（中）和 Kullback–Leibler散度（右）得到。该变分近似基于块神经自回归流，而目标 P 来自对生化需氧量模型的贝叶斯分析。经许可转载自 Fisher 等人 [2021]。

6.4.5 学习变分自编码器

构建表达能力强的生成模型，一种优雅的方法是从扩展状态空间上的参数化生成模型入手；设 $p_\theta(x, z)$ 是一个参数化的概率密度，其中 $x \in \mathbb{R}^d$ 是可观测变量，$z \in \mathbb{R}^{d'}$ 是潜变量。即使 $p_\theta(x, z)$ 是一个相对简单的模型，其边缘分布的密度为

$$p_\theta(x) = \int p_\theta(x, z)\, \mathrm{d}z \tag{6.37}$$

能够刻画所观测变量之间相当复杂的依赖关系。因此，目标是学习 θ，例如基于数据集 $\{x_i\}_{i=1}^n$ 使用最大似然。然而，这可能颇具挑战，因为（6.37）一般不具有闭式。

要继续进行，需要显式的近似。一种解决方案是训练一个编码器网络 $q_\phi(z|x)$，以近似条件分布 $p_\theta(z|x)$。这是变分自编码器（VAE）背后的主要思想，即最大化证据下界

$$\mathcal{L}_{\theta,\phi}(x) = \int q_\phi(z|x) \log \frac{p_\theta(x, z)}{q_\phi(z|x)}\, \mathrm{d}z \tag{6.38}$$

联合训练模型和编码器参数 θ 以及 ϕ。事实上，我们可以将（6.38）改写为

$$\mathcal{L}_{\theta,\phi}(x) = \log p_\theta(x) - \mathrm{KL}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$$

这清楚地表明，我们在提升似然 $p_\theta(x)$ 较大值的同时，也要求编码器 $q_\phi(z|x)$ 能够良好地近似条件分布 $p_\theta(z|x)$。然而，使用（6.38）通常需要 $q_\phi(z|x)$具有可处理密度，这与我们此前遇到的问题相同

在第6.4.4节讨论变分贝叶斯时遇到过这一问题。因此，编码器网络的表达能力，以及 VAE 发现隐藏结构的潜力，都可能受到限制。

一种可能的解决方案是 Stein 变分自编码器 [Pu 等人，2017年，Feng 等人， 2017]。Stein 变分自编码器的思路是，不指定编码器网络，而是使用 SVGD 通过一组粒子 $\{z_{i,\ j}\}_{j=1}^{m}$隐式表示条件分布 $p_\theta(z|x_i)$。与标准变分自编码器相比，后者假设 $q\phi(\cdot \mid x)$ 是在给定 x 条件下的高斯分布，而非参数粒子近似能够捕获更丰富的后验结构。

Feng 等人提出的方法 [2017] 随后根据以下方式更新模型参数 θ：

$$\theta \leftarrow \theta + \epsilon \sum_{i=1}^{n} \sum_{j=1}^{m} \nabla_\theta \log p_\theta(x_i, z_{i,j}),$$

这可以解释为对联合似然进行梯度上升，其中将这些配对 $(x_i,\ z_{i,\ j})$ 视为一个伪数据集。根据 Fisher 恒等式

$$\nabla_\theta \log p_\theta(x) = \mathbb{E}_{z\sim p_\theta(z|x)}\Big[\nabla_\theta \log p_\theta(x, z)\Big].$$

因此，如果粒子 $\{z_{i,\ j}\}_{j=1}^{m}$ 是来自后验分布 $p_\theta(z \mid x_i)$ 的精确样本，那么上述更新将恢复边缘对数似然$\sum_i \log p_\theta(x_i)$ 的精确梯度（存在蒙特卡洛误差）。为加快计算速度，可以训练一个识别网络 $N_\beta(\cdot)$ 来模仿 SVGD 粒子动力学，即：

$$z_{ij} = N_\beta(\xi_j, x_i), \qquad \forall i = 1, \ldots, n, \quad j = 1, \ldots, m,$$

其中 ξ_i 从噪声分布（例如标准高斯分布）中采样，而参数β 则基于最小二乘法进行训练，并按照式（6.34）通过梯度下降实现。这种摊销 避免了对数据集中的每个条目分别运行 SVGD 的需要 x_i。

6.5 梯度估计

数值计算梯度的计算任务

$$\nabla_\phi \mathbb{E}_{X\sim Q_\phi}[f(X)] \tag{6.39}$$

关于分布 Q_ϕ 的参数 ϕ，在许多学习问题中都会遇到这种情况，包括变分推断 [（参见第 6.4.4 节和 Paisley 等人， 2012]）以及训练变分自编码器 [（参见第 6.4.5 节和 Kingma 和 Welling， 2013], ）；在这些问题中，人们希望对证据下界执行梯度上升；此外，在强化学习中，人们也希望通过遵循策略梯度来改进策略 [（参见第 6.5.2 节和 Williams， 1992]）。从根本上说，困难在于无法准确计算期望（6.39），并且计算 f 的成本很高。因此，许多作者提出了数值方法，使用少量 m 样本（通常为 $m \in \{1, 2, 3\}$）来近似计算式（6.39）；我们将看到，在这一背景下，斯坦算子能够发挥有益作用，重点讨论 Q_ϕ 为离散分布的情形（第 6.5.1 节），以及强化学习应用中所遇到的策略梯度情形（第 6.5.2 节）。

6.5.1 使用离散 Stein 算子进行梯度估计

本节重点讨论 Q_ϕ 是在集合 $\mathcal{X}$上的离散分布，且 $\mathcal{X}$ 的规模使得精确计算（6.39）变得不可行。在这种情况下，一种常用策略是将梯度改写为

$$\begin{aligned}\nabla_\phi \mathbb{E}_{X\sim Q_\phi}[f(X)] &= \nabla_\phi \sum_{x\in\mathcal{X}} f(x) q_\phi(x) \\ &= \sum_{x\in\mathcal{X}} f(x) \frac{\nabla_\phi q_\phi(x)}{q_\phi(x)} q_\phi(x) \\ &= \mathbb{E}_{X\sim Q_\phi}[f(X)\nabla_\phi \log q_\phi(X)] \end{aligned} \tag{6.40}$$

其中 q_ϕ 是 Q_ϕ的概率质量函数。对（6.40）直接进行蒙特卡洛近似，

$$\frac{1}{m}\sum_{i=1}^{m} f(X_i)\nabla_\phi \log q_\phi(X_i), \qquad X_1,\dots,X_m \overset{\text{i.i.d.}}{\sim} Q_\phi$$

