利用流形结构的高效贝叶斯推理方法研究

（申请清华大学工学博士学位论文）

培 养 单 位：计 算 机 科 学 与 技 术 系
学　　科：计 算 机 科 学 与 技 术
研　究　生：刘　畅
指 导 教 师：朱　军　教　授

二〇一九年六月

A Study on Efficient Bayesian Inference Methods Using Manifold Structures

Dissertation Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in

Computer Science and Technology

by

Liu Chang

Dissertation Supervisor : Professor Zhu Jun

June, 2019

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：(1) 已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；(2) 为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容；(3) 根据《中华人民共和国学位条例暂行实施办法》，向国家图书馆报送可以公开的学位论文。

本人保证遵守上述规定。

（保密的论文在解密后应遵守此规定）

作者签名：_____　　导师签名：_____

日　　期：_____　　日　　期：_____

摘 要

贝叶斯模型因其灵活的建模能力和稳定的学习表现使得它在人工智能及机器学习领域中得到了广泛应用，而当前大数据环境的特征则为贝叶斯模型的学习过程，即贝叶斯推理，带来了新的挑战和需求。多样的数据形式要求贝叶斯推理方法可高效处理变量的流形结构，复杂的模型结构需要推理方法具有更强的近似灵活性，繁重的后续任务要求推理方法的粒子高效性，而巨大的数据规模则需要推理方法可高效利用推理时间和计算资源。另外，新的高效推理方法的设计与开发也需要对现有种类繁多的方法其背后的根本原理和联系进行分析。而流形这一数学概念因其具有包容性的定义、丰富多样的结构和对本质几何特征的描述，可为分析和解决这些问题提供根本的视角和有力的工具。本文针对这些挑战和需求，利用数据和模型的显式或隐式的流形结构，面向马尔可夫链蒙特卡罗（Markov chain Monte Carlo, MCMC）及基于粒子的变分推理（particle-based variational inference, ParVI）这两个贝叶斯推理的关键领域，为增强贝叶斯推理方法的高效性展开理论和实践上的研究。具体贡献包括：

1. 提出了随机梯度测地线 MCMC 方法，使得针对流形变量的 MCMC 方法处理大规模数据的时间效率有了本质提高；

2. 开发了黎曼-斯坦因变分梯度下降方法，提高了现有 ParVI 方法的迭代效率，并为处理流形变量的贝叶斯推理任务首次带来了兼具近似灵活性和粒子高效性的 ParVI 方法；

3. 分析了 ParVI 方法所依赖的假设，揭示了现有各 ParVI 方法之间的联系，并依此理论开发了两个新的 ParVI 方法以及可适用于所有 ParVI 方法的加速框架和带宽选择方法，增强了 ParVI 方法的算力效率和粒子高效性；

4. 建立了将一般 MCMC 方法描述为沃瑟斯坦空间上的流（flow）的统一理论框架，系统地解释了现有各 MCMC 方法的行为机理，并将其与 ParVI 方法建立了一般性的联系，进而依此理论开发了两个新的 ParVI 方法，提高了 ParVI 方法的算力效率和 MCMC 方法的粒子高效性。

关键词： 贝叶斯推理；马尔可夫链蒙特卡罗方法；变分推理方法；微分流形；信息几何

Abstract

Bayesian models have drawn notable attention in artificial intelligence and machine learning, for their flexible modeling ability and remarkable robustness in learning. Its learning task, i.e. Bayesian inference, is faced with new challenges and requirements in this Big Data era. The variety of data formats poses the demand for Bayesian inference methods to efficiently tackle variables with manifold structures; powerful but complicated models require a high approximation flexibility of inference methods; intense downstream tasks need particle efficiency; and the immense amount of data asks inference methods for efficiency in training time and computation resource. Moreover, designing and developing new inference methods also require a clear knowledge on the fundamental principles of and relations among various existing methods. On the other hand, the mathematical concept of manifold could provide a fundamental perspective and powerful tools for these problems, thanks to its inclusive definition, rich structures and reflection on the intrinsic geometry of a space. In addressing all these challenges and requirements on the efficiency of inference methods from various aspects, we present in this thesis a study on enhancing the efficiency of Bayesian inference methods for both theoretical and practical concerns, by utilizing the explicit or implicit manifold structures of data and models. The study focuses on the two vital fields in Bayesian inference of Markov chain Monte Carlo (MCMC) and particle-based variational inference (ParVI). In this thesis, the main contributions are highlighted in the following.

1. We propose stochastic gradient geodesic MCMC methods, so that we substantially improve the time efficiency of manifold-variable-targeted MCMC methods for processing large scale data sets.

2. We develop Riemannian Stein variational gradient descent methods, which on one hand enhance the iteration efficiency of existing ParVI methods, and on the other hand introduce the first ParVI method for manifold variable inference tasks, with both approximation flexibility and particle efficiency.

3. We make a theoretical analysis on the assumptions that ParVI methods are based on, which also reveals the relation between existing ParVI methods and inspires two new ParVI methods. For all ParVI methods in practice, we propose an acceleration framework and a bandwidth selection method based on the theory, so that the computation efficiency

and particle efficiency of ParVI methods are further improved.

4. We propose a unified theoretical framework for describing general MCMC methods as flows on the Wasserstein space, which systematically explains the behavior of existing MCMC methods, and bridging general MCMC methods with ParVI methods. Based on the theory, we develop two novel ParVI methods that improve the computation efficiency of ParVI methods, and enhance the particle efficiency of MCMC methods.

Key Words: Bayesian inference; Markov chain Monte Carlo; variational inference; differential manifold; information geometry

目 录

主要符号对照表

BLR	贝叶斯逻辑回归模型（Bayesian logistic regression）
BNN	贝叶斯神经网络（Bayesian neural network）
fGH 流	纤维梯度哈密顿流（fiber-gradient Hamiltonian flow）
fRP 流形	纤维黎曼-泊松流形（fiber-Riemannian Poisson manifold）
GFSD	平滑密度的梯度流方法（gradient flow with smoothed density）
GFSF	平滑函数的梯度流方法（gradient flow with smoothed functions）
GMC	测地线蒙特卡罗方法（geodesic Monte Carlo）
gSGNHT	测地线随机梯度诺泽-胡佛恒温器方法（geodesic stochastic gradient Nosé-Hoover thermostats）
HE 方法	热方程带宽选择方法（heat equation bandwidth selection method）
HMC	哈密顿蒙特卡罗（Hamiltonian Monte Carlo）
KL 散度	库尔贝克-莱布勒散度（Kullback-Leibler divergence）
KSD	核化斯坦因差异量（kernelized Stein discrepancy）
LD	郎之万动力学系统（Langevin dynamics）
LDA	隐式狄利克雷分配模型（latent Dirichlet allocation）
MCMC	马尔可夫链蒙特卡罗（Markov chain Monte Carlo）
MH	梅特罗波利斯-海斯廷斯方法（Metropolis-Hastings method）
MMS	最小移动量框架（minimizing movement scheme）
ModVI	基于模型的变分推理（model-based variational inference）
ParVI	基于粒子的变分推理（particle-based variational inference）
PO	粒子优化方法（particle optimization method）
pSGHMC-det	基于粒子使用等价确定性动力学系统模拟的随机梯度哈密顿蒙特卡罗方法（particle-based stochastic gradient Hamiltonian Monte Carlo with equivalent deterministic dynamics）
pSGHMC-fGH	基于粒子使用 fGH 流模拟的随机梯度哈密顿蒙特卡罗方法（particle-based stochastic gradient Hamiltonian Monte Carlo with fGH flow）
RAG	黎曼加速梯度方法（Riemannian accelerated gradient）
RKHS	再生核希尔伯特空间（reproducing kernel Hilbert space）
RKSD	黎曼-核化斯坦因差异量（Riemannian kernelized Stein discrepancy）

RNes	黎曼-涅斯捷洛夫方法（Riemannian Nesterov's method）
RSVGD	黎曼-斯坦因变分梯度下降方法（Riemannian Stein variational gradient descent）
SAM	球面混合模型（spherical admixture model）
SGGMC	随机梯度测地线蒙特卡罗方法（stochastic gradient geodesic Monte Carlo）
SGHMC	随机梯度哈密顿蒙特卡罗方法（stochastic gradient Hamiltonian Monte Carlo）
SGLD	随机梯度郎之万动力学系统（stochastic gradient Langevin dynamics）
SGNHT	随机梯度诺泽-胡佛恒温器方法（stochastic gradient Nosé-Hoover thermostats）
SSI	对称分解积分器（symmetric splitting integrator）
SVGD	斯坦因变分梯度下降方法（Stein variational gradient descent）
VI	变分推理（variational Inference）
WAG	沃瑟斯坦加速梯度方法（Wasserstein accelerated gradient）
WNes	沃瑟斯坦-涅斯捷洛夫方法（Wasserstein Nesterov's method）
A	流形上的外微分式，k-形式（exterior differential form, k-form）
$\mathcal{A}$	一般化斯坦因算符（generalized Stein's operator）
$\mathscr{A}^k(\mathscr{M})$	流形 $\mathscr{M}$ 的 k-形式空间（the space of k-forms on manifold $\mathscr{M}$）
a, b（上下标）	张量的分量指标（indices for tensor components）
a, b	模型参数（model parameters）
$B_t(x)$	标准布朗运动（standard Brownian motion）
$\mathcal{B}$	巴伯生成器（Barbour's generator）
Bern(θ)	伯努利分布（Bernoulli distribution）
C	一般对称正定矩阵，正实数（a general positive-definite matrix, a positive real number）
$\mathscr{C}^\infty(\mathscr{M}), C^\infty(\mathscr{M})$	流形 $\mathscr{M}$ 上的光滑向量值（标量值）函数空间（the space of vector-valued (scalar-valued) functions on manifold $\mathscr{M}$）
$\mathscr{C}_c^\infty(\mathscr{M}), C_c^\infty(\mathscr{M})$	流形 $\mathscr{M}$ 上的具有紧致支撑集的光滑向量值（标量值）函数空间（the space of compactly-supported vector-valued (scalar-valued) functions on manifold $\mathscr{M}$）
c	非负标量值函数，模型参数（non-negative scalar-valued function, model parameter）

const	公式中出现的常数（constant）
D	MCMC 动力学系统的扩散矩阵（diffusion matrix of an MCMC dynamics）
$\mathscr{D}, \widetilde{\mathscr{D}}$	数据集，其随机选取的子数据集（data set, its randomly selected subset）
$\mathrm{Dir}(\alpha)$	狄利克雷分布（Dirichlet distribution）
d	函数的微分，k-形式的外微分（differential of functions, exterior differential of k-forms）
d（上下标）	数据集中数据点的指标（index for data points in a data set）
$d(\cdot,\cdot)$	距离函数（distance）
$d_W(\cdot,\cdot)$	沃瑟斯坦距离（Wasserstein distance）
div	黎曼流形上向量场的散度（divergence of a vector field on a Riemannian manifold）
$\mathbb{E}_q[f]$	关于分布 q 函数 f 的期望（expectation of function f wrt. distribution q）
$\mathrm{Exp}_x(v)$	黎曼流形上的指数映射（exponential map on a Riemannian manifold）
exp	实数上的指数函数（exponential function on real numbers）
expm	矩阵的指数映射（exponential map for matrices）
$F_t(x)$	流形上的流（flows on a manifold）
$\mathcal{F}, \mathcal{H}$	沃瑟斯坦空间上的函数（functions on Wasserstein space）
$\mathcal{F}_f, \mathcal{H}_h$	沃瑟斯坦空间上的线性函数（linear functions on Wasserstein space）
$\mathscr{F}$	纤维丛的公共纤维空间（the common fiber of a fiber bundle）
f, h	一般函数（general functions）
$G, (g_{ij})$	坐标系中的黎曼度量矩阵（Riemannian metric matrix in a coordinate space）
$\mathcal{G}$	目标函数（objective function）
$\mathscr{G}$	$\mathscr{C}_c^\infty(\mathscr{M})$ 经核函数平滑后的空间（the space of kernel-smoothed $\mathscr{C}_c^\infty(\mathscr{M})$）
g	黎曼结构（Riemannian structure）
$\tilde{g}$	纤维黎曼结构（fiber-Riemannian structure）
grad	黎曼流形上函数的梯度（gradient of a function on a Riemannian manifold）
$\mathrm{grad}_{\mathrm{fib}}$	纤维黎曼流形上函数的纤维梯度（fiber-gradient of a function on