是（6.39）的无偏估计量，但除非使用极大量的 m蒙特卡洛样本，否则可能具有很高的方差。一种简单的降方差技巧称为 REINFORCE [Glynn，1990，Williams，1992]，

$$\frac{1}{m}\sum_{i=1}^{m}[f(X_i) - B_i]\nabla_\phi \log q_\phi(X_i) \tag{6.41}$$

只要满足以下条件，就能保持无偏性：

$$\mathbb{E}[B_i \nabla_\phi \log q_\phi(X_i)] = 0, \tag{6.42}$$

也就是说，只要 $B_i\nabla_\phi \log q_\phi(X_i)$ 是一个控制变量，即一个均值已知（此处为零）的随机变量。幸运的是，控制变量条件（6.42）很容易构造。例如，$B_i^{\text{ind}}\nabla_\phi \log q_\phi(X_i)$ 在 B_i^{ind}与 X_i 相互独立时均值为零，因为

$$\begin{aligned}\mathbb{E}[B_i^{\text{ind}}\nabla_\phi \log q_\phi(X_i)] &= \mathbb{E}[B_i^{\text{ind}}] \sum_{x\in\mathcal{X}} \frac{\nabla_\phi q_\phi(X)}{q_\phi(X)} q_\phi(X) \\ &= \mathbb{E}[B_i^{\text{ind}}] \sum_{x\in\mathcal{X}} \nabla_\phi q_\phi(X) \\ &= \mathbb{E}[B_i^{\text{ind}}]\nabla_\phi \sum_{x\in\mathcal{X}} q_\phi(X) = \mathbb{E}[B_i^{\text{ind}}]\nabla_\phi 1 = 0.\end{aligned}$$

类似地，只要 $\mathbb{E}[B_i^{\text{zero}} \mid X_i] = 0$，$B_i^{\text{zero}}\nabla_\phi \log q_\phi(X_i)$ 的均值就为零，因为根据塔式性质，

$$\mathbb{E}[B_i^{\text{zero}}\nabla_\phi \log q_\phi(X_i)] = \mathbb{E}[\mathbb{E}[B_i^{\text{zero}} \mid X_i]\nabla_\phi \log q_\phi(X_i)] = 0.$$

更一般地， $B_i = B_i^{\text{ind}} + B_i^{\text{zero}}$也满足协变量条件（6.42），接下来我们将充分利用这种分解的灵活性。

这一原始 REINFORCE 算法将每个 B_1，……， B_m 设为一个常数，其取值被选择为使（6.41）的方差最小化。对于 $m > 1$ ，一种流行的变体是通过将 B_i 替换为函数值的留一平均而得到的，从而得到**R**EINFORCE **L**eave-**O**ne-**O**ut（RLOO）估计器 [Salimans 和 Knowles，2014年，Kool 等人，2019，Richter 等人， 2020]，

$$\frac{1}{m}\sum_{i=1}^{m}\left(f(X_i) - \frac{1}{m-1}\sum_{j\neq i} f(X_j)\right)\nabla_\phi \log q_\phi(X_i). \tag{6.43}$$

RLOO 估计同样在构造上对目标梯度是无偏的。然而，尽管它改进了先前的估计量，除非使用大量 m 样本，RLOO 的方差仍可能顽固地居高不下。

斯坦算子可以作为一种有用的工具，在不牺牲无偏性的情况下进一步降低 RLOO 的方差 [Shi 等人，2022b]。其思想是，我们希望为 B_i 设计一个更好的选择，理想情况下使得 $B_i \approx f(X_i)$，从而使（6.41）的方差相应地较小。为此，令 T^{Q_ϕ} 为 Q_ϕ 的 Stein 算子；第 3.5 节讨论了若干可能的离散 Stein 算子。然后，由于 $\mathbb{E}[(\mathrm{T}^{Q_\phi} g_j)(X_j) \mid X_i] = 0$ 对 $j \neq i$ 以及 T^{Q_ϕ} 定义域中的任意 g_j 成立，我们可以考虑取 B_i 为如下形式

$$B_i = B_i^{\text{ind}} + B_i^{\text{zero}} = \frac{1}{m-1}\sum_{j\neq i} f(X_j) + \frac{1}{m-1}\sum_{j\neq i}(\mathcal{T}_{Q_\phi} g_j)(X_j). \tag{6.44}$$

由于我们的目标是近似 $f(X_i)$，原则上可以通过求解 Stein 方程（6.19）来为每个 g_i 选择合适的

$$f(x) = \mathbb{E}_{X\sim Q_\phi}[f(X)] + (\mathcal{T}_{Q_\phi} g)(x), \qquad \forall x \in \mathcal{X},$$

正如第6.3.5节所讨论的那样，但对于我们考虑的情形，$\mathcal{X}$ 的规模会使直接求解 Stein 方程变得不切实际。作为一种可行的替代方案，Shi 等人[2022b]建议使用 Gibbs Stein 算子（3.20），并选择 g_i 具有以下形式

$$g_j(x) = \frac{1}{m-1}\sum_{k\neq j} N_\beta(f(X_k), \nabla f(X_k)^\top (x - X_k)) \tag{6.45}$$

其中， N_β 是一个参数为 β 、需要进行优化的神经网络。请注意，尽管每个函数 g_j 都依赖于 X_i，我们仍有 $\mathbb{E}[(\mathcal{T}^{Q_\phi} g_j)(X_j) \mid X_i] = 0$，因此该梯度估计量仍然无偏。因此，可以通过对梯度估计量的样本方差进行梯度下降来优化 β，这与 Grathwohl 等人 [2018]的方法类似。在各分量 X 取值较少（例如 $\mathcal{X} = \{0,1\}^d$）的情形下，使用 Gibbs Stein 估计量的动机是数值稳定性；而在各分量 X 取值各不相同的情形下，推荐使用 Barker Stein 算子（这是……的一个特例