	a fiber-Riemannian manifold）
H	MCMC 动力学系统的确定性漂移部分的向量场（the vector field of the deterministic drift part of an MCMC dynamics）
$\mathfrak{H}$	哈密顿量（Hamiltonian）
$\mathscr{H}$	再生核希尔伯特空间（reproducing kernel Hilbert space）
I_m	$m \times m$ 维单位矩阵（$(m \times m)$-dimensional identity matrix）
$\mathcal{I}_k$	k 阶第一类修正贝塞尔函数（modified Bessel function of the first kind in order k）
$\mathscr{I}, \mathscr{J}$	流形上的子集（subsets on a manifold）
i, j, k, l（上下标）	张量的分量指标，有限集中元素的指标（indices for tensor components, indices for elements in a finite set）
id	单位映射（identity map）
J	嵌入映射的雅可比矩阵（Jacobian of an embedding map）
$\mathcal{J}$	目标函数（objective function）
Jac ϕ	映射 ϕ 的雅可比矩阵（Jacobian of map ϕ）
K	核函数（kernel）
$\mathrm{KL}(q\|\|p), \mathrm{KL}_p(q)$	分布 q 关于分布 p 的库尔贝克-莱布勒散度（Kullback-Leibler divergence of distribution q wrt. distribution p）
k	迭代步数，离散参数（iteration count, discrete parameter）
L	有限集大小（size of a finite set）
$\mathcal{L}$	函数空间上的线性映射（linear map between function spaces）
$\mathfrak{L}$	拉格朗日量（Lagrangian）
$\mathscr{L}^2_q(\mathscr{M}), L^2_q(\mathscr{M})$	流形 $\mathscr{M}$ 上的关于分布 q 二次可积的向量值（标量值）函数空间（the space of 2nd-order q-integrable vector-valued (scalar-valued) functions on manifold $\mathscr{M}$）
ℓ, m, n	空间维度（dimensions of spaces）
M	一般矩阵（a general matrix）
$\mathscr{M}$	一般流形（a general manifold）
$\mathcal{M}$	费舍尔信息矩阵（Fisher information matrix）
$\max \cdot \operatorname{argmax}, \min \cdot \operatorname{argmin}$	目标函数最优值与最优变量的数乘（the scalar product of the optimal value and the optimal argument）
N	有限集大小（size of a finite set）
$\mathcal{N}(\lambda, \Sigma)$	高斯分布（Gaussian distribution）
$\mathcal{O}$	大 O 记号（渐进上界记号）（big O notation (asymptotic upper bound)）

o	小 o 记号（渐进非紧上界记号）（small o notation (asymptotic non-tight upper bound)）
$\mathscr{P}(\mathscr{M})$	流形 $\mathscr{M}$ 上的分布空间（the space of all distributions on manifold $\mathscr{M}$）
$\mathscr{P}_2(\mathscr{M})$	流形 $\mathscr{M}$ 上的沃瑟斯坦空间（Wasserstein space on manifold $\mathscr{M}$）
$\mathscr{P}_{\mathscr{H}}(\mathscr{M})$	用来解释 SVGD 的由核函数定义结构的分布流形（the manifold for explaining SVGD with structure defined by a kernel）
P	正交补空间的标准正交基所构成的矩阵（the matrix of orthonormal basis of an orthogonal complement）
p	目标分布，后验分布（a target distribution, posterior distribution）
Q	MCMC 动力学系统的卷曲矩阵（curl matrix of an MCMC dynamics）
q	一般分布，变分分布（a general distribution, variational distribution）
$\hat{q}$	经验分布（empirical distribution）
$\tilde{q}$	平滑经验分布（smoothed empirical distribution）
q_t	连续演化的分布，分布曲线（continuously evolving distribution, distribution curve）
$\mathbb{R}$	实数集（real numbers）
$\mathbb{R}^m$	m 维欧氏空间（m-dimensional Euclidean space）
r, s	余切向量，动量，辅助变量（cotangent vectors, momentums, auxiliary variables）
$\mathbb{S}^m$	m 维超球面（m-dimensional hypersphere）
T	乘积流形重数，SAM 模型的话题数（order of product manifold, number of topics of SAM）
$T\mathscr{M}, T^*\mathscr{M}$	流形 $\mathscr{M}$ 的（余）切丛（(co)tangent bundle of manifold $\mathscr{M}$）
$T_x\mathscr{M}, T_x^*\mathscr{M}$	流形 $\mathscr{M}$ 在点 x 处的（余）切空间（(co)tangent space of manifold $\mathscr{M}$ at x）
$\mathcal{T}_q^\rho$	从分布 q 到分布 ρ 的最优传输映射（optimal transport map from distribution q to distribution ρ）
$\mathcal{T}_{\#}q$	分布 q 在可测映射 $\mathcal{T}$ 下的前推（push-forward of distribution q under measurable map $\mathcal{T}$）
$\mathscr{T}(\mathscr{M})$	流形 $\mathscr{M}$ 上的向量场空间（space of vector fields on manifold $\mathscr{M}$）
t	时间，连续标量参数（time, continuous parameter）
U, V	向量场（vector fields）

$\mathcal{U}$	沃瑟斯坦空间上的纤维梯度哈密顿流（fGH 流）（fiber-gradient Hamiltonian flow on Wasserstein space）
u, v	切向量，速度（tangent vectors, velocities）
V_f	一般流形上以 f 为哈密顿量的哈密顿向量场（Hamiltonian vector field with Hamiltonian f on general manifolds）
$\mathcal{V}_{\mathcal{F}}$	沃瑟斯坦空间上以 $\mathcal{F}$ 为哈密顿量的哈密顿向量场（Hamiltonian vector field with Hamiltonian $\mathcal{F}$ on Wasserstein spaces）
vMF(λ, κ)	冯 · 米塞斯-费舍尔分布（von Mises-Fisher distribution）
W	MCMC 动力学系统的等价确定性向量场（equivalent deterministic vector field of an MCMC dynamics）
w	核函数的带宽（kernel bandwidth）
X	数据变量，数据集（data variable, data set）
$\mathfrak{X}$	向量场空间的线性子空间（linear subspace of the space of vector fields）
x, y, z	空间上一般的点或其坐标（general points on a space, or their coordinates）
$\tilde{x}, \tilde{y}$	点的坐标（coordinates of points）
$\tilde{x}^i, x^i$	点 x 坐标的第 i 个分量（the i-th component of the coordinate of x）
$x^{(i)}$	有限集中第 i 个元素（the i-th element in a finite set）
Y	数据响应变量（如类别等）（response variable (e.g. labels)）
Z	隐变量，目标变量（latent variable, target variable）
$\mathscr{Z}$	隐空间（latent space）
α	模型参数（model parameter）
β	模型参数, SAM 模型的话题变量（model parameter, the topic variable of SAM）
$\Gamma(\cdot)$	伽马函数（Gamma function）
Γ_q^ρ	从 q 到 ρ 的平行移动（parallel transport from q to ρ）
γ_t	一般流形上的曲线（curve on a manifold）
Δ	贝尔特拉米-拉普拉斯算符（Beltrami-Laplace operator）
$\delta_x(\cdot)$	集中在点 x 处的狄拉克分布（测度）（Dirac distribution (measure) concentrated on x）
δ_{ij}, δ_j^i	克罗内克-德尔塔张量（Kronecker delta tensor）
ε	离散步长参数（discrete step size）
ζ	推导中所使用的向量场（vector field used in deductions）

η	一般的（协变）张量（a general (covariant) tensor）
θ	模型参数，SAM 模型的话题配比变量（model parameter, topic proportion variable of SAM）
ϑ	角度，一般向量（an angle, a general vector）
ι	线性空间之间的等距同构映射（isometric isomorphism between linear spaces）
κ, λ	模型参数（model parameter）
Λ	有限维线性空间中的正交投影（orthogonal projection in a finite-dimensional linear space）
μ, ν	流形切空间上的 k 次外形式（exterior form of degree k on a tangent space of a manifold）
Ξ	嵌入映射（embedding map）
ξ	MCMC 方法的恒温器变量（thermostats variable of an MCMC method）
Π	坐标空间中的子集（subsets in a coordinate space）
π_q	$\mathscr{L}_q^2(\mathscr{M})$ 向沃瑟斯坦切空间 $T_q\mathscr{P}_2(\mathscr{M})$ 的正交投影（orthogonal projection of $\mathscr{L}_q^2(\mathscr{M})$ onto the tangent space $T_q\mathscr{P}_2(\mathscr{M})$ of Wasserstein space）
ϖ	纤维丛向其基空间的投影（projection of a fiber bundle onto its base space）
ρ	一般分布，辅助分布（a general distribution, auxiliary distribution）
ϱ	二重乘积流形上的联合分布（joint distribution on a 2-fold product manifold）
Σ	协方差矩阵（covariance matrix）
σ	排列，sigmoid 函数，模型参数（permutation, sigmoid function, model parameter）
ς	模型参数（model parameter）
τ	SAM 模型的话题变量下标（index for the topic variable of SAM）
Φ, Ψ	坐标映射（coordinate maps）
ϕ	变换，向量值函数（transformation, vector-valued function）
φ	标量值函数（scalar-valued function）
χ	流形上的二重向量场，泊松结构（bivector field on a manifold, Poisson structure）
Ω	坐标空间中的子集（subsets in a coordinate space）
ω	流形上的体积形式（volume form on a manifold）

$\top$（上标）	矩阵转置（transpose of a matrix）
$a \cdot b$	两实数乘积或两欧氏向量内积（the multiplication of two real numbers, or the common inner product of two vectors in an Euclidean space）
$\times$	两线性空间或两流形的直积（direct product of two linear spaces or two manifolds）
$\circ$	两映射的复合（composition of two maps）
$\lvert\cdot\rvert$	矩阵行列式，有限集大小（determinant of a matrix, size of a finite set）
$\lVert\cdot\rVert$	赋范空间中的范数（默认为欧氏空间中的 2-范数）（norm of a normed linear space (default is the 2-norm in an Euclidean space)）
∇	欧氏空间中的梯度（gradient in Euclidean spaces）
∂_i	函数的偏导数，切空间的基向量（partial derivative of a function, basis vector of a tangent space）
$*$	函数的卷积（convolution of functions）
$\langle\cdot,\cdot\rangle$	内积（inner product）
$\{\cdot,\cdot\}$	泊松括号，泊松结构（Poisson bracket, Poisson structure）
$\otimes$	两张量的张量积，线性空间或流形的多重直积（tensor product of two tensors, multi-folded direct product of a linear space or a manifold）
$\wedge$	外形式的外向积，楔积（exterior product, wedge product）

第 1 章　引言

本文所关心的任务是增强贝叶斯推理方法的高效性，以及利用数据和模型的流形结构解决此任务的理论和方法。本章首先介绍贝叶斯模型和贝叶斯推理的概念及其在当下应用环境中的重要意义，并强调贝叶斯推理当下所面临的高效性需求，然后介绍利用流形结构增强贝叶斯推理方法高效性的独特价值和广阔前景，随后对研究现状进行综述并提出仍需解决的具体问题，最后对本文工作内容及其所解决的高效性需求进行概述。

1.1　研究背景

近年来，由信息科技产业方面带来的大规模数据以及深度学习等技术的成功[1-6] 引发了人工智能及机器学习领域在研究及应用上的迅猛发展。虽然深度学习在处理这些大规模数据的实际任务中取得了令人瞩目的成功，但是当下实际任务中不断浮现出的新需求也为深度学习和机器学习领域带来了新的挑战，例如抵抗对抗样例（adversarial examples）[7] 的干扰，进行鲁棒决策[8-9]，实现小样本学习（few-shot learning）[10] 和元学习（meta-learning）[11-12]，提高模型可解释性（意味着可靠性和可控性）[13] 等。针对这些，历史悠久的贝叶斯模型（Bayesian models）因其强大的建模能力，特别是将知识与数据结合的能力及其表现出的鲁棒性和避免过拟合等优势，而在新时代依然受到人们的高度关注。结合深度学习的贝叶斯模型可比传统深度模型有更好的表现[14]，并且已经成功地在防御对抗样例攻击[15]、小样本学习[16] 和元学习[17] 等方面取得喜人进展。

贝叶斯模型　不同于一般模型的是，贝叶斯模型是以概率性的方式来对数据的隐含表示进行建模的，并以一个称为隐变量（latent variable）的随机变量 Z 表示。因此在学习贝叶斯模型的过程中，便需要以概率性的方式来表示数据的隐含特征，即需要进行贝叶斯推理。具体来说，一个贝叶斯模型可抽象为使用隐变量 Z 对数据 X 进行建模的过程（可参见图 1.1）。贝叶斯模型依据经验观念或领域知识（domain knowledge）为隐变量 Z 指定一个先验分布（prior distribution）$p(Z)$，并指定一个基于给定隐变量产生数据 X 的过程，即似然分布（likelihood distribution）$p(X|Z)$。贝叶斯模型因此建立了隐变量与数据的联合分布 $p(X,Z)=p(Z)p(X|Z)$，其中体现了数据与隐变量之间的关系，而对数据的建模则可由 $p(X)=\int p(X,Z)\,\mathrm{d}Z$ 给出。

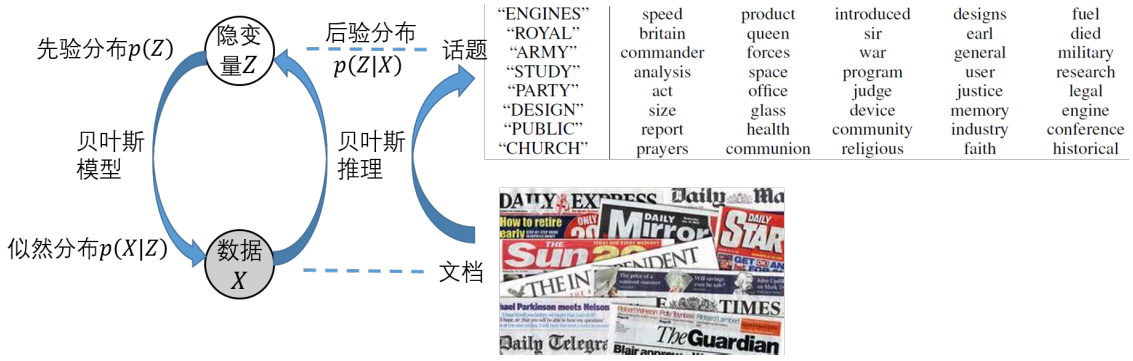


图 1.1　贝叶斯模型结构及贝叶斯推理的概念和意义（部分取自 [18]）。

贝叶斯推理　在拿到一组数据之后，人们往往会关心一个贝叶斯模型可以从数据中提取出来的知识，并希望得到一个可以更好匹配这组数据的贝叶斯模型。贝叶斯推理（Bayesian inference）正是这些需求的核心任务。具体来说，贝叶斯推理这个任务是要在给定一组数据 X 之后，求得隐变量的后验分布（posterior distribution）$p(Z|X)$，因而它有时也被称为后验推理。一方面，后验分布体现了观测数据通过两者之间的相关性对隐变量带来的更新（相较于先验分布），因此它通过隐变量表达了数据所带来的知识。对于很多贝叶斯模型，隐变量往往就是人们所关心的量，例如可反映文档数据特征的话题这个量[19-22]（参见图 1.1），或是手写字符的笔画构成[16]，因而由后验分布所体现的隐变量信息便可满足人们对所关心量的需求。在强化学习（reinforcement learning）领域中也有直接将动作（action）作为隐变量并使用其后验分布进行决策的工作[23]。对于另一些贝叶斯模型，隐变量虽然不会直接向人们汇报数据的特征，但可通过在新数据点上进行预测[14-15]、在新条件下给出决策[24-26] 或者生成新数据样例[3,27-28] 等方式为人们服务。在这些任务中使用隐变量的后验分布即可体现训练数据为这些任务所提供的信息。另一方面，在寻找一个与数据更加匹配的贝叶斯模型时（例如寻找参数化似然分布 $p_\theta(X|Z)$ 的最优参数 θ 时）也需要进行贝叶斯推理。例如通过著名的期望最大化算法（expectation-maximization algorithm, EM algorithm）[29] 寻找最优模型参数时，在其 E 步即需要求解关于后验分布的期望[30-31]。

贝叶斯推理意义重大，但同时也是一个很难的任务。由贝叶斯公式，可以得知后验分布可以表示为 $p(Z|X) = \dfrac{p(X,Z)}{p(X)} = \dfrac{p(Z)p(X|Z)}{p(X)}$，但对于一个一般的（即非共轭的）贝叶斯模型，由于 $p(X) = \displaystyle\int p(X,Z)\,\mathrm{d}Z = \int p(Z)p(X|Z)\,\mathrm{d}Z$ 通常无法闭式（closed form）求得，因而其后验分布往往无法通过闭形式来表达，从而难以在实际中使用。这种情况也被称为后验分布是不可行的（intractable）。为解决这个问题，各种贝叶斯推理方法便从各自不同角度来估计后验分布。变分推理方法

（variational inference, VI）是通过从一个可行的（tractable）分布族中选出与后验分布最接近的分布作为对它的近似，其中“可行分布”是指此分布有密度函数的闭形式，或者可以方便直接地从中采取样本。而蒙特卡罗方法（Monte Carlo）则希望直接从后验分布中采样。由于一般后验分布的复杂性以及准确的后验密度函数不可知等因素，蒙特卡罗方法通常会模拟一条以后验分布为平稳分布的马尔可夫链（Markov chain）来近似采样，即马尔可夫链蒙特卡罗方法（Markov chain Monte Carlo, MCMC）。这两类方法近年来取得了诸多令人瞩目的进展，并仍在活跃发展中。

当下的大数据环境以及机器学习在更多更细粒度领域的广泛应用为贝叶斯推理带来了新的高效性需求（可参见图 1.2）。首先，数据的多样性带来了处理流形数据以及结构化数据的需求，而贝叶斯模型为更好地为这些数据建模，希望将其隐变量取在一个特定的流形上。例如针对超球面（hypersphere）数据（每个数据点都是一个具有单位长度的高维向量）的球面混合模型（spherical admixture model）[22] 和超球面变分自编码器（hyperspherical variational auto-encoder）[32] 及类似模型[33] 也将隐变量取在超球面上，针对具有树状结构数据的庞加莱变分自编码器（Poincaré variational auto-encoder）[34] 及类似模型[35-37] 将隐变量取在双曲空间（hyperbolic space）[38] 中，而针对矩阵补全（matrix completion）任务的贝叶斯矩阵补全方法[39-41] 则将隐变量取在斯蒂菲尔流形（Stiefel manifold）[42-43] 上。对于这些模型进行贝叶斯推理便要求推理方法能够高效处理具有流形结构的隐变量。其次，当下的贝叶斯模型为提高建模能力，先验和似然的选取已不再限于成共轭关系的分布，而会十分灵活多样，使得后验分布十分复杂，特别是与深度学习方法结合的一些模型[3,14]。更复杂的后验分布会使推理方法在取得合理结果时需要使用更多的资源，例如需要更多的时间和算力才可收敛到复杂后验上，或者需要更多的存储空间保存足够多的样本才能合理地近似复杂后验，甚至在使用样本用于后续任务时也需要更多的时间和算力。因此推理方法需要进一步增强资源高效性以满足复杂模型的实际需求。再次，为了实现复杂建模的目的，这些复杂模型也要求贝叶斯推理方法能够很好地估计出后验分布的复杂情况，即要求推理方法具有更强的近似灵活性（approximation flexibility）。最后，贝叶斯推理领域近来已涌现出非常多的新方法，但是针对这些方法所基于的假设以及它们之间的联系等方面的分析尚属初步。只有对这些方法有了充分理解并建立联系，才能清楚哪些方法有重复，而哪些方面尚未探索，进而指导新的高效推理方法的开发。

利用流形结构的意义　针对贝叶斯推理所面对的这些挑战与高效性需求，流形（manifold）这一数学概念及其相关理论可以提供有效的解决方案（可参见图 1.2）。

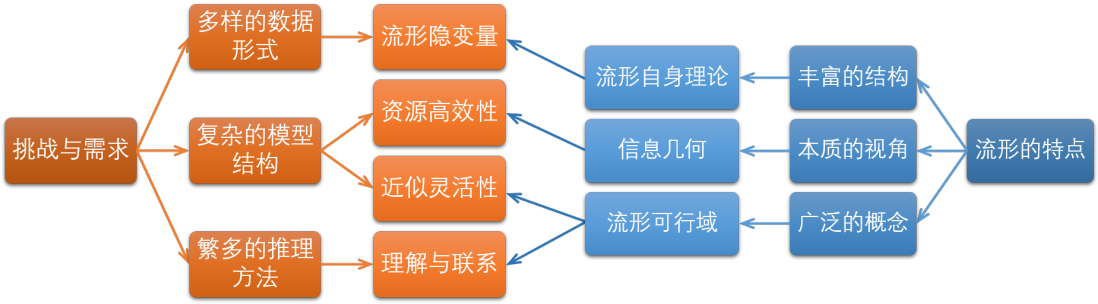


图 1.2　贝叶斯推理面临的挑战与需求，并在其中利用流形结构的意义。

大致来说，流形是局部表现为某一欧氏空间的几何空间。一方面，这一概念可以允许空间的扭曲与弯折，因而它比线性空间要广泛得多，而另一方面，流形上也可以定义各种丰富的结构并可显式写为表达式，因而在流形上也可以进行各种分析和计算。在贝叶斯推理中考虑流形结构，首先一种情况自然是需要处理流形隐变量的贝叶斯推理方法。流形结构一方面为这些方法的开发给出了限定性的原则，而另一方面流形自身的理论则可帮助这些推理方法更加高效地处理隐变量的流形结构。例如流形嵌入空间中的结构可帮助 MCMC 方法提高对应马尔可夫链的模拟效率[44]。其次，流形及其结构可以在其坐标空间中“参数化”，也就是使用欧氏空间中的坐标来表示，但无论是哪个坐标空间中的表达式，它所反映的都是同一个流形的本质几何性质，而不会随坐标空间的不同而不同。与之形成对比的是，若将欧氏隐变量 $Z \in \mathbb{R}^m$ 视作似然分布 $p(X|Z)$ 的参数，则似然分布的不同参数化形式将会产生不同的隐变量，进而会产生不同的推理过程。这些过程中有些可能会收敛得快，而另一些则会收敛得慢。人们希望可以建立一种与参数化形式无关的推理过程，进而可以从更加本质的角度提高推理方法的效率，而流形的观念正好可以解决这个问题。为此，可以考虑所有似然分布 $\{\, p(X|Z) \mid Z \in \mathbb{R}^m \,\}$ 所构成的流形，将隐空间 $\mathbb{R}^m$ 视为此流形的一个坐标系，并用流形的结构来确定在 $\mathbb{R}^m$ 中的推理过程。这就是信息几何（information geometry）方法[45-47] 的思想。考虑信息几何的方法已在实践中展现出了更高的迭代效率，例如著名的自然梯度（natural gradient）方法[47-51]。再次，为提高变分推理方法的近似灵活性，它所使用的可行分布族应选为尽可能大的分布集合。而当可行分布族大到无法使用一个线性空间来表示时，便只能以流形这个更广泛的概念来描述，即流形可行域的情况。沃瑟斯坦空间（Wasserstein space）[52-54] 便是一个例子。它大到可以包含隐空间上所有方差有限的分布，但其丰富的结构仍然允许对应的变分推理方法得出可以实现的算法[55-57]，并取得比使用参数分布族更高的近似灵活性。最后，由于流形的概念十分广泛，因而很多推理方法都有望在流形可行域的视角下统一起来，例如郎之

万动力学系统（Langevin dynamics）[58-60] 这个 MCMC 方法与一些基于粒子的变分推理方法[55-56] 都可被视作是对沃瑟斯坦空间上的梯度流这个过程的模拟[56,61]。从这个视角出发，人们便可研究这些推理方法的特性及其之间的联系，并启发更多方法的设计。

基于这些论点，本文说明了流形结构可为贝叶斯推理所面临的高效性需求提供分析问题的本质视角和解决问题的有力工具。而另一方面，贝叶斯推理高效性的很多问题仍然没有得到解决，而利用流形结构这个观念仍然有广阔空间可以继续挖掘。本文将深入研究通过流形结构来解决贝叶斯推理在当下环境所面临的高效性需求。

1.2　研究现状

本节首先介绍一般贝叶斯推理方法的划分与发展，然后介绍利用流形结构的贝叶斯推理方法，最后总结贝叶斯推理领域中仍然有待研究的问题。

1.2.1　一般贝叶斯推理方法

如上一节所提，贝叶斯推理方法可以划分为变分推理方法（variational inference, VI）和马尔可夫链蒙特卡罗方法（Markov chain Monte Carlo, MCMC）。VI 方法会从一个可行分布族中找出与后验分布尽可能接近的分布作为对后验分布的估计。在 VI 领域中，这个可行分布族中的分布又被称为变分分布（variational distribution）。MCMC 方法则会通过模拟一个合适的马尔可夫链从后验分布中采样。

基于模型的变分推理方法（ModVI）　经典的变分推理方法所使用的可行分布族通常是一个参数分布族，或者称为统计模型，因而这类变分推理方法可称为基于模型的变分推理方法（model-based variational inference, ModVI）。ModVI 领域中有一大类方法为隐变量 $Z \in \mathbb{R}^m$ 选取的变分分布是各维之间或者各子部分之间相互独立的，使得变分分布可以写为一些因子乘积的形式，例如 $q(Z) = \prod_{i=1}^{m} q(Z_i)$。这称为平均场假设（mean-field assumption）。这些因子进一步被指数分布族（exponential family）参数化，从而将推理问题变为参数优化问题。这类方法可见于各经典贝叶斯模型的推理方法中[19-22]，并可参见综述 [30,62]。这些方法依赖于被处理的模型的结构，而近来一些方法[63-64] 则通过将变分分布选为混合高斯分布（mixture of Gaussian）使之可用于任一模型。另一类称为期望传播（expectation propagation, EP）[65] 的推理方法则借鉴独立同分布数据 $\mathscr{D} := \{X_d\}_{d=1}^{|\mathscr{D}|}$ 所对应的后验分布的形式 $p(Z|\mathscr{D}) \propto$

$p(Z)\prod_{d=1}^{|\mathscr{D}|}p(X_d|Z)$，选择具有同样因子形式的变分分布 $q(Z)\propto q_0(Z)\prod_{d=1}^{|\mathscr{D}|}q_d(Z)$，并逐个更新其中的每个因子。近年来，EP 方法也得到了一些改进，以解决贝叶斯神经网络的推理任务[66-67]。在深度学习得到人们的关注之后，ModVI 领域便引入了由神经网络所描述的变分分布，例如变分自编码器（variational auto-encoder）[3] 及其变种[68]。由于神经网络的强大拟合能力[69-71]，使用神经网络的 ModVI 方法可比传统方法有更强的近似灵活性。这些方法都是通过神经网络直接描述变分分布的密度函数，而另有一些方法以使用神经网络生成样本的方式来表示变分分布，例如标准化流（normalizing flows）方法[72] 及其变种[73]。这些方法为实现计算，仍然可以得到变分分布的密度函数，而近来出现的一些方法[74-75] 通过密度比值估计的手段抛弃了这个限制，从而得到更强的近似灵活性。

ModVI 方法使用参数化形式来表示变分分布，并希望最小化与后验分布之间的差别（通常以 KL 散度（KL divergence）衡量），因此它们最终可表示为一个参数优化问题。最常用而有效的优化方法是梯度下降方法（gradient descent），而在此基础之上也有仍可保持一阶（即只使用梯度信息）但可收敛更快的加速方法，例如波利亚克动量方法（Polyak's momentum）[76] 和涅斯捷洛夫加速方法（Nesterov's acceleration method）[77]。随着深度学习的兴起，优化领域近年来也得到了快速发展，出现了很多新的优化方法[78-80]。