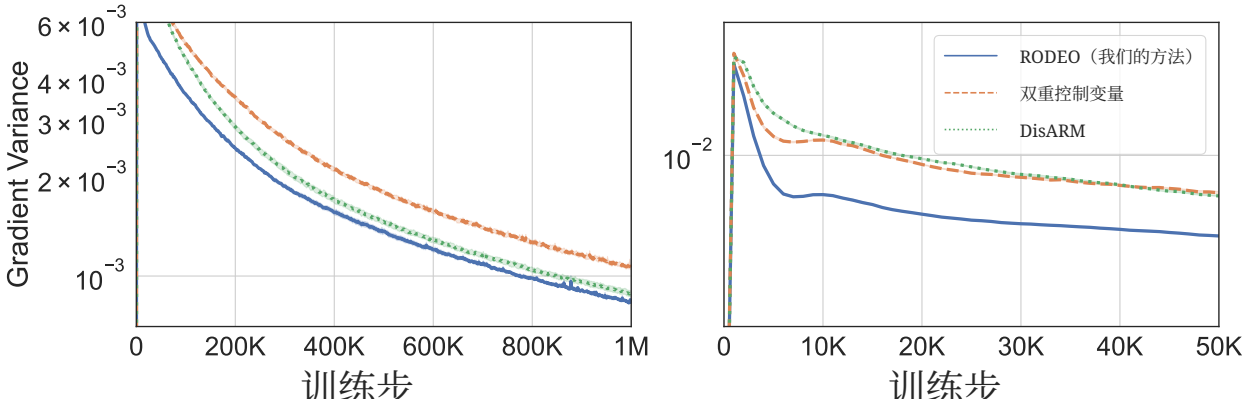

图 6.10： 在 Fashion MNIST 数据集上训练变分自编码器；采用了维度为 z 的二元潜变量$d' = 200$，并使用 $m = 2$ 个样本来评估每个梯度。经授权转载自 Shi 等人 [2022b]。

Zanella Stein 算子；定义 3.8），Shi 等人 [2022b]对此作出了推荐。此外，在某些应用中（例如变分自编码器；第 6.4.5节），计算 ∇f 无需额外成本 [Titsias 和 Shi， 2022]，这意味着，计算（6.45）所需的 f 评估次数不多于 RLOO。

在式（6.44）中的局部 Stein 控制变量（参见注 6.3）之上，Shi 等人 [2022b]引入了第二组全局 Stein 控制变量，并加入额外学习得到的检验函数$\tilde{g}_k$，其形式为（6.45），以进一步缩减方差：

$$\frac{1}{m}\sum_{k=1}^{m}\Big[(f(x_k) - \frac{1}{m-1}\sum_{j\neq k}(f(x_j) + (\mathcal{T}_{Q_\phi} g_j)(x_j)))\nabla_\eta \log q_\eta(x_k) \tag{6.46}$$

$$+ (\mathcal{T}_{Q_\phi}\tilde{g}_k)(x_k)\Big]. \tag{6.47}$$

这种将 Stein 方程用作双重方差缩减工具的整体方法，在 Shi 等人 [2022b] 的工作中被称为 RODEO（基于离散 Stein 算子的 RLOO）。作者将 RODEO 应用于训练二元潜变量变分自编码器（参见第 6.4.5 节以及 Kingma 和 Welling， 2013]）。潜变量的维度为 z，并使用了 Fashion MNIST 数据集（Xiao 等人， 2017]）。从图 6.10 可以看出，与 DisARM（Dong 等人， []）和双重控制变量（Titsias 和 Shi， 2022]）等其他梯度估计方案相比，RODEO 显著降低了训练路径上的梯度方差。这反过来通常会提升变分自编码器的性能；更多结果请参见 Shi 等人 [2022b]。

6.5.2 策略梯度的控制变量

我们介绍的最后一个应用涉及强化学习，其起始点是一个马尔可夫决策过程，由一组状态 $\mathcal{S}$、一组动作 $\mathcal{A}$以及一个马尔可夫

过程 $(S_t)_{t=0}^{\infty} \subset \mathcal{S}$，使得 S_{t+1} 仅取决于 S_t 和一个动作 $A_t \in \mathcal{A}$。此外，还有一系列称为奖励 $R(S_t, A_t)$ 的随机变量，它们仅取决于当前状态 S_t 和动作 A_t。强化学习旨在寻找一个（随机）策略 $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$，即在时间 t 决定采取哪个动作时，我们对 $A_t \sim \pi(S_t)$进行采样，其期望折扣累积奖励

$$J(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \middle| S_0 = s_0\right]$$

达到最大值。这里将初始状态 s_0 视为固定值，而 $\gamma \in (0, 1)$ 是一个折扣因子，用于优先考虑奖励序列中较早的部分，其值由用户指定。

为简化表述，设$\mathcal{S}$ 为一个可数集，设$\mathcal{A} = \mathbb{R}^d$，并设 P_π 为折扣状态访问分布，其概率质量函数为

$$p_\pi(s) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}_\pi(S_t = s | S_0 = s_0),$$

其中，$\mathbb{P}_\pi$ 表示动作 A_t 按照策略 $\pi(S_t)$选取，因此我们可以使用简写

$$J(\pi) = \mathbb{E}_\pi[R(S, A)] = \mathbb{E}_{S \sim P_\pi}[\mathbb{E}_{A|S \sim \pi(S)}[R(S, A)|S]].$$

在下文中，我们会稍微扩展符号的用法，使得对每个 $s \in \mathcal{S}$，$\pi(\cdot|s)$ 表示分布 $\pi(s)$ 的密度函数。

在策略梯度方法中，策略 π 由参数化表示，记作 π_θ，参数 θ 通过对 $\theta \mapsto J(\pi_\theta)$进行梯度上升来更新。根据策略梯度定理[Williams, 1992], 该目标的梯度为

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\pi_\theta}\left[Q^{\pi_\theta}(S_t, A_t)(\nabla_\theta \log \pi_\theta)(A_t|S_t)\right],$$

其中，动作价值函数 $Q^\pi(s, a)$ 表示初始状态为$s \in \mathcal{S}$ 且初始动作为 $a \in \mathcal{A}$时的期望回报，之后每一步均采用策略 π。随后，在动作价值函数已知或已估计出的前提下，可以使用蒙特卡洛方法近似该梯度。一项关键挑战是，对该梯度进行蒙特卡洛估计往往会遭遇高方差。

为解决这一问题，通常会采用 REINFORCE 等技术（参见第 6.5.1 节），将其转化为一个等价的期望，并降低相应蒙特卡洛估计量的方差。直接应用可得

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}\left[\left(Q^{\pi_\theta}(S, A) - B(S)\right)(\nabla_\theta \log \pi_\theta)(A|S)\right],$$

其中 $B(S)$ 取决于状态 S，但不取决于动作 A。通过类似于（6.43）的论证，可以确保梯度的值不变。B 的常见取值是常数，或 价值函数 $B(s) = \max_{a \in \mathcal{A}} Q(s, a)$ [Sutton 和 Barto, 1998]。然而，由于

这些传统基线仅基于状态 S 的函数，因此它们大幅降低方差的能力有限。Stein 恒等式使得 B 可以同时依赖于状态和动作，而不改变梯度的值。

对于任意元素 t g_s 属于 $\pi_\theta(s)$ 的 Langevin Stein 算子定义域内（定义3.2），

$$\begin{aligned} 0 &= \mathbb{E}_{A\sim\pi_\theta(s)}[(\mathcal{T}_{\pi_\theta(s)}g_s)(A)] \\ &= \mathbb{E}_{A\sim\pi_\theta(s)}\left[g_s(A)\cdot(\nabla\log\pi_\theta)(A|s) + (\nabla g_s)(A)\right] = 0. \end{aligned}$$