基于粒子的变分推理方法（ParVI）　ModVI 方法使用参数化形式来表示变分分布，而这终究会限制其近似能力。近年来出现的一类基于粒子的变分推理方法（particle-based variational inference, ParVI）则使用变分分布的一组样本，或者称为粒子，来表示此变分分布。它遵从变分推理方法的原理，即通过最小化与后验分布的差别来更新变分分布，因而可以得到一个确定性的粒子更新规则。斯坦因变分梯度下降方法（Stein variational gradient descent, SVGD）[81] 是这类方法的典型代表。它是通过在特定的函数空间中最大化 KL 散度的减小率来得到粒子更新规则的。其后出现了 SVGD 的各变种[55,82-84] 以及基于沃瑟斯坦空间上梯度流而开发的 w-SGLD 和 Blob 方法[56]。与 MCMC 方法类似，ParVI 方法使用的是粒子这样一种非参数化形式，而只要使用更多的粒子它们便可取得更强的近似灵活性。另外，近来也有工作表明它们具有渐进准确性[85]，即当所使用的粒子数趋于无穷时其近似结果是准确的。这些是 ParVI 方法相较 ModVI 方法的优势。而它们所遵从的优化原理和确定性更新规则使得 ParVI 的迭代效率比 MCMC 方法更高，并且由于它们直接考虑和利用了粒子间相互作用，使得它们相较 MCMC 方法具有更强的粒子高效性，亦即达到相同的近似效果只需要使用更少的粒子。这意味着推理结果可节省更多的

表 1.1　三类贝叶斯推理方法的比较。

推理方法	ModVI	ParVI	MCMC
渐进准确性	无	有	有
近似灵活性	有限	无限	无限
迭代有效性	强	强	弱
粒子高效性	（不适用）	强	弱

存储空间，并且在使用推理结果（通常是计算各粒子所给出的目标量的均值）时可节省更多的算力和时间。

马尔可夫链蒙特卡罗方法（MCMC）　较早期的 MCMC 方法，例如梅特罗波利斯（Metropolis）方法[86] 和后来海斯廷斯（Hastings）对其改进[87]，重要性采样（importance sampling）[88-89] 以及粒子滤波方法（particle filtering）[90]，都是基于取舍提议样本（proposal）的方法。由于提议样本与当前样本通常会很靠近，而且被拒绝的情况时有发生，因而这些方法的采样过程收敛很慢，并且样本之间具有很高的自相关性（auto-correlation），粒子效率（或有效样本比例）很低。另外一类称为吉布斯采样（Gibbs sampling）的方法[91] 通过不断地从各变量或变量的各维分量的条件分布中交替采样以完成对整个目标分布的采样。由于这些条件分布相对容易进行采样，因此它可适用于各种场景，例如话题模型的推理任务[92-95]。但它所产生的样本仍然十分靠近，所以它也面临着收敛慢和粒子效率低的问题。近年来得到迅速发展的基于动力学系统（dynamics-based）的 MCMC 方法则可以利用信息量更加丰富的后验分布对数梯度 $\nabla_Z \log p(Z|\mathscr{D})$ 来构建一个合适的连续时间动力学系统，从而在离散模拟中更快收敛，并产生自相关性更低的样本从而节省存储空间并减轻后续任务的负担。郎之万动力学系统（Langevin dynamics, LD）方法[58-59] 是最早的实例，而之后逐渐为人熟知的哈密顿蒙特卡罗（Hamiltonian Monte Carlo, HMC）方法[96-98] 则更加高效，因为它引入了动量这个辅助变量，可使提议样本在保持高接受率的同时离开当前样本更远，这样便可以更快速地探索样本空间，从而加速收敛并提高粒子效率。之后各种方法[99-104] 则进一步分析和改进了 HMC 方法。由于 MCMC 方法所模拟的马尔可夫链可保证以后验分布为平稳分布，因此它们都具有渐进准确性的保证。一些特定的 MCMC 方法也有可证的收敛阶分析，例如 LD 方法[58,105-107]，HMC 方法[104,108-109] 以及 HMC 的变种[110]。本文将主要考虑基于动力学系统的 MCMC 方法，因此若无特别说明，MCMC 方法均指这一类 MCMC 方法。这三类贝叶斯推理方法的特点对比可参见表 1.1。

基于动力学系统的 MCMC 方法中有一类值得专门加以介绍，即具有可扩展性

的随机梯度 *MCMC* 方法。注意到独立同分布数据 $\mathscr{D}=\{X_d\}_{d=1}^{|\mathscr{D}|}$ 对应的后验分布的对数梯度为：

$$\nabla_Z \log p(Z|\mathscr{D}) = \nabla_Z \log p(Z) + \sum_{d=1}^{|\mathscr{D}|} \nabla_Z \log p(X_d|Z),$$

因此对它的一次准确计算需要遍历一次整个数据集 $\mathscr{D}$，而当数据集非常大时，此计算代价是巨大的。为使 MCMC 方法具有可扩展性（scalability），即能够以亚线性于数据规模 $|\mathscr{D}|$ 的时间复杂度高效处理大规模数据，一个可行的方法是在每次需要计算梯度时，转而在原数据集 $\mathscr{D}$ 的一个随机选出的具有固定大小的子数据集 $\widetilde{\mathscr{D}}$ 上进行对梯度的估计：

$$\tilde{\nabla}_Z \log p(Z|\mathscr{D}) := \nabla_Z \log p(Z) + \frac{|\mathscr{D}|}{|\widetilde{\mathscr{D}}|} \sum_{d\in\widetilde{\mathscr{D}}} \nabla_Z \log p(X_d|Z). \tag{1-1}$$

由于子数据集 $\widetilde{\mathscr{D}}$ 具有随机性，因而此估计也是一个随机变量。这个估计便被称为随机梯度（stochastic gradient）。一方面，随机梯度的计算代价可显著小于准确梯度的代价，且随机子数据集大小 $|\widetilde{\mathscr{D}}|$ 的选择对原数据集大小 $|\mathscr{D}|$ 是不敏感的。因而可以使用随机梯度的方法（只要保证可正确收敛）都可以实现可扩展性。另一方面，随机梯度具有清楚的理论描述，这极大地方便了对使用随机梯度的方法的分析，使这些方法的正确性和高效性有理论保证。具体地，由数据点 $\{X_d\}_{d=1}^{|\mathscr{D}|}$ 的独立同分布性质，当子数据集 $\widetilde{\mathscr{D}}$ 是均匀随机挑选而来时，可知随机梯度是准确梯度的无偏估计。进一步由中心极限定理，可将随机梯度视作服从以准确梯度为期望的高斯分布，因而可形式化地写出关系 $\tilde{\nabla}_Z \log p(Z|\mathscr{D}) = \nabla_Z \log p(Z|\mathscr{D}) + \mathcal{N}(0,\Sigma(Z))$，其中 $\mathcal{N}(0,\Sigma(Z))$ 表示以 0 为期望以 $\Sigma(Z)$ 为协方差矩阵的高斯分布。这两方面的好处使得随机梯度已在机器学习领域中得到了广泛应用。它最早在基于优化的学习任务中受到关注，其中最有名的优化方法是随机梯度下降方法（stochastic gradient descent, SGD）[111]。随后，在贝叶斯推理领域，同样基于优化的 ModVI 方法中也逐渐有了可使用随机梯度的方法[30,67]。

而在 MCMC 领域中，随机梯度所带来的噪声却有可能从根本上破坏原动力学系统的平稳分布。随机梯度 MCMC 方法最早由 Welling 等人[60] 提出。他们在 LD 的模拟过程中使用随机梯度，并将对应方法称为随机梯度郎之万动力学系统方法（stochastic gradient Langevin dynamics, SGLD）。此方法之后被发现是可以在使用小步长时直接使用随机梯度进行模拟的[112-114]。但是对于 HMC 方法，使用随机梯度进行模拟则会带来根本性的问题[115-116]，即无论后验分布如何，对应的模拟过程的平稳分布会变得无限接近于均匀分布。为解决此问题，必须构造新的动力学

系统。随机梯度哈密顿蒙特卡罗方法（stochastic gradient Hamiltonian Monte Carlo, SGHMC）[115] 为原来的哈密顿动力学系统加入了随机扩散项以及用来平衡随机梯度噪声影响的摩擦力项，使所得动力学系统在使用随机梯度时仍然可以正确采样。随后，SGHMC 方法被不断地加以改进[103,117-118]，其中值得一提的是随机梯度诺泽-胡佛恒温器方法（stochastic gradient Nosé-Hoover thermostats, SGNHT）[119]。它为 SGHMC 动力学系统引入了恒温器变量（thermostats），使得摩擦力项可以更好地匹配随机梯度噪声以抵消其扰动，从而提高采样效率。最后，Ma 等人[120] 为这些 MCMC 方法的动力学系统给出了一个统一的完备表示形式，特别是这一形式中的动力学系统都可以后验分布为平稳分布。至于 ParVI 方法，由于它们也是基于优化的方法，并且与 LD 具有紧密的联系[55-56,85]，因而通常都可以直接使用随机梯度进行模拟。

1.2.2　利用流形结构的贝叶斯推理方法

如上一节所述，在贝叶斯推理方法中结合流形的观念可以应对当前环境所带来的高效性需求。本节将按照利用流形结构的方式分别介绍现有研究工作。

流形自身理论　ModVI 方面，各模型通常使用流形上的可行分布来构造基于平均场近似的方法[22,33]，或者通过流形的坐标表示来使用基于神经网络的方法[32,34]。对于前者，推理问题往往会被转化为流形上的优化问题。在此方面，基于（随机）梯度下降的方法[121-122] 及其一阶加速方法[123-124] 均已被开发。MCMC 方面，Brubaker 等人[125] 考虑了以欧氏空间中的限制形式而定义的流形上的采样，而 Byrne 等人[44] 则考虑在流形的嵌入空间中进行采样。这两个方法都可以处理没有全局坐标系的流形，例如超球面。此外也有一些针对特定流形而开发的 MCMC 方法[41,126]。上述 MCMC 方法均基于哈密顿动力学系统。

信息几何　信息几何的思想和技术最先在 ModVI 领域得到应用，因为求解对应优化问题可以方便地使用自然梯度[47]。近年来，自然梯度也在基于随机子数据集的 ModVI 方法[30]、针对深度模型的快速实现[48-50] 以及利用沃瑟斯坦空间几何结构[51] 等方面取得发展。考虑信息几何的 MCMC 方法则都是基于在似然分布流形上进行采样的思想而开发的，最典型的代表是黎曼流形郎之万动力学系统方法（Riemann manifold Langevin dynamics）和黎曼流形哈密顿蒙特卡罗方法（Riemann manifold Hamiltonian Monte Carlo）[127]。这两个方法的改进方法[128] 和随机梯度版本[18,120] 也已被考虑。需要注意的是这些 MCMC 方法也可用于隐变量处在一个具有全局坐标系的流形上的情况。

流形可行域　由于 ParVI 方法使用粒子这样十分一般化的方式来代表变分分布，因此对 ParVI 方法所使用的可行域（变分分布族）的分析需要考虑十分广泛的分布空间，而这通常是一个流形。Liu[85] 首先从此角度出发分析了 SVGD 方法，发现它可看作是对 $\mathbb{R}^m$ 上的一个由核函数（kernel）定义的分布流形 $\mathscr{P}_{\mathscr{H}}(\mathbb{R}^m)$ 上 KL 散度的梯度流进行模拟的过程，即在此分布流形上最小化 KL 散度的过程。随后出现的 w-SGLD 方法和 Blob 方法则是从模拟 $\mathbb{R}^m$ 上的沃瑟斯坦空间 $\mathscr{P}_2(\mathbb{R}^m)$ 上 KL 散度的梯度流的角度而开发的。这些理解为 ParVI 方法的原理分析与技术改进打下了基础。而另一方面，MCMC 领域中的 LD 方法也被识别为是对沃瑟斯坦空间 $\mathscr{P}_2(\mathbb{R}^m)$ 上梯度流的模拟[61]，从而与 ParVI 方法建立了联系。

1.2.3　有待研究的问题

综上所述，利用流形结构的贝叶斯推理方法已经取得了令人瞩目的进展，但贝叶斯推理中仍然有很多高效性的需求还未得到满足，并且利用流形结构的实践尚处初步，仍然有很多方向值得探索从而满足这些高效性需求。

第一，目前仍然没有可以高效处理大规模数据的流形隐变量 MCMC 方法。现有的针对流形隐变量的 MCMC 方法都是基于哈密顿动力学系统的，而它却与随机梯度不相容，因此这些方法都不具有可扩展性。虽然针对信息几何而开发的 MCMC 方法可以用于隐变量所在流形具有全局坐标系的情况，但现实中仍然有一大类模型无法满足此要求，例如隐空间为超球面的模型[22,32-33]。这些模型在处理大规模数据集时，尚无高效的 MCMC 方法，而现有的 ModVI 方法则由于其有限的近似灵活性，往往无法取得很好的结果。这些因素掩盖了这类模型的优势。

第二，流形隐变量方面仍然没有近似灵活性强且粒子高效的推理方法，并且 ParVI 方法也会受模型参数化形式的影响而无法快速收敛。针对流形隐变量，虽然有一些近似灵活性强的 MCMC 方法，但它们的粒子高效性仍然有限，而具有粒子高效性的 ParVI 方法则尚不可处理流形变量。ParVI 方法目前也无法利用信息几何方法达到与参数化形式无关的收敛速度，影响了它们在实际应用中的表现。

第三，ParVI 方法所做的假设及其之间的关系尚不清晰，且它们仍然没有充分利用梯度信息。虽然现有工作已经将一些 ParVI 方法识别为对分布流形 $\mathscr{P}_{\mathscr{H}}(\mathbb{R}^m)$[85] 或沃瑟斯坦空间 $\mathscr{P}_2(\mathbb{R}^m)$[56] 上 KL 散度的梯度流的模拟，但仍然有很多问题尚不清晰，例如这两个流形上的梯度流的关系，各 ParVI 方法使用有限多个粒子对梯度流模拟时是否对变分分布做了假设，这些可能的假设之间的关系以及是否可能摆脱这些假设等问题。这些问题关系到人们对 ParVI 方法的认识，包括与其他方法对比的优势与劣势，适用场景，影响 ParVI 方法性能的关键因素，以及