然而，此处的梯度是对动作 A求取的，而策略梯度涉及对参数 θ求取梯度。重参数化技巧可以弥合这一差距：也就是说，如果我们能将从 $\pi_\theta(s)$ 采样的动作表示为 $A = f_\theta(s,\xi)$，其中ξ 是与 θ 和 s相互独立的随机噪声，那么Liu 等人 [的定理 3.1 2018] 表明

$$\mathbb{E}_{A\sim\pi_\theta(s)}\left[g_s(A)\cdot(\nabla_\theta\log\pi_\theta)(A|s)\right] = \mathbb{E}\left[\nabla_\theta f_\theta(s,\xi)\cdot(\nabla g_s)(f_\theta(s,\xi))\right].$$

其中，最后一个期望是关于创新噪声 ξ取的。这就引出了一种基于 Stein 算子的 REINFORCE 替代方法，

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\pi_\theta}\left[\begin{array}{r}(Q^{\pi_\theta}(S,A) - B(S,A))(\nabla_\theta\log\pi_\theta)(A|S) \\ +\nabla_\theta f_\theta(S,\xi)\cdot\nabla_A B(S,A)\end{array}\right],$$

其中 $B(s,a) = g_s(a)$ 且每个 g_s 都属于 $\pi(s)$对应的 Langevin Stein 算子定义域，其中每个 $s \in\mathcal{S}$。由此，我们推导出了依赖于状态和动作的函数，这些函数可以通过优化来降低相关蒙特卡洛估计量的方差，同时确保所得估计量仍然无偏。Liu 等人Liu 等人 [2018] 的实证研究发现，在一系列强化学习应用中，与较简单的依赖状态的 $B(S,A)$ 相比，使用 $B(S)$可以显著降低方差，从而提高样本效率。

7 结论

本专著汇集了严谨的定义与结果，它们为 Stein方法近期及新兴的方法学应用奠定了基础。通过此举，我们希望为概率推断和学习领域的从业者提供了一份便捷且全面的参考资料。由于这一主题仍是活跃的研究领域，我们至多只能希望呈现写作之时该领域知识现状的一个缩影；我们对 Stein方法方法学方面的观点和理解必将进一步发展。事实上，即便在这一点上，我们也无意作出真正全面的论述，而是将范围限定于最经典的 Stein差异以及 Stein方法最直接的方法学应用。目前已有大量且日益增长的文献，其中一些我们曾顺带提及，希望读者能受到启发，深入探索。

参考文献

L. Ambrosio、N. Gigli 和 G. Savaré。《梯度流：度量空间与概率测度空间中的梯度流》。Springer，2005年。A. Anastasiou、A. Barp、F.-X. Briol、B. Ebner、R. E. Gaunt、F. Ghaderinezhad、J. Gorham、A. Gretton、C. Ley、Q. Liu、L. Mackey、C. J. Oates、G. Reinert 和 Y. Swan。Stein方法与统计学相遇：近期若干进展综述。《style id='3'>统计科学》，38(1)：120–139，2023年。R. Assaraf 和 M. Caffarel。量子蒙特卡洛中观测量的零方差零偏差原理：力的应用。《style id='5'>化学物理杂志，119(20)：10536–10552，2003年。F. Bach、S. Lacoste-Julien 和 G. Obozinski。论herding算法与条件梯度算法的等价性。载于《style id='7'>第29届国际机器学习会议论文集》，2012年。S. Banerjee、K. Balasubramanian 和 P. Ghosal。Stein变分梯度下降的有限粒子收敛速率改进。载于《style id='9'>第13届国际学习表征会议》，2025年。A. D. Barbour。Stein方法与泊松过程收敛。《style id='11'>应用概率杂志，25(A)：175–184，1988年。A. D. Barbour。扩散逼近的Stein方法。《style id='13'>概率论及相关领域，84(3)：297–322，1990年。A. D. Barbour 和 L. H. Y. Chen。《style id='15'>Stein方法导论》。World Scientific，2005年。A. A. Barker。质子－电子等离子体径向分布函数的蒙特卡洛计算。《style id='17'>澳大利亚物理学杂志，18(2)：119–134，1965年。A. Barp、F.-X. Briol、A. Duncan、M. Girolami 和 L. Mackey。最小Stein差异估计量。载于《style id='19'>第33届神经信息处理系统会议论文集》，2019年。A. Barp、C. J. Oates、E. Porcu 和 M. Girolami。一种Riemann–Stein核方法。《style id='21'>Bernoulli，28(4)：2181–2208，2022年。

A. Barp、C.-J. Simon-Gabriel、M. Girolami 和 L. Mackey。利用核差异实现目标分布分离与收敛。《style id='1'>机器学习研究期刊，25(378)：1–50，2024年。J. Baum、H. Kanagawa 和 A. Gretton。用于序列模型拟合优度的核Stein检验。载于《style id='3'>第40届国际机器学习会议论文集》，2023年。D. Belomestny、L. Iosipoi、E. Moulines、A. Naumov 和 S. Samsonov。马尔可夫链的方差缩减及其在马尔可夫链蒙特卡洛中的应用。《style id='5'>《统计与计算》，30：973–997，2020年。D. Belomestny、A. Goldman、A. Naumov 和 S. Samsonov。马尔可夫链蒙特卡洛中神经控制变量的理论保证。《style id='7'>数学与计算机仿真，220：382–405，2024年。D. M. Blei、A. Kucukelbir 和 J. D. McAuliffe。变分推断：统计学家综述。美国统计学会会刊，112(518)：859–877，2017年。F. Bleile、S. Lumpp 和 M. Drton。利用Stein型差异高效学习平稳扩散，2026年。I. M. Bomze、F. Rinaldi 和 D. Zeffiro。Frank–Wolfe方法及其相关方法：投影自由一阶优化方法探究。《style id='11'>运筹学年鉴，343(2)：607–638，2024年。G. Bresler 和 D. Nagaraj。马尔可夫链平稳分布的Stein方法及其在伊辛模型中的应用。《style id='13'>应用概率年刊，29(5)：3230–3265，2019年。F.-X. Briol、T. Karvonen、A. Gessner 和 M. Mahsereci。闭式核均值嵌入词典。载于《style id='15'>第1届国际概率数值会议论文集》，2025年。S. Brooks、A. Gelman、G. Jones 和 X.-L. Meng。《style id='17'>马尔可夫链蒙特卡洛手册》。CRC出版社，2011年。T. C. Brown 和 A. Xia。Stein方法与生灭过程。《style id='19'>概率年刊，29(3)：1373–1403，2001年。G. Buttazzo、V. Ferone 和 B. Kawohl。凹函数集合上的最小化问题及相关问题。《style id='21'>数学通讯，173(1)：71–89，1995年。C. Carmeli、E. De Vito 和 A. Toigo。可积函数的向量值再生核希尔伯特空间与Mercer定理。《style id='23'>分析与应用，4(04)：377–408，2006年。C. Carmeli、E. De Vito、A. Toigo 和 V. Umanitá。向量值再生核希尔伯特空间与普适性。《style id='25'>分析与应用，8(01)：19–61，2010年。

K.-F. Chang。严格正定函数。逼近理论期刊，87(2): 148–158，1996年。C. Chen、R. Zhang、W. Wang、B. Li 和 L. Chen。用于可扩展贝叶斯采样的统一粒子优化框架。载于第*34*届人工智能不确定性会议论文集，2018a。L. H. Chen、L. Goldstein 和 Q.-M. Shao。斯坦因方法的正态逼近。施普林格科学与商业媒体，2010a。P. Chen、K. Wu、J. Chen、T. O'Leary-Roseberry 和 O. Ghattas。投影斯坦变分牛顿法：一种高维快速且可扩展的贝叶斯推断方法。载于第*33*届神经信息处理系统会议论文集，2019a。W. Y. Chen、L. Mackey、J. Gorham、F.-X. Briol 和 C. J. Oates。Stein 点。载于第*35*届国际机器学习会议论文集，2018b。W. Y. Chen、A. Barp、F.-X. Briol、J. Gorham、M. Girolami、L. Mackey 和 C. J. Oates。斯坦因点马尔可夫链蒙特卡罗。载于第*36*届国际机器学习会议论文集，2019b。Y. Chen、M. Welling 和 A. Smola。核驱赶生成的超样本。载于第*26*届人工智能不确定性会议论文集，2010b。P. Chew。一个几乎与完全图一样好的平面图。载于第*2*届年度计算几何研讨会论文集，1986。S. Chewi、T. Le Gouic、C. Lu、T. Maunu、P. Rigollet 和 A. Stromme。镜像朗之万扩散的指数遍历性。载于第*34*届神经信息处理系统会议论文集，2020年。K. Chwialkowski、H. Strathmann 和 A. Gretton。核拟合优度检验。载于第*33*届国际机器学习会议论文集，2016。K. P. Chwialkowski、A. Ramdas、D. Sejdinovic 和 A. Gretton。利用概率测度的解析表示进行快速双样本检验。载于第*29*届神经信息处理系统会议论文集，2015。P. Clement 和 W. Desch。Wasserstein 度量三角不等式的初等证明。美国数学会论文集，136(1):333–339，2008。C. Conca 和 M. Vanninathan。不可压缩流体方程中的周期均匀化问题。数学流体动力学手册，4:649–698，2007年。

G. Detommaso、T. Cui、Y. Marzouk、R. Scheichl 和 A. Spantini。斯坦变分牛顿方法。载于第*22*届神经信息处理系统会议论文集，2018年。L. Dinh、J. Sohl-Dickstein 和 S. Bengio。使用 Real NVP 进行密度估计。载于第*5*届国际学习表征会议论文集，2017年。Z. Dong、A. Mnih 和 G. Tucker。DisARM：用于二元潜变量的反变量梯度估计量。载于第*34*届神经信息处理系统会议论文集，2020年。R. M. Dudley。实分析与概率。CRC出版社，2018年。A. B. Duncan、T. Lelievre 和 G. Pavliotis。使用非可逆朗之万采样器进行方差缩减。统计物理学杂志，163(3):457–491，2016。N. Dunford。向量值函数的积分。美国数学会通报，第43页，1937年。A. Eberle。扩散过程的反射耦合与收缩率。概率论及相关领域，166:851–886，2015。B. Ebner、A. Fischer、R. E. Gaunt、B. Picker 和 Y. Swan。Stein 矩方法。斯堪的纳维亚统计学杂志，52(4):1594–1624，2025年。P. Eichelsbacher 和 G. Reinert。离散吉布斯测度的 Stein 方法。应用概率年刊，18(4):1588–1618，2008。S. N. Ethier。群体遗传学中出现的一类退化扩散过程。纯粹与应用数学通讯，29(5):483–493，1976年。Y. Feng、D. Wang 和 Q. Liu。利用摊销式 Stein变分梯度下降学习生成样本。载于第*33*届人工智能不确定性会议论文集，2017年。M. Fisher、T. Nolan、M. Graham、D. Prangle 和 C. J. Oates。利用核 Stein 差异进行测度传输。载于第*24*届国际人工智能与统计会议论文集，第1054–1062页，2021年。M. A. Fisher 和 C. J. Oates。无梯度核Stein差异。载于第*37*届神经信息处理系统会议论文集，2023年。M. Fromont、B. Laurent、M. Lerasle 和 P. Reynaud-Bouret。用于双样本问题的基于核的检验与非渐近自助法。载于第*25*届学习理论会议论文集，2012年。

F. Futami、Z. Cui、I. Sato 和 M. Sugiyama。通过贪心粒子优化进行贝叶斯后验近似。载于*AAAI*人工智能会议论文集，第33卷，第3606–3613页，2019。M. A. Gallegos-Herrada、D. Ledvinka 和 J. S. Rosenthal。马尔可夫链几何遍历性的等价性。理论概率杂志，37(2):1230–1256，2024。A. Gelman、J. Carlin、H. Stern、D. Dunson、A. Vehtari 和 D. Rubin。贝叶斯数据分析。统计科学丛书。CRC出版社，博卡拉顿，佛罗里达州，第三版，2014。ISBN 978-1-4398-4095-5。C. J. Geyer。马尔可夫链蒙特卡洛最大似然。载于计算机科学与统计学：计算机科学与统计学接口第*23*届研讨会论文集，1991。P. W. Glynn。随机系统的似然比梯度估计。*ACM*通讯，33(10):75–84，1990。C. Gong、J. Peng 和 Q. Liu。用于并行贝叶斯优化的分位数Stein变分梯度下降。载于第*36*届国际机器学习会议论文集，2019。W. Gong、Y. Li 和 J. M. Hernández-Lobato。切片核化Stein差异。载于第*9*届学习表征会议论文集，2021。I. Goodfellow、J. Pouget-Abadie、M. Mirza、B. Xu、D. Warde-Farley、S. Ozair、A. Courville 和 Y. Bengio。生成对抗网络。载于第*28*届神经信息处理系统会议论文集，2014。J. Gorham 和 L. Mackey。用 Stein方法衡量样本质量。载于第*29*届神经信息处理系统会议论文集，第226–234页，2015。J. Gorham 和 L. Mackey。用核衡量样本质量。载于第*34*届国际机器学习会议论文集，2017。J. Gorham、A. B. Duncan、S. J. Vollmer 和 L. Mackey。用扩散衡量样本质量。应用概率年刊，29(5):2884–2928，2019。J. Gorham、A. Raj 和 L. Mackey。随机 Stein 差异。载于第*37*届神经信息处理系统会议论文集，2020。F. Gotze。关于多元中心极限定理的收敛速度。概率年刊，19(2):724–739，1991。