表 1.2　现有贝叶斯推理方法总结及本文主要贡献

推理方法		ModVI	ParVI		MCMC	
			常规方法	加速方法	常规方法	随机梯度方法
一般方法		[3,19-21,30,62-68,72-80,111]	SVGD[81,85]，w-SGLD/Blob[56]，[55,82-84]，**第 5 章**：GFSD/GFSF，**第 6 章**：pSGHMC	**第 5 章**	LD ([58-59,105-107])，[86-91,96-104,108-109]	SGHMC ([110,115])，[60,103,112-114,116-120]
利用流形结构的方法	流形自身理论	[22,32-34,121-124]	**第 4 章**：RSVGD		[41,44,125-126]	**第 3 章**
	信息几何	[30,47-51]	**第 4 章**：RSVGD		[127-128]	[18,120]
	流形优化域		• [85]：SVGD 模拟 $\mathscr{P}_{\mathscr{H}}(\mathbb{R}^m)$ 上的梯度流； • [56]：w-SGLD/Blob 模拟 $\mathscr{P}_2(\mathbb{R}^m)$ 上的梯度流； • **第 5 章**：SVGD/w-SGLD/Blob/GFSD/GFSF 通过平滑性假设模拟 $\mathscr{P}_2(\mathbb{R}^m)$ 上的梯度流； • **第 6 章**：RSVGD 模拟 $\mathscr{P}_2(\mathscr{M})$ 上的梯度流，pSGHMC 模拟 $\mathscr{P}_2(\mathbb{R}^m)$ 上的 fGH 流		• [61]：LD 模拟 $\mathscr{P}_2(\mathbb{R}^m)$ 上的梯度流； • **第 6 章**：任一 MCMC 方法模拟 $\mathscr{P}_2(\mathscr{M})$ 上的 fGH 流，其中 $\mathscr{M}$ 是一个 fRP 流形	
			• [56]：w-SGLD/Blob 与 LD 等同； • **第 5 章**：SVGD/GFSD/GFSF 与 LD 等同； • **第 6 章**：任一 MCMC 方法都可有一个与之等同的 ParVI 方法，特别地，pSGHMC 与 SGHMC 等同			

开发新 ParVI 方法所需遵循的原则等。而另一方面，优化领域中已经发现，只需要梯度信息的方法可以具有比梯度流模拟更快的收敛速度，即一阶加速方法[76-77,123-124]。但目前的 ParVI 方法则都是基于梯度流模拟的方法，因此尚未充分利用梯度信息，而计算梯度正是计算代价的主要来源。这影响了 ParVI 方法在实际中利用计算资源的效率。

第四，一般 MCMC 动力学系统的行为及其与 ParVI 方法的关系尚不清晰，且一般 MCMC 方法尚无粒子高效的模拟方法，而 ParVI 方法中也缺乏更有效率的动力学系统。LD 已被发现是沃瑟斯坦空间 $\mathscr{P}_2(\mathbb{R}^m)$ 上 KL 散度的梯度流[61]，这一关键发现使得它的收敛行为可以得到深入挖掘[106-107]，并且也可以与 ParVI 方法建立起对应关系[56]。但对于其他众多的 MCMC 方法，目前却还没有类似的认识。这限

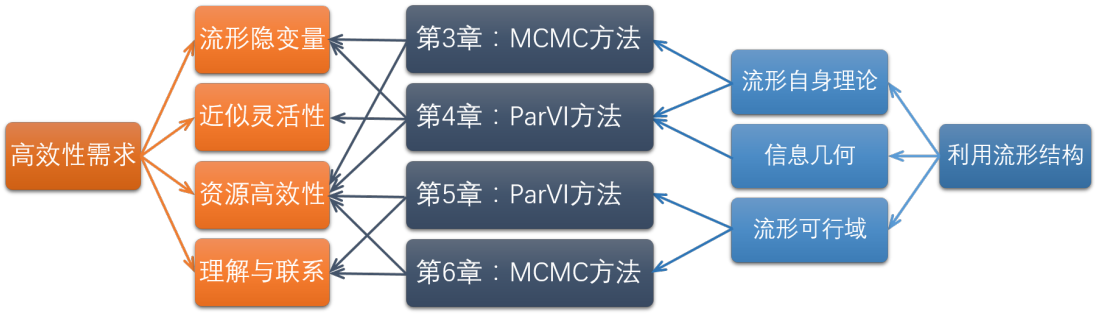


图 1.3　本文各章研究内容所关注对象和利用流形结构的方式及其所解决的高效性需求。

制了人们对 MCMC 方法行为的认识，包括保证它们收敛的因素，以及面对随机梯度时的不同表现（例如 LD 可直接使用而 HMC 则无法使用）。另外，这也制约了除 LD 外的其他 MCMC 方法与 ParVI 方法建立联系，使得其他 MCMC 方法尚无具有更强粒子高效性的 ParVI 形式的模拟方法。而从 ParVI 方法的角度来看，目前方法只对应着 MCMC 方法中 LD 这一种动力学系统，而其他比 LD 更高效的动力学系统（例如 SGHMC，SGNHT 等）则无法在 ParVI 领域发挥优势，从而影响了 ParVI 方法利用计算资源的效率。

本文接下来将针对这些问题提出具体的解决方案或理论分析。本节所提的相关文献组织在表 1.2 中，其中也列出了本文主要贡献，以便对比。

1.3　研究内容及主要贡献

针对上述贝叶斯推理方法仍然面临的高效性需求，本文将深入挖掘利用流形结构的思想和技术，从使用流形自身的理论和方法、结合信息几何技术以及流形可行域观念等三个方面出发，进一步提高各推理方法处理流形隐变量的能力、资源高效性以及近似灵活性，并对各推理方法的基础、行为和联系进行全面的分析和理解。本文所提理论具有一般性和统一性的优势，使得所做分析和所得结论可推广至一般方法，并可将各特定方法建立联系，以及启发新的统一分析和新方法的开发。本文所提方法都具有坚实的理论基础和明晰的直觉，并可有效解决当前环境所带来的高效性需求。主要研究内容根据所关注对象和利用流形结构的方式的不同而分为四个部分，如图 1.3 所示。各部分研究工作的主要贡献由表 1.3 概述，并由表 1.2 进行了与已有工作的对比。

第一部分（第 3 章）提出了随机梯度测地线 MCMC 方法，通过使用流形嵌入技术和随机梯度，提高了处理流形隐变量的 MCMC 方法的可扩展性，增强了其处理大规模数据的时间高效性。为保证使用随机梯度时 MCMC 动力学系统的正确性，本文根据 Ma 等人[120] 的完备表示形式对其进行设计。能够使用随机梯度可使

表 1.3　本文各章贡献描述

章节	贡献描述
第 3 章	提高处理流形隐变量的 MCMC 方法的可扩展性
第 4 章	提高处理流形隐变量的方法的近似灵活性和粒子高效性，提高 ParVI 方法的迭代效率
第 5 章	分析 ParVI 方法所作假设及联系，提高 ParVI 方法的迭代效率及粒子高效性
第 6 章	分析 MCMC 方法的行为及与 ParVI 方法的联系，提高 MCMC 方法的粒子高效性，提高 ParVI 方法的迭代效率

所得方法具有可扩展性，而流形嵌入技术的使用则可保证所得方法可适用于超球面这类没有全局坐标系的流形。所提的两个方法分别是 SGHMC 及 SGNHT 在流形隐变量任务中的推广，但这个推广不能依照直觉进行替换，而需要依照流形的结构进行推导。另外，此部分也开发了使用所提两个方法解决球面混合模型的推理任务的方法，而实验结果表明，使用所提 MCMC 方法可以取得显著优于 ModVI 方法的结果，而其收敛速度则显著地快于所有可适用的 MCMC 方法。

第二部分（第 4 章）提出了黎曼-斯坦因变分梯度下降方法（表 1.2 中 RSVGD 方法），使用流形自身的理论为 ParVI 方法实现了处理流形隐变量的能力，从而为处理流形隐变量的任务首次引入了同时具有近似灵活性和粒子高效性的方法，并利用信息几何方法提高了现有 ParVI 方法的迭代效率。此部分所提方法是将 SVGD 的思想用于流形上所开发的，可看作是 SVGD 的推广。但经过分析，SVGD 方法所采用的一些技术不再适用于流形的情况，例如它在确定最优更新规则时所选择的函数族不再能给出流形上所允许的解。为此，本文基于流形及核函数的理论，设计了新的处理方法，得到了适合流形结构的方法。对于非流形隐变量，将此方法用于似然分布流形上即可通过信息几何得到与隐变量参数化形式无关的高效算法，而对于流形隐变量，为便于在超球面这类没有全局坐标系的流形上使用，本文也推导出了此方法在流形嵌入空间中的表达式。实验结果表明，所提方法在非流形任务上可取得比 SVGD 更高的迭代效率，而在球面混合模型的推理任务上则可取得优于 MCMC 方法（包括第一部分所提方法）的粒子高效性。

第三部分（第 5 章）首先在理论方面，从沃瑟斯坦空间作为 ParVI 方法可行域的视角出发，揭示了各 ParVI 方法使用有限多个粒子时所采用的近似和假设以及它们之间的等价性，并依据此理论开发了两个新的 ParVI 方法，然后利用沃瑟斯坦空间的理论，为所有 ParVI 方法提出了一个加速框架和一个带宽选择方法，从而提高了 ParVI 方法的算力高效性。具体地，所提理论发现 ParVI 方法都是对沃瑟斯坦空间上的梯度流近似模拟的方法，因而都可与 LD 等同，而它们使用有限多个

粒子近似模拟梯度流时要么需要平滑密度要么需要平滑函数。这两种平滑形式所暗藏的等价性表明了各 ParVI 方法之间的等价性，而平滑操作被揭示的必需性则表明 ParVI 方法依赖一个平滑性假设。两种平滑形式也启发了两个新的 ParVI 方法（表 1.2 中 GFSD/GFSF 方法）。在指导实践方面，为提高 ParVI 方法利用沃瑟斯坦梯度信息的效率，本文考虑将流形上的一阶加速优化方法用于沃瑟斯坦空间从而为 ParVI 方法开发了一个加速框架。为此，本文对沃瑟斯坦空间的流形结构进行了深入挖掘。针对 ParVI 方法所使用的平滑核函数，本文基于对平滑操作目的的分析，提出了一个具有原则性的核函数带宽选择方法。实验结果表明，所提加速框架可明显提高各 ParVI 方法的迭代效率，而带宽选择方法则可进一步提高 ParVI 方法的粒子高效性。

第四部分（第 6 章）首先根据流形结构理论，为任一 MCMC 动力学系统在沃瑟斯坦空间上的表示建立了一个统一的理论框架，并根据此表示的结构分析和解释了现有 MCMC 方法的行为，然后在流形可行域的观念下将任一 MCMC 动力学系统与 ParVI 方法建立了联系，从而使 MCMC 方法可以使用具有粒子高效性的 ParVI 方法模拟，同时也使 ParVI 方法可以利用比 LD 更高效的 MCMC 动力学系统来提高算力高效性。为使所提统一理论可涵盖一般的 MCMC 动力学系统，本文建立了一些新的数学概念和关系，进而发现一个一般的 MCMC 动力学系统与一个 fRP 流形的沃瑟斯坦空间上的 fGH 流有一一对应的关系。由于构成 fRP 流的一部分可保持 KL 散度不变而另一部分可在特定子流形上最小化 KL 散度，因此这个沃瑟斯坦空间上的流的表示可为一般 MCMC 动力学系统的收敛性和稳定性等行为提供一个直观的解释。此理论可将 MCMC 动力学系统分为三类。本文依此具体分析了现有的 MCMC 方法，特别是找出了导致 LD 可直接使用随机梯度而 HMC 无法使用随机梯度的根本原因。另外，所提理论也为 MCMC 方法和 ParVI 方法建立起了桥梁，即根据 MCMC 方法的流的表示可以找到对应的 ParVI 模拟方法。特别地，本文为 SGHMC 这个 MCMC 方法开发了对应的 ParVI 方法（表 1.2 中 pSGHMC 方法），从而为 SGHMC 动力学系统带来了具有粒子高效性的模拟方法，同时也为目前只利用了 LD 的 ParVI 领域引入了比 LD 更高效的 SGHMC 动力学系统。实验结果验证了所提 ParVI 方法优于 SGHMC 原方法的粒子高效性以及胜于现有 ParVI 方法的迭代效率。

1.4　论文组织

本文各章节所考虑的方法和技术及其所解决的问题可见图 1.3。具体组织结构如下：

第 1 章为引言部分，介绍了贝叶斯推理的问题和需求，强调了在其中考虑流形结构的意义和价值，梳理了现有研究的状态和问题，并概括了本文的工作内容和贡献。

第 2 章为背景知识，提供了后续各章所需知识的描述，包括流形及其上结构的概念及 MCMC 动力学系统的完备表示形式等。

第 3 章提出了随机梯度测地线 MCMC 方法，利用随机梯度和流形嵌入技术实现了具有可扩展性的处理流形隐变量的 MCMC 方法。

第 4 章提出了黎曼-斯坦因变分梯度下降方法，利用信息几何提高了现有 ParVI 方法的迭代效率，同时为处理流形隐变量的推理任务带来了具有近似灵活性和粒子高效性的方法。

第 5 章为一般 ParVI 方法提出了有限多个粒子近似的理论框架，统一分析了各 ParVI 方法所作假设及其之间关系，并根据此理论框架开发了两个新 ParVI 方法、ParVI 方法的加速框架以及带宽选择方法，提高了 ParVI 方法的迭代效率和粒子高效性。

第 6 章为一般 MCMC 方法提出了作为沃瑟斯坦空间上 fGH 流的统一表示形式，利用 fGH 流的结构分析和解释了各 MCMC 方法的表现和行为，并根据此统一表示形式将一般 MCMC 方法与 ParVI 方法建立联系，进而提出了两个新的 ParVI 方法，提高了 MCMC 模拟方法的粒子高效性，并为 ParVI 领域引入了具有更高迭代效率的动力学系统。

第 7 章对本文的工作和贡献进行总结，并列出了本文工作在未来可以启发的实际方法和理论分析。

为保持文章的整体性，各引理定理命题的证明会直接附于其后。对于一些不足以总结为命题的结论，本文会将它们的推导过程列在“推导”环境中。

第 2 章　背景知识

为正确描述和利用流形结构，本章将从一般流形和黎曼流形这两方面介绍流形概念及其上的结构和性质。针对文中处理 MCMC 方法的工作，本章也将介绍一般 MCMC 动力学系统的完备表示形式。

2.1　流形及其结构

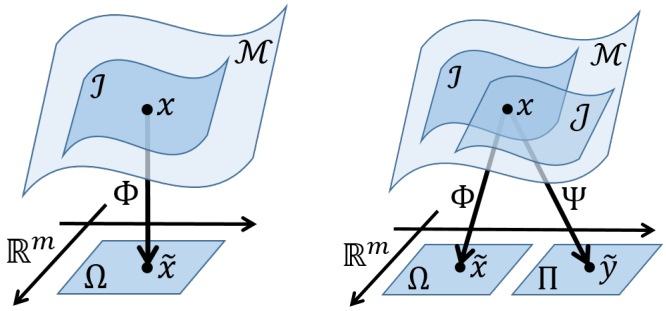
关于流形的详细基础知识，可参见文献[129-133]。

2.1.1　一般流形

2.1.1.1　流形和坐标系

本节首先介绍流形的基本概念。相关概念可参见图 2.1。一个拓扑空间（topological space）[①] $\mathscr{M}$ 可被称为是一个 m 维*流形*（manifold），如果对于它上面的任一点 $x \in \mathscr{M}$，都存在这个点的一个邻域（即包含点 x 的开子集）$\mathscr{I} \subseteq \mathscr{M}$ 使得这个邻域同胚（homeomorphic）于欧氏空间 $\mathbb{R}^m$ 的一个开子集 $\Omega \subseteq \mathbb{R}^m$。其中，拓扑空间可理解为定义了开子集族的集合，进而可定义拓扑空间上函数的连续性以及拓扑空间之间映射的连续性；而两个拓扑空间同胚意味着这两个空间之间存在一个双射（bijection），且这个双射及其逆映射都是连续的。这个定义可以直观地理解为，一个流形可看作是任一局部都表现为欧氏空间的集合。注意到一个流形不需要线性性质，例如加法和数乘，因而它可以包含线性空间所无法描述的集合，例如 m 维超球面 $\mathbb{S}^m := \left\{ x \in \mathbb{R}^{m+1} \,\middle|\, \|x\| = 1 \right\}$。对于上面所考虑的点 x，记从其邻域 $\mathscr{I}$ 到欧氏空间开子集 Ω 的同胚映射为 $\Phi : \mathscr{I} \to \Omega$。由此映射的双射性质便可将开子集 $\mathscr{I}$——这个流形 $\mathscr{M}$ 的局部——用 $\mathbb{R}^m$ 中的元素 $\tilde{x} := \{\tilde{x}^i\}_{i=1}^m$ 来表示，其中的上标 i 表示 $\tilde{x}$ 的（逆变）分量（(contravariant) component）的指标（index）。因此称 $(\mathscr{I}, \Phi)$（或记为 $(\mathscr{I}, \{\tilde{x}^i\}_{i=1}^m)$）为流形 $\mathscr{M}$ 在点 x 处的一个*局部坐标系*（local coordinate system），$\tilde{x} \in \mathbb{R}^m$ 为点 $x \in \mathscr{M}$ 的坐标（coodinates），而 Ω 为流形 $\mathscr{M}$ 的一个（局部）坐标空间（coordinate space）。流形的定义可保证局部坐标系总是存在，但全局坐标系——使整个流形 $\mathscr{M}$ 同胚于 $\mathbb{R}^m$ 的一个开子集的同胚映射——则不一定存在，例如 m 维超球面 $\mathbb{S}^m$ 就不存在全局坐标系。通过坐标系，定义在

① 更加严格地，是一个第二类可数的豪斯多夫空间（second-countable Hausdorff space），即任一开子集都可写为一个固定的可数的开子集族中若干集合的并集的（第二类可数的）、且任意两点都存在不相交邻域的（满足豪斯多夫条件的，即第二类可分的）拓扑空间。



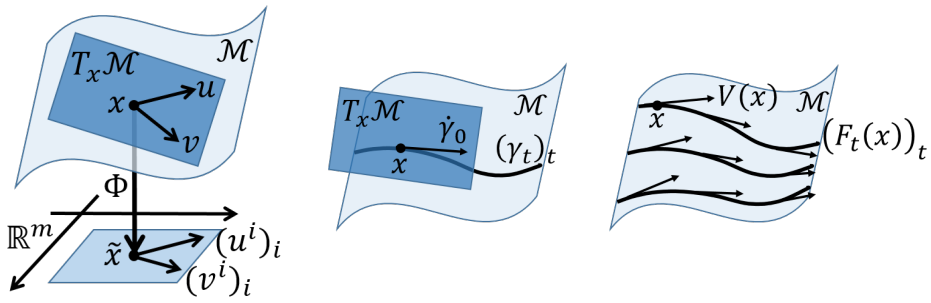
(a) 流形概念及其局部坐标系　(b) 相交坐标系及光滑流形概念

图 2.1　流形、坐标系及光滑流形的概念示意图。

流形上的函数 $f : \mathscr{M} \to \mathbb{R}$（至少局部地）便可被表示为常见的欧氏空间上的函数 $f \circ \Phi^{-1} : \Omega \to \mathbb{R}$，从而便于进一步的分析和显式计算。

人们希望所考虑的流形具有一定的光滑（smooth）性质，并且这个光滑性质可由坐标系体现但又不依赖于坐标系的选取。由于拓扑空间的开子集之间通常都会有交集，因此若两个坐标系 $(\mathscr{I}, \Phi)$ 和 $(\mathscr{J}, \Psi)$（对应坐标空间为 Ω 和 Π）存在交集 $\mathscr{I} \cap \mathscr{J} \neq \varnothing$，则此交集上的点 $x \in \mathscr{I} \cap \mathscr{J}$ 将会有 $\tilde{x}$ 和 $\tilde{y}$ 这两个坐标表示，而人们希望在交集 $\mathscr{I} \cap \mathscr{J}$ 上这两个坐标系所体现的光滑性是等价的。出于此考虑，可要求映射 $\Psi \circ \Phi^{-1} : \Omega \to \Pi, \tilde{x} \mapsto \tilde{y}$ 与映射 $\Phi \circ \Psi^{-1} : \Pi \to \Omega, \tilde{y} \mapsto \tilde{x}$ 都是光滑的。这样的坐标系被称为是相容的。注意到这两个映射都是欧氏空间 $\mathbb{R}^m$ 的开子集上的向量值函数，因此其光滑性可用无穷阶连续可微这个性质来准确描述。相容的坐标系 $(\mathscr{I}, \Phi)$ 和 $(\mathscr{J}, \Psi)$ 在体现流形 $\mathscr{M}$ 的局部光滑性上是等价的。特别地，对于流形上的函数 f，若 $f \circ \Phi^{-1}$ 这个欧氏空间开子集上的函数在 $\mathscr{I} \cap \mathscr{J}$ 上是光滑的，则 $f \circ \Psi$ 也是光滑的，因此由相容坐标系所表现的光滑性就是 f 自身的光滑性，而与具体的坐标系无关。如果一个流形具有一组可覆盖整个流形的相容坐标系族，那么这样的流形可被称为是一个*光滑流形*（smooth manifold）。有了光滑流形的定义，便可进而定义其上函数 f 的光滑性，即为 f 在任一相容坐标系中作为欧氏空间开子集上的函数的光滑性。记 $C^\infty(\mathscr{M})$ 为光滑流形 $\mathscr{M}$ 上所有光滑函数的集合。由于流形的常见性质在非光滑流形的情况下会涉及过多细节，并且常见的流形都是光滑流形，因此本文将只考虑光滑流形、其相容坐标系及其上光滑函数，并将它们分别简称为流形、坐标系和函数。

另外，由于流形上的点 x 和它在坐标系中的坐标 $\tilde{x}$ 可通过坐标映射 Φ 相互转化，这使得某一概念或关系在流形上的表示和在坐标系中的表示等价。因此出于对简化符号并增加可读性的考虑，在不造成歧义的情况下，下文有时也会将流形上的



(a) 切向量、切空间及坐标表示　(b) 光滑曲线的切向量　(c) 向量场及其对应的流

图 2.2　切向量、切空间、向量场和流的概念示意图。

点和它在坐标系中的坐标用同一个符号 x 表示。其具体含义可依据其语境来明确。例如，对于流形上的函数 $f(x)$，偏导数 $\dfrac{\partial}{\partial x^i} f(x)$ 表示的是 $\dfrac{\partial}{\partial \tilde{x}^i} f(\Phi^{-1}(\tilde{x}))\Big|_{\tilde{x}=\Phi(x)}$。

2.1.1.2　切向量和切空间

流形上切向量及切空间的图示可参见图 2.2(a)。流形 $\mathscr{M}$ 在点 $x \in \mathscr{M}$ 处的**切向量**（tangent vector）v 这个概念的正式定义是以 $\mathscr{M}$ 上的微分算符的形式来描述的。具体地，它被定义为光滑函数[①]上的函数 $v : C^\infty(\mathscr{M}) \to \mathbb{R}$，且满足（1）线性性：$v[af + bh] = av[f] + bv[h], \forall f, h \in C^\infty(\mathscr{M}), a, b \in \mathbb{R}$；（2）莱布尼兹法则（Leibniz rule）：$v[fh] = f(x)v[h] + h(x)v[f]$。可以证明，所有据此定义的切向量构成一个 m 维线性空间，称为流形 $\mathscr{M}$ 在点 x 处的**切空间**（tangent space），记为 $T_x\mathscr{M}$。进一步，设点 x 处有一个坐标系 $(\mathscr{I}, \Phi)$，其坐标表示为 $\{\tilde{x}^i\}_{i=1}^m$。据此可以定义点 x 处的 m 个切向量 $\{\partial_i\}_{i=1}^m$，其中 $\partial_i[f] := \dfrac{\partial}{\partial \tilde{x}^i} f \circ \Phi^{-1}(\tilde{x}^1, \cdots, \tilde{x}^m)\Big|_{\tilde{x}=\Phi(x)}$。为符合通常习惯，这个量在文中也会被记为 $\partial_i f$。由坐标映射 Φ 的同胚性质可以证明，这 m 个切向量是线性无关的，因而可以作为切空间 $T_x\mathscr{M}$ 的一组基底，称为**自然基底**。在自然基底下，切空间中任一切向量 v 都可以表示为分量形式：$v = \sum_{i=1}^m v^i \partial_i$，其中（逆变，contravariant）分量可表示为 $v^i = v[\tilde{x}^i] \in \mathbb{R}$，这里 $\tilde{x}^i : \mathscr{I} \to \mathbb{R}$ 被视为流形上的一个函数。基于这个分量表示以及切向量的线性性，可以将切向量 v 在函数 f 上的作用表示为：$v[f] = \sum_{i=1}^m v^i \partial_i f$。需要注意的是，这些表示形式都是需要指定一个坐标系的，因而称它们为**坐标表达式**（coordinate expression）。但是无论是哪个具体的坐标系，这些坐标表达式都是可以适用的，或者说，这些表达式在不同坐标系中给出的是同一个量。这就是一个坐标表达式的**坐标不变性**（coordinate

① 更加严格地，是在 x 的某一邻域上有定义且在此邻域上光滑的函数，其对应函数的集合通常被记为 $C_x^\infty(\mathscr{M})$。

invariance）。流形上合法定义的概念都具有坐标不变的坐标表达式。最后，流形上所有点处的切空间的并集称为*切丛*（tangent bundle）$T\mathscr{M} := \bigcup_{x\in\mathscr{M}} T_x\mathscr{M}$，其中一个元素可表示为 (x, v)（$v \in T_x\mathscr{M}$）。它是一个 $2m$ 维流形。

为简化符号，本文将采用一个已在流形领域中广泛采用的记号规则，称为*爱因斯坦求和规则*（Einstein's summation convention）。这个规则规定，对于在下标（协变分量，covariant component）和上标（逆变分量，contravariant component）中重复出现的指标，自动为其求和，并省略求和记号。使用此规则，上面的坐标表达式可分别写作 $v = v^i\partial_i$ 以及 $v[f] = v^i\partial_i f$。后面将会有更多的例子。另一种紧凑的表达形式是矩阵形式，例如可将 $v[f]$ 写为 $v^\top\nabla f$。基于爱因斯坦求和规则的分量表达式在表达复杂的表达式，特别是同时涉及微分和矩阵运算（例如 MCMC 动力学系统的完备表示形式[120]）时，可给出简洁而清晰的形式，而矩阵形式则通常更易于理解，特别是从实现角度或与常见形式对比。因此，为提高可读性，本文将在不同章节使用不同形式的符号。为将分量形式与矩阵形式联系，本文记 (v^i) 或 $(v^i)_i$ 或 $(v^i)_{m\times 1}$ 为对应的 m 维（列）向量。