W. Grathwohl、D. Choi、Y. Wu、G. Roeder 和 D. Duvenaud。穿过虚空的反向传播：优化控制变量以进行黑箱梯度估计。载于第*6*届国际学习表征会议论文集，2018。O. Hagrass、B. Sriperumbudur 和 K. Balasubramanian。基于核 Stein 差异的极小极大最优拟合优度检验。*Bernoulli*，32(1):299–324，2026。J. Han 和 Q. Liu。Stein变分自适应重要性采样。载于第*33*届人工智能不确定性会议论文集，2017。J. Han 和 Q. Liu。无梯度的Stein变分梯度下降。载于第*37*届国际机器学习会议论文集，2018。S. Har-Peled 和 M. Mendel。低维度量空间中网的快速构造及其应用。载于第*21*届年度计算几何研讨会论文集，2005。S. G. Henderson。通过近似马尔可夫过程实现方差缩减。博士论文，斯坦福大学，1997。G. E. Hinton。通过最小化对比散度训练专家乘积。神经计算，14(8):1771–1800，2002。L. Hodgkinson、R. Salomone 和 F. Roosta。用于事后校正采样的再生 Stein 核方法，2020。S. Holmes。生灭链的Stein方法。载于*Stein*方法：专题讲座与应用，第45–67页。数理统计学会，2004。A. Horowitz。二阶朗之万方程与数值模拟。《核物理*B*》，280:510–522，1987。J. Huggins 和 L. Mackey。随机特征 Stein 差异。载于第*32*届神经信息处理系统会议论文集，2018。A. Hyvärinen。通过得分匹配估计非归一化统计模型。机器学习研究期刊，6(4):695–709，2005。N. Ikeda 和 S. Watanabe。《随机微分方程与扩散过程》。爱思唯尔，2014。W. Jitkrittum、W. Xu、Z. Szabó、K. Fukumizu 和 A. Gretton。线性时间核拟合优度检验。载于第*31*届神经信息处理系统大会论文集，2017。

H. Kanagawa、W. Jitkrittum、L. Mackey、K. Fukumizu 和 A. Gretton。《用于比较潜变量模型的核Stein检验》。英国皇家统计学会会刊*B*辑：统计方法论，85(3):986–1011，2023年。H. Kanagawa、A. Barp、A. Gretton 和 L. Mackey。《利用核 Stein 差异控制矩》。应用概率年刊，35(6):3818–3843，2025年。S. Karlin 和 J. McGregor。《生灭过程的分类》。美国数学会汇刊，86(2):366–400，1957年。D. P. Kingma 和 M. Welling。《自编码变分贝叶斯》。载于第*1*届国际学习表征会议论文集，2013年。D. P. Kingma、T. Salimans、R. Jozefowicz、X. Chen、I. Sutskever 和 M. Welling。《利用逆自回归流改进变分推断》。载于第*30*届神经信息处理系统会议论文集，2016年。W. Kool、H. V. Hoof 和 M. Welling。《购买4个 REINFORCE 样本，免费获赠一个基线！》。载于*DeepRLStructPred@ICLR*研讨会论文集，2019年。A. Korba、P.-C. Aubin-Frankowski、S. Majewski 和 P. Ablin。《核Stein差异下降》。载于第*40*届国际机器学习会议论文集，2021年。C. Landim、S. Olla 和 H. Yau。《具有时空遍历随机流的对流—扩散方程》。概率论及相关领域，112(2):203–220，1998年。H. Le、A. Lewis、K. Bharath 和 C. Fallaize。《黎曼流形上 Stein 方法的扩散方法》。*Bernoulli*，30(2):1079–1104，2024年。R. Leluc、F. Portier、J. Segers 和 A. Zhuman。《加速蒙特卡洛积分：用于实现最优收敛的控制近邻》。*Bernoulli*，31(2):1160–1180，2025年。A. Leucht 和 M. H. Neumann。《退化 U 统计量和 V统计量的相依野生自助法》。多元分析杂志，117:257–280，2013年。C. Ley 和 Y. Swan。《Stein 密度方法与信息不等式》。《概率电子通信》，18(7):1–14，2013年。C. Ley、G. Reinert 和 Y. Swan。《用于比较一元分布的 Stein 方法》。概率综述，14:1–52，2017年。L. Li、Y. Li、J.-G. Liu、Z. Liu 和 J. Lu。《用于高效抽样的随机版 Stein变分梯度下降》。应用数学与计算科学通讯，15(1):37–63，2020年。

L. Li、R. Dwivedi 和 L. Mackey。《去偏分布压缩》。载于 R. Salakhutdinov、Z. Kolter、K. Heller、A. Weller、N. Oliver、J. Scarlett 和 F. Berkenkamp 编，第*41*届国际机器学习会议论文集，第235卷，收录于机器学习研究论文集，第27675–27731页。PMLR，2024年7月21日至27日。网址：https://proceedings.mlr.press/v235/li24r.html。C. Liu 和 J. Zhu。《用于贝叶斯推断的黎曼斯坦变分梯度下降》。载于第*32*届*AAAI*人工智能会议论文集，2018年。H. Liu、Y. Feng、Y. Mao、D. Zhou、J. Peng 和 Q. Liu。《通过 Stein 恒等式实现用于策略优化的动作依赖控制变量》。载于第*6*届国际学习表征会议论文集，2018年。J. Liu、N. Mehta、C. Tao 和 L. Carin。《利用斯坦分数匹配对未归一化统计模型进行估计与抽样》。载于*ICML 2019*机器学习与统计学斯坦方法研讨会，2019年。Q. Liu 和 J. Lee。《黑箱重要性采样》。载于第*20*届国际人工智能与统计会议论文集，2017年。Q. Liu 和 D. Wang。《Stein变分梯度下降：一种通用的贝叶斯推断算法》。载于第*30*届神经信息处理系统年会论文集，2016年。Q. Liu 和 D. Wang。《深度能量模型学习：对比散度与摊销最大似然估计》，2017年。Q. Liu 和 D. Wang。《将 Stein变分梯度下降视为矩匹配》。载于第*32*届神经信息处理系统会议论文集，2018年。Q. Liu、J. Lee 和 M. Jordan。《用于拟合优度检验的核化斯坦差异》。载于第*35*届国际机器学习会议论文集，2016年。X. Liu 和 F.-X. Briol。《关于核拟合优度检验的稳健性》。 机器学习研究期刊，26(262)，2025年。X. Liu、A. Duncan 和 A. Gandy。《利用扰动改进基于核化斯坦差异的拟合优度检验》。载于第*40*届国际机器学习会议论文集，2023年。L. Lorch、A. Krause 和 B. Schölkopf。《利用平稳扩散进行因果建模》。载于第*27*届国际人工智能与统计会议论文集，2024年。Y. Ma、T. Chen 和 E. Fox。《随机梯度马尔可夫链蒙特卡罗的完整方案》。载于第*29*届神经信息处理系统会议论文集，2015年。