注意这里给出的切向量的正式定义与通常意义下所理解的线性空间中的切向量是不同的，但它却是对后者所体现的直觉的抽象和推广。为说明这两者之间的关系，可考虑流形 $\mathscr{M}$ 上的一条经过点 x 的光滑曲线 $\gamma : (-\delta, \delta) \to \mathscr{M}$，其中 $\delta \in \mathbb{R}^+$，且 $\gamma(0) = x$（参见图 2.2(b)）。此曲线在 x 处可定义一个切向量

$$\dot\gamma_0 : C^\infty(\mathscr{M}) \to \mathbb{R}, f \mapsto \left.\frac{\mathrm{d}}{\mathrm{d}t} f(\gamma(t))\right|_{t=0}. \tag{2-1}$$

由上面的定义可验证 $\dot\gamma_0 \in T_x\mathscr{M}$。按此定义，$\dot\gamma_0$ 是作用在函数上的沿着曲线 γ_t 在 x 处的方向导数算符。而通常意义下的线性空间中的切向量则是指极限 $\lim\limits_{t\to 0}\dfrac{\gamma_t - \gamma_0}{t}$，但由于流形上没有定义减法，因此这个定义失效。但此定义还是可以在 x 附近的坐标空间中使用，并得到 $\tilde{\dot\gamma}_0 := \lim\limits_{t\to 0}\dfrac{\Phi(\gamma_t) - \Phi(\gamma_0)}{t} \in \mathbb{R}^m$，而这个向量恰好就是切向量 $\dot\gamma_0$ 在坐标空间中的表示 $(\dot\gamma_0^i)$。而由于 $\dot\gamma_0[f] = \dot\gamma_0^i\partial_i f$，因此正式定义的方向导数也与通常意义下的方向导数表达式 $\tilde{\dot\gamma}_0^\top\nabla f$ 一致。

另外，对于一般的流形，其切空间只在流形上的一点处才有意义，或者说不同点处的切空间是不同的。但若 $\mathscr{M}$ 是一个线性空间，则其各点处的切空间 $T_x\mathscr{M}$ 都同构于它自己 $\mathscr{M}$，对应的同构映射为 $\mathscr{M} \to T_x\mathscr{M}, y \mapsto v_y, v_y[f] := \left.\dfrac{\mathrm{d}}{\mathrm{d}t} f(x + ty)\right|_{t=0}$。而由于一般的流形没有线性结构，因此无法通过这种方式将不同点上的切向量与流形上的点建立联系。一般流形不同点上的切空间的相异性使得不同点上的切向量无法直接进行向量操作，例如减法。流形的这个特点也有其实际的物理意义，例如在广义相对论中，距离较远或者处于强引力场中的两个不同位置上的粒子无法

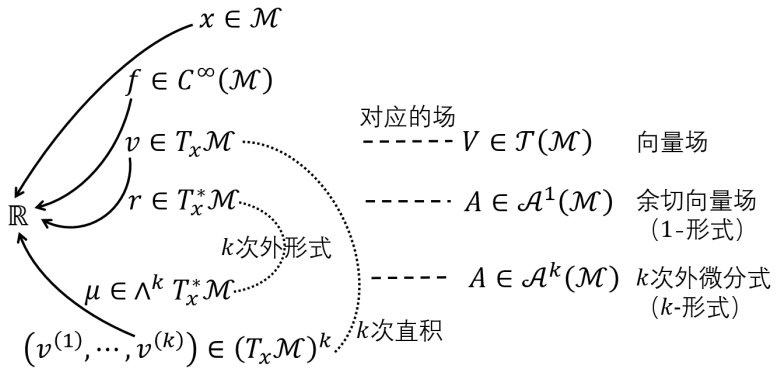


图 2.3　流形上各概念的定义及其之间关系的示意图。

合理定义“相对速度”的概念[134]。（由红移效应所表现出的相对速度其实也体现了时空自身的扭曲，因而也不是通常意义下的相对速度；特别地，这个表象的相对速度可能会超过光速。）

2.1.1.3　向量场，流，以及动力学系统

参见图 2.2(c)，若流形 $\mathscr{M}$ 上的每一点 x 处都有一个切向量 $V(x)$，并且切向量 $V(x)$ 对 x 的依赖关系是光滑的（例如可通过坐标系中的表示 $V^i(\Phi(x))$ 描述），那么称 V 是流形 $\mathscr{M}$ 上的一个光滑切向量场，简称向量场（vector field）。流形 $\mathscr{M}$ 上的所有向量场的集合记为 $\mathscr{T}(\mathscr{M})$。给定一个向量场 V，可以定义与之关联的流（flow），为流形 $\mathscr{M}$ 上的一组曲线 $\{(F_t(x))_t \mid x \in \mathscr{M}\}$，其中穿过 $x \in \mathscr{M}$ 曲线 $(F_t(x))_t$ 满足条件 $F_0(x) = x$，且它在 x 处的切向量 $\dot{F}_0(x)$（定义参见式 (2-1)）正是向量场 V 在点 x 处所给出的切向量 $V(x)$。对于任一向量场，它所对应的流至少局部存在。流或者流中的一条曲线有时也被称为积分曲线（integral curve）。

流形 $\mathscr{M}$ 上的一个向量场也可以描述一个*确定性动力学系统*（deterministic dynamics），即流形上的粒子的位置随时间连续演化的规律 $x(t)$。给定向量场 V，它可定义一个演化规律为 $\dot{x}(t) = V(x(t))$，即粒子在 t 时刻的速度由 $V(x(t))$ 给定。可发现这一定义与流的定义十分类似，因而有 $x(t) = F_t(x(0))$，只不过将流中曲线 $F_t(x)$ 解释为 0 时刻处在 x 位置的粒子随后的演化规律。因此，在本文中涉及流或者动力学系统的地方，将直接使用对应的向量场进行描述。

2.1.1.4　外微分式和流形上的测度

此处将引入更多概念。图 2.3 对这些概念的定义及其之间的关系进行了梳理。

点 x 处的切空间 $T_x\mathscr{M}$ 的（代数）对偶空间（algebraic dual space），即 $T_x\mathscr{M}$ 上所有线性函数的空间，被称为余切空间（cotangent space），并记为 $T_x^*\mathscr{M}$。其中的向量

被称为余切向量（cotangent vector）。类似切丛，流形上所有点处的余切空间的并集称为余切丛（cotangent bundle）$T^*\mathscr{M} := \bigcup_{x\in\mathscr{M}} T_x^*\mathscr{M}$，它也是一个 $2m$ 维流形。对于 $\mathscr{M}$ 上的函数 $f \in C^\infty(\mathscr{M})$，可以定义一个余切向量 $\mathrm{d}f$ 为：$\mathrm{d}f[v] := v[f], \forall v \in T_x\mathscr{M}$。特别地，给定 x 附近的一个坐标系 $(\mathscr{I}, \{x^i\}_{i=1}^m)$，可以定义 m 个余切向量 $\{\mathrm{d}x^i\}_{i=1}^m$，其中 $\mathrm{d}x^i[v] := v[x^i]$，这里 $x^i : \mathscr{I} \to \mathbb{R}$ 被视为流形上的一个函数。这组向量线性无关，因而可构成余切空间 $T_x^*\mathscr{M}$ 的一组基底。在这组基底下，余切向量 $\mathrm{d}f$ 可表示为 $\mathrm{d}f = \partial_i f\, \mathrm{d}x^i$。

接下来考虑切空间 $T_x\mathscr{M}$ 上的 k 重反对称线性函数 μ，或称 k 次外形式（exterior form of degree k）。这就是说，$\mu[v^{(1)}, \cdots, v^{(k)}]$ 是以 k 个向量为参数的标量值函数，它对任一参数 $v^{(i)}$ 都是线性的，且交换任意两个参数的值得到的函数值会改变符号。它是 $T_x\mathscr{M}$ 上的 k 阶张量（tensor）——即 $T_x\mathscr{M}$ 上的 k 重线性函数——的一种特殊情况。特别地，余切向量可以看作是 1 次外形式。所有这样的函数 μ 所构成的线性空间记为 $\wedge^k T_x^*\mathscr{M}$。注意对于 $k > m$，由 μ 的反对称性以及 k 个切向量的线性相关性可知，$\mu = 0$，因而这样的 $\wedge^k T_x^*\mathscr{M}$ 只有一个零元素。

为找到 $\wedge^k T_x^*\mathscr{M}$ 的一组基底，首先考虑 k 次外形式 μ 与 l 次外形式 ν 的外向积（exterior product），或称楔积（wedge product），定义为一个 $\ell = k + l$ 次外形式：

$$\mu \wedge \nu := \frac{(k+l)!}{k!l!}\mathrm{alt}(\mu \otimes \nu) \in \wedge^\ell T_x^*\mathscr{M},$$

其中 $\otimes$ 为张量积（tensor product），即 $(\mu \otimes \nu)[v^{(1)}, \cdots, v^{(k)}, v^{(k+1)}, \cdots, v^{(k+l)}] := \mu[v^{(1)}, \cdots, v^{(k)}]\nu[v^{(k+1)}, \cdots, v^{(k+l)}]$，而 ℓ 重张量 η 的反对称化算符（alternation）$\mathrm{alt}(\eta)$ 则会将 η 映射为一个 ℓ 次外形式：$\mathrm{alt}(\eta)[v^{(1)}, \cdots, v^{(\ell)}] := \frac{1}{\ell!}\sum_{\sigma\in\mathrm{perm}(1,\cdots,\ell)} \mathrm{sign}(\sigma)\eta[v^{(\sigma(1))}, \cdots, v^{(\sigma(\ell))}]$，其中 $\mathrm{perm}(1, \cdots, \ell)$ 表示序列 $(1, \cdots, \ell)$ 的所有排列（permutation）（共有 $\ell!$ 个），而 $\mathrm{sign}(\sigma)$ 则表示排列 σ 的符号，即若序列 $(\sigma(1), \cdots, \sigma(\ell))$ 可由 $(1, \cdots, \ell)$ 经偶数次对换而得，则 $\mathrm{sign}(\sigma) = 1$，否则 $\mathrm{sign}(\sigma) = -1$。由此概念，可对满足 $1 \leqslant i_1 < \cdots < i_k \leqslant m$ 的指标集定义一个 k 次外形式 $\mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k} := \Big(\big((\mathrm{d}x^{i_1} \wedge \mathrm{d}x^{i_2}) \wedge \mathrm{d}x^{i_3}\big) \wedge \cdots\Big) \wedge \mathrm{d}x^{i_k} = \sum_{\sigma\in\mathrm{perm}(i_1,\cdots,i_k)} \mathrm{sign}(\sigma)\, \mathrm{d}x^{\sigma(1)} \otimes \cdots \otimes \mathrm{d}x^{\sigma(k)}$。现考虑任一 k 次外形式 μ。由于它总是一个 k 阶张量，因而它可以表示为：$\mu = \mu_{i_1,\cdots,i_k}\, \mathrm{d}x^{i_1} \otimes \cdots \otimes \mathrm{d}x^{i_k}$，其中 $\mu_{i_1,\cdots,i_k} := \mu[\partial_{i_1}, \cdots, \partial_{i_k}] \in \mathbb{R}$。由外形式 μ 的反对称性，若这 k 个指标 $i_1, \cdots, i_k$ 中有重复的，那么对应的 $\mu_{i_1,\cdots,i_k} = 0$。而对于包含互不相等的 k 个指标的指标集，可以按照它们是否互成排列进行划分。对于一类互成排列的 $k!$ 个指标集，它们给出的 μ 的分量值具有相同的绝对值，因而可以选择满足条件 $1 \leqslant< i_1 < \cdots < i_k \leqslant m$

的指标集作为代表，并只考虑其他指标集所对应的符号。根据这个划分，k 次外形式 μ 可以写为：$\mu = \sum_{1\leqslant<i_1<\cdots<i_k\leqslant m} \mu_{i_1,\cdots,i_k} \sum_{\sigma\in\mathrm{perm}(i_1,\cdots,i_k)} \mathrm{sign}(\sigma)\,\mathrm{d}x^{\sigma(1)}\otimes\cdots\otimes\mathrm{d}x^{\sigma(k)}$。而由 $\mathrm{d}x^{i_1}\wedge\cdots\wedge\mathrm{d}x^{i_k}$ 的定义，上式可表示为：

$$\mu = \sum_{1\leqslant<i_1<\cdots<i_k\leqslant m} \mu_{i_1,\cdots,i_k}\,\mathrm{d}x^{i_1}\wedge\cdots\wedge\mathrm{d}x^{i_k}.$$

注意若仍以爱因斯坦求和规则表示，上式应写为 $\mu = \frac{1}{k!}\mu_{i_1,\cdots,i_k}\,\mathrm{d}x^{i_1}\wedge\cdots\wedge\mathrm{d}x^{i_k}$。注意满足条件 $1\leqslant<i_1<\cdots<i_k\leqslant m$ 的指标集共有 $\binom{m}{k}$ 个，因而任一 k 次外形式都可以用这 $\binom{m}{k}$ 个 k 次外形式 $\mathrm{d}x^{i_1}\wedge\cdots\wedge\mathrm{d}x^{i_k}, 1\leqslant<i_1<\cdots<i_k\leqslant m$ 线性表示。而由于对于这 $\binom{m}{k}$ 个系数 $\mu_{i_1,\cdots,i_k}, 1\leqslant<i_1<\cdots<i_k\leqslant m$，已没有任何条件可限制它们之间的关系，因此这个线性表示也是唯一的。因此，这 $\binom{m}{k}$ 个 k 次外形式便构成了线性空间 $\wedge^k T_x^*\mathscr{M}$ 的一组基底。特别地，$\wedge^k T_x^*\mathscr{M}$ 是 $\binom{m}{k}$ 维的。