L. Mackey 和 J. Gorham. 一类强对数凹分布的多元 Stein 因子。《概率电子通信》，21:1–14，2016。D. Martinez-Taboada 和 A. Ramdas. 序贯核化 Stein 差异。载于第*28*届国际人工智能与统计会议，2025。T. Matsubara、J. Knoblauch、F.-X. Briol 和 C. J. Oates. 不可处理似然的稳健广义贝叶斯推断。《皇家统计学会杂志：*B*辑》，84 (3):997–1022，2022。T. Matsubara、J. Knoblauch、F.-X. Briol 和 C. J. Oates. 离散不可处理似然的广义贝叶斯推断。美国统计学会会刊，119(547): 2345–2355，2024。A. Mira、R. Solgi 和 D. Imparato. 贝叶斯估计量的零方差马尔可夫链蒙特卡罗。《统计与计算》，23(5):653–662，2013。K. Muandet、K. Fukumizu、B. Sriperumbudur 和 B. Schölkopf. 分布的核均值嵌入：综述及后续进展。《基础与趋势》® 《机器学习》，10 (1-2):1–141，2017。A. Müller. 积分概率度量及其生成函数类。《应用概率进展》，29(2):429–443，1997。R. M. Neal. 使用马尔可夫链蒙特卡罗方法进行概率推断。1993。M. E. Newman 和 G. T. Barkema.《统计物理中的蒙特卡罗方法》。克拉伦登出版社，1999。C. J. Oates. 最小核差异估计量。载于科学计算中的蒙特卡罗与拟蒙特卡罗方法第*15*届国际会议论文集，2022。C. J. Oates、M. Girolami 和 N. Chopin. 用于蒙特卡罗积分的控制泛函。《英国皇家统计学会杂志》，*B*辑，79:695–718，2017。C. J. Oates、J. Cockayne、F.-X. Briol 和 M. Girolami. 基于 Stein 方法的一类估计量的收敛速率。*Bernoulli*，25(2):1141–1159，2019。B. Oksendal.《随机微分方程：导论及其应用》。施普林格科学与商业媒体，2013。A. B. Owen. 对可扩展求积规则的一项约束。《数值数学》，132: 511–518，2016。J. Paisley、D. M. Blei 和 M. I. Jordan. 通过随机搜索进行变分贝叶斯推断。载于第*31*届国际机器学习会议论文集，2012。

S. Patterson 和 Y. Teh. 概率单纯形上的随机梯度黎曼朗之万动力学。载于第*27*届神经信息处理系统大会论文集，2013。G. A. Pavliotis. 随机过程及其应用。Springer，2016。D. Peleg 和 A. A. Schäffer. 图跨度器。图论杂志，13(1):99–116，1989。W. F. Pfeffer.《散度定理与有限周长集》。CRC出版社，博卡拉顿，2012。S. Pigola 和 A. G. Setti. 非线性偏微分方程与几何学中的全局散度定理。数学论文集，26(1-77):2，2014。Y. Pu、Z. Gan、R. Henao、C. Li、S. Han 和 L. Carin. Stein变分自编码器。载于第*31*届神经信息处理系统大会论文集，2017。X. Qu 和 B. C. Vemuri. 黎曼流形上核 Stein 差异的理论与应用，2025。A. Rahimi 和 B. Recht. 面向大规模核机器的随机特征。载于第*21*届神经信息处理系统大会论文集，2007。R. Ranganath、D. Tran、J. Altosaar 和 D. Blei. 算子变分推断。载于第*30*届神经信息处理系统大会论文集，2016。G. Reinert 和 N. Ross. 快速混合 Glauber 动力学平稳分布的近似及其在指数随机图中的应用。应用概率年刊，29(5):3201–3229，2019。L. Rey-Bellet 和 K. Spiliopoulos. 不可逆朗之万采样器与方差缩减：一种大偏差方法。非线性，28(7):2081–2103，2015。D. Rezende 和 S. Mohamed. 使用归一化流进行变分推断。载于第*34*届国际机器学习会议论文集，2015。M. Riabiz、W. Chen、J. Cockayne、P. Swietach、S. A. Niederer、L. Mackey 和 C. J. Oates. MCMC输出的最优稀疏化。《英国皇家统计学会杂志》，*B*辑，84 (4):1059–1081，2022。L. Richter、A. Boustati、N. Nüsken、F. Ruiz 和 O. D. Akyildiz. VarGrad：一种低方差梯度估计量，用于变分推断。载于第*34*届神经信息处理系统会议论文集，2020。N. Ross. Stein方法基础。概率综述，8:210–293，2011。

W. Rudin. 实分析与复分析. McGraw-Hill, Inc.，美国，第3版，1987。F. J. Ruiz、M. K. Titsias与D. M. Blei。广义重参数化梯度。载于第*30*届神经信息处理系统大会论文集，2016。T. Salimans与D. A. Knowles。论将控制变量与随机逼近用于变分贝叶斯及其与随机线性回归的联系，2014年。A. Schrab、B. Guedj与A. Gretton。KSD聚合拟合优度检验。载于 第*36*届神经信息处理系统大会论文集，2022年。S. Schwabik与G. Ye。巴拿赫空间积分专题。World Scientific，2005年。X. Shao。相依野生自助法。美国统计学会会刊，105 (489):218–235，2010。J. Shi与L. Mackey。Stein变分梯度下降的有限粒子收敛速率。神经信息处理系统进展，36:26831–26844，2023年。J. Shi、C. Liu与L. Mackey。使用镜像Stein算子进行抽样。载于 第*10*届国际学习表征会议论文集，2022a。J. Shi、Y. Zhou、J. Hwang、M. Titsias与L. Mackey。基于离散 Stein 算子的梯度估计。载于 第*36*届神经信息处理系统大会论文集，2022b。S. Si、C. J. Oates、A. B. Duncan、L. Carin与F.-X. Briol。通过随机优化实现蒙特卡洛方法的可扩展控制变量。载于 第*14*届科学计算中蒙特卡洛与准蒙特卡洛方法国际会议论文集，2020年。C.-J. Simon-Gabriel与B. Schölkopf。核分布嵌入：分布上的全称核、特征核与核度量。机器学习研究期刊，19(44):1–29，2018年。C.-J. Simon-Gabriel、A. Barp、B. Schölkopf与L. Mackey。用最大均值差异度量弱收敛。机器学习研究期刊，24(184):1–20，2023年。J. Sohl-Dickstein、P. Battaglino与M. R. DeWeese。最小概率流学习，2011。L. F. South、T. Karvonen、C. Nemeth、M. Girolami与C. J. Oates。源自Sard方法的半精确控制泛函。《生物统计学》，109(2):351–367，2022a。L. F. South、M. Riabiz、O. Teymur与C. J. Oates。MCMC后处理。《统计学及其应用年度评论》，9(1):529–555，2022b。

L. F. South、C. J. Oates、A. Mira与C. Drovandi。正则化零方差控制变量。贝叶斯分析，18(3):865–888，2023年。B. K. Sriperumbudur、K. Fukumizu与G. R. Lanckriet。普适性、特征核与测度的再生核希尔伯特空间嵌入。机器学习研究期刊，12:2389–2410，2011。B. K. Sriperumbudur、K. Fukumizu、A. Gretton、B. Schölkopf与G. R. Lanckriet。论积分概率度量的经验估计。电子统计学期刊，6: 1550–1599，2012年。C. Stein。相依随机变量之和的分布的正态近似误差界。载于 第*6*届伯克利数学统计与概率研讨会论文集，第*2*卷：概率论，第583–602页。加州大学出版社，1972。C. Stein。期望的近似计算。讲义—专著系列。数理统计学会，1986。C. Stein、P. Diaconis、S. Holmes、G. Reinert等。可交换对在模拟分析中的应用。载于 *Stein*方法，第1–25页。数理统计学会，2004。I. Steinwart与A. Christmann。支持向量机。施普林格科学与商业媒体，2008。I. Steinwart、D. Hush与C. Scovel。高斯径向基函数核的再生核希尔伯特空间的显式描述。*IEEE*信息论汇刊，52(10): 4635–4643，2006年。A. Stuart、J. Voss、P. Wilberg等。随机微分方程的条件路径采样与朗之万马尔可夫链蒙特卡罗方法。数学科学通讯，2(4):685–697，2004。Z. Sun、A. Barp与F.-X. Briol。向量值控制变量。载于 第*42*届国际机器学习会议论文集，2023年。R. S. Sutton与A. G. Barto。强化学习：导论。麻省理工学院出版社，1998。O. Teymur、J. Gorham、M. Riabiz与C. J. Oates。使用最大均值差异对概率测度进行最优量化。载于 第*24*届国际人工智能与统计会议论文集，2021年。X. Tian、N. Bi与J. Taylor。Magic：一种通用、强大且易于处理的选择性推断方法。*arXiv*预印本 *arXiv:1607.02630*，2016。

M. Titsias 和 J. Shi. 用于离散潜变量模型梯度估计的双重控制变量. 收录于第*25*届人工智能与统计国际会议论文集，2022年。M. J. Wainwright、M. I. Jordan 等. 图模型、指数族与变分推断. 基础与趋势® 机器学习，1(1–2):1–305，2008年。C. Wang、W. Chen、H. Kanagawa 和 C. J. Oates. Stein π-重要性采样. 收录于第*37*届神经信息处理系统会议论文集，2023年。D. Wang 和 Q. Liu. 学习生成样本：及其在生成对抗学习摊销最大似然估计中的应用，2016年。D. Wang 和 Q. Liu. 用于学习多样化混合模型的非线性 Stein 变分梯度下降. 收录于第*36*届国际机器学习会议论文集，2019年。D. Wang、Z. Zeng 和 Q. Liu. 连续图模型的 Stein 变分消息传递. 收录于第*34*届国际机器学习会议论文集，2017年。D. Wang、Z. Tang、C. Bajaj 和 Q. Liu. 使用矩阵值核的 Stein 变分梯度下降. 收录于第*33*届神经信息处理系统会议论文集，2019年。M. Welling 和 Y. W. Teh. 通过随机梯度朗之万动力学进行贝叶斯学习. 收录于国际机器学习会议，2011年。H. Wendland. 散乱数据逼近. 剑桥大学出版社，2004年。R. J. Williams. 用于连接主义强化学习的简单统计梯度跟随算法. 机器学习，8(3):229–256，1992年。G. Wynne 和 A. B. Duncan. 函数型数据的核双样本检验. 机器学习研究期刊，23(73):1–51，2022年。G. Wynne、M. J. Kasprzak 和 A. B. Duncan. 核 Stein 差异的傅里叶表示及其在无限维希尔伯特空间上的测度拟合优度检验中的应用. *Bernoulli*，31(2):868–893，2025年。H. Xiao、K. Rasul 和 R. Vollgraf. Fashion-MNIST：用于机器学习算法基准测试的新型图像数据集，2017年。T. Xifara、C. Sherlock、S. Livingstone、S. Byrne 和 M. Girolami. Langevin扩散与 Metropolis 调整的朗之万算法. *Statistics & Probability Letters*，91:14–19，2014年。

W. Xu 和 T. Matsuda. 方向分布的 Stein 拟合优度检验. 收录于第*23*届人工智能与统计国际会议论文集，2020年。W. Xu 和 T. Matsuda. 黎曼流形上的可解释 Stein 拟合优度检验. 收录于第*38*届国际机器学习会议论文集，2021年。J. Yang、Q. Liu、V. A. Rao 和 J. Neville. 通过 Stein 差异检验离散分布的拟合优度. 收录于第*35*届国际机器学习会议论文集，2018年。F. Yu、Y. Zhang、S. Song、A. Seff 和 J. Xiao. LSUN：利用人在回路中的深度学习构建大规模图像数据集，2015年。G. Zanella. 离散空间中局部马尔可夫链蒙特卡罗的知情提议. 美国统计学会会刊，115(530):852–865，2019年。K. S. Zhang、G. Peyré、J. Fadili 和 M. Pereyra. 镜像朗之万蒙特卡罗的 Wasserstein 控制. 收录于第*33*届学习理论年会论文集，2020年。J. Zhuo、C. Liu、J. Shi、J. Zhu、N. Chen 和 B. Zhang. Stein 变分梯度下降的消息传递. 收录于第*35*届国际机器学习会议论文集，2018年。