上面的考虑都是在一个点 x 处的情况。若流形 $\mathscr{M}$ 上的每一点 x 处都有一个 k 次外形式 $A(x)$，并且它对 x 的依赖关系是光滑的（例如可通过坐标系中的表示描述），那么称 A 是流形 $\mathscr{M}$ 上的一个 k 次外微分式（exterior differential form of degree k），或简称为k-形式（k-form）。流形 $\mathscr{M}$ 上所有 k-形式构成的集合记为 $\mathscr{A}^k(\mathscr{M})$。在 k-形式 A 上可以定义外微分（exterior differential）使之成为一个 $(k+1)$-形式 $\mathrm{d}A := \sum_{1\leqslant<i_1<\cdots<i_k\leqslant m} \mathrm{d}A_{i_1,\cdots,i_k}\wedge\mathrm{d}x^{i_1}\wedge\cdots\wedge\mathrm{d}x^{i_k}$，其中 $\mathrm{d}A_{i_1,\cdots,i_k}$ 即为上面所给出的由函数而定义的余切向量场，或者视作 1-形式。

除了直接考虑在带边区域及其边界上的积分（参见 4.3.1 节）外，外微分式在本文中的重要意义在于它可以定义流形上的测度。这将在 2.1.2.5 节中具体介绍。

2.1.2　黎曼流形

2.1.2.1　黎曼结构

注意到点 x 处的切空间 $T_x\mathscr{M}$ 是一个线性空间，因而可以在其中定义一个内积（inner product）$g_x : T_x\mathscr{M}\times T_x\mathscr{M}\to\mathbb{R}$，即 $T_x\mathscr{M}$ 上的二重对称正定线性函数。若在流形上的每一点处的切空间都有此内积定义，且此内积以光滑的方式依赖于 x，那么这个“内积场”结构 g_x 便被称为*黎曼结构*（Riemannian structure），而具有黎曼结构的流形便被称为*黎曼流形*（Riemannian manifold）。黎曼结构 g_x 是一个二阶张量场，因此它可以在点 x 附近的坐标系中表示为 $g_x = g_{ij}(x)\,\mathrm{d}x^i\otimes\mathrm{d}x^j$，从而使得

切空间 $T_x\mathscr{M}$ 中两向量 $u = u^i\partial_i, v = v^j\partial_j$ 的内积开业表示为 $g_x(u,v) = g_{ij}(x)u^iv^j$。黎曼结构的坐标表示 (g_{ij})（通常也被记为 G）是一个处处对称正定的矩阵，被称为*黎曼度量矩阵*（Riemannian metric matrix）或黎曼度量张量。

2.1.2.2　梯度

仅仅是这一个结构的引入，便可使一个流形具有丰富的结构。首先，切空间 $T_x\mathscr{M}$ 此时变为一个希尔伯特空间（Hilbert space），因此由瑞兹表示定理（Riesz representation theorem）可以建立它与余切空间 $T_x^*\mathscr{M}$ 之间的一个明确的同构关系。特别地，对于任一余切向量 $r \in T_x^*\mathscr{M}$，都在切空间 $T_x\mathscr{M}$ 中存在唯一的切向量，记为 $r^\#$，使得 $g_x(r^\#, v) = r[v], \forall v \in T_x\mathscr{M}$。使用坐标形式表达，可写出 $g_{ij}(r^\#)^iv^j = r_jv^j$ 对任意 $v \in T_x\mathscr{M}$ 成立，因而可以解得 $(r^\#)^i = g^{ij}r_j$，其中 g^{ij} 是 (g_{ij}) 的逆矩阵的矩阵元，二者具有关系 $g^{ik}g_{kj} = \delta^i_j$，这里 $\delta^i_j = \begin{cases} 1, \text{若 } i = j, \\ 0, \text{若 } i \neq j, \end{cases}$ 是克罗内克-德尔塔张量（Kronecker delta tensor）。进一步，由 2.1.1.4 节的介绍，对于函数 $f \in C^\infty(\mathscr{M})$，可定义余切向量 $\mathrm{d}f = \partial_i f(x)\,\mathrm{d}x^i$，因而可以导出它所对应的切向量 $(\mathrm{d}f(x))^\#$。由于这个操作可以在流形上的任一点进行，因而函数 f 定义了流形上的一个向量场 $(\mathrm{d}f)^\#$，被称为*梯度*（gradient），并记为 $\operatorname{grad} f$，即：

$$\operatorname{grad} f := (\mathrm{d}f)^\#.$$

其坐标表达式为：

$$\operatorname{grad} f = g^{ij}\,\partial_i f\,\partial_j.$$

向量场 $-\operatorname{grad} f$ 所对应的流即为*梯度流*（gradient flow）。

这个抽象的定义是与通常意义下的梯度概念等价的。通常意义下函数 f 在点 x 处的梯度被定义为可使 f 最速上升的方向，即：

$$v^* := \underset{v\in T_x\mathscr{M}, \|v\|^2_{T_x\mathscr{M}} := g_x(v,v)=1}{\max\cdot\operatorname{argmax}} \frac{\mathrm{d}}{\mathrm{d}t} f(\gamma_t)\bigg|_{t=0},$$

其中曲线 $\gamma : (-\delta, \delta) \to \mathscr{M}$ 满足 $\gamma_0 = x$，$\dot\gamma_0 = v$（其中 $\dot\gamma_0$ 在 2.1.1.2 节第三段式 (2-1) 中定义），而“$\max\cdot\operatorname{argmax}$”表示目标函数最大值与使目标函数取最大值的向量的数乘。由求导的链式法则，目标函数可以写为：

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} f(\gamma_t)\bigg|_{t=0} &= \frac{\mathrm{d}}{\mathrm{d}t}(f\circ\Phi^{-1})(\Phi(\gamma_t))\bigg|_{t=0} = \partial_i(f\circ\Phi^{-1})(x)\frac{\mathrm{d}}{\mathrm{d}t}(\Phi^i(\gamma_t))\bigg|_{t=0} \\
&= \partial_i f(x)\dot\gamma_0^i = \delta_i^k\partial_k f(x)v^i = g^{kj}g_{ji}\partial_k f(x)v^i = g_{ij}(g^{jk}\partial_k f(x))v^i \\
&= g_x(\operatorname{grad} f(x), v),
\end{aligned}$$

即是一个内积的形式。因此得到：

$$v^* = \underset{v\in T_x\mathscr{M},\|v\|^2_{T_x\mathscr{M}}:=g_x(v,v)=1}{\max\cdot\operatorname{argmax}} g_x(\operatorname{grad} f(x), v) = \operatorname{grad} f(x),$$

这与上面所给出的梯度的定义是一致的。所以从这个角度来说，f 的梯度流也可解释为 f 的最速下降曲线族。

2.1.2.3　距离与测地线

有了黎曼结构之后，便可以度量黎曼流形上光滑曲线的长度。具体地，考虑曲线 $\gamma : [a,b] \to \mathscr{M}$，则可将其长度定义为 $L(\gamma) := \sqrt{\int_a^b \|\dot\gamma_t\|^2_{T_{\gamma_t}\mathscr{M}}\,\mathrm{d}t} = \sqrt{\int_a^b g_{\gamma_t}(\dot\gamma_t,\dot\gamma_t)\,\mathrm{d}t}$，其中曲线的切向量 $\dot\gamma_t \in T_{\gamma_t}\mathscr{M}$ 可参照式 (2-1) 定义。进而可以定义流形上两点之间的距离，为连接两点的所有光滑曲线中距离最短的值，即 $d(x,y) := \inf_{\gamma:\gamma_a=x,\gamma_b=y} L(\gamma)$。若在此距离下此流形是完备的，则由霍普夫-里诺定理（Hopf–Rinow theorem）[135] 可知，存在取得最短距离的曲线。这样的曲线被称为测地线（geodesic），它是欧氏空间中直线概念的推广。

更加准确地说，测地线是由流形上的一个仿射联络（affine connection）这个独立于黎曼结构的结构而定义的。一个仿射联络，或者等价地，一个协变导数（covariant derivative），为两个邻近点上的切向量建立了一个对应关系，进而可以由此对应关系，将一点上的切向量沿着给定曲线对应到另一点上的切向量，即平行移动（parallel transport）或平行输运（parallel translation）。对于给定两点以及连接两点的给定曲线，如果这两点上的两个切向量符合这个对应关系，则称它们是平行的。而测地线则可定义为一类特殊的曲线 γ，使得曲线上任两点 γ_a,γ_b 处的切向量 $\dot\gamma_a,\dot\gamma_b$ 关于所截曲线 $\gamma_{t\in[a,b]}$ 是平行的。此定义的直观解释是，流形上“直线”的各个点上的切向量应该都是相互平行的。这个定义可表示为一个常微分方程的初值问题，因而存在唯一解。对于黎曼流形，其黎曼结构可定义一个特殊的联络，称为黎曼联络（Riemannian connection）或列维-奇维塔联络（Levi-Civita connection）。此联络所定义的测地线刚好就是上一部分中通过最短长度的方式所定义的测地线。此外，平行移动本身，以及与测地线相关的测地流（geodesic flow）和指数映射（exponential map）等概念也会分别在下面各章中加以考虑。这些概念将在对应的章节中加以介绍。

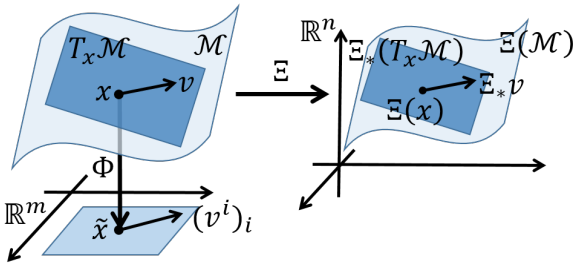


图 2.4　流形及其嵌入空间的示意图。

2.1.2.4　嵌入空间

首先讨论无需黎曼结构的一般流形的情况。相关概念在图 2.4 中予以示意。人们所熟知的流形概念通常都是作为一个欧氏空间 $\mathbb{R}^n$ 的子集而出现的，特别是在对这些流形进行可视化时。而且很多常见的流形也都是以此方式定义的，例如 $n-1$ 维超球面 $\mathbb{S}^{n-1} := \{ x \in \mathbb{R}^n \mid \|x\| = 1 \}$。但在 2.1.1.1 节中所介绍的流形概念则没有借用这个外部的包裹空间，而是直接考察这个集合本身的性质。不过，这种内禀（intrinsic）的定义方式也可以与通常的基于外部空间的流形观念建立起联系，即通过流形的*嵌入*（embedding）来描述。具体地说，对 m 维流形 $\mathscr{M}$ 进行嵌入是指通过一个光滑的单射 $\Xi : \mathscr{M} \to \mathbb{R}^n$ 将它映射到欧氏空间 $\mathbb{R}^n$ 中。由嵌入映射 Ξ 的单射性质，可知如果这样的 Ξ 存在，那么 $n \geqslant m$。对于按 2.1.1.1 节中所介绍的内禀的方式所定义的 m 维流形，惠特尼嵌入定理（Whitney embedding theorem）[136-137] 证明了它总可以嵌入到 $2m$ 维欧氏空间 $\mathbb{R}^{2m}$ 中，从而可以作为欧氏空间 $\mathbb{R}^{2m}$ 的子集。由此，这两种定义流形的观念是等价的。对于具有全局坐标系的流形，其全局坐标系即可视为一个嵌入，并且此种情况下有 $n = m$。而对于没有全局坐标系的流形，其嵌入空间的维度 n 一定大于 m。例如 $n-1$ 维超球面 $\mathbb{S}^{n-1}$ 没有全局坐标系，但它可自然地嵌入在定义它的欧氏空间 $\mathbb{R}^n$ 中。

上述讨论是针对一般流形而言的，即不需要黎曼结构。而当流形 $\mathscr{M}$ 具有黎曼结构时，可以进一步考虑此黎曼结构与嵌入空间 $\mathbb{R}^n$ 为 $\mathscr{M}$ 所带来的结构之间的关系。对于流形 $\mathscr{M}$ 在点 x 处的一个切向量 $v \in T_x\mathscr{M}$，可通过嵌入映射定义一个嵌入空间中的切向量 $\Xi_* v[h] := v[h \circ \Xi], \forall h \in C^\infty(\Xi(\mathscr{M}))$，称为切向量 v 在光滑映射 Ξ 下的前推（push-forward）。令欧氏空间 $\mathbb{R}^n$ 的一组坐标系为 $\{y^a\}_{a=1}^n$（它是全局的），则切向量 $\Xi_* v$ 可通过坐标表示为 $(\Xi_* v)^a \tilde{\partial}_a$，其中 $\tilde{\partial}_a := \dfrac{\partial}{\partial y^a}$ 是 $\mathbb{R}^n$ 中自然基底中的一个向量，而 $(\Xi_* v)^a = \Xi_* v[y^a] = v[y^a \circ \Xi] = v^i \partial_i [y^a \circ \Xi] =: v^i \dfrac{\partial y^a}{\partial x^i}$，这里 $\left(\dfrac{\partial y^a}{\partial x^i} \right)$ 可看作嵌入映射 Ξ 在两个流形各自坐标系下的雅可比矩阵（Jacobian matrix）。由于嵌入映射 Ξ 是单射，因此矩阵 $\left(\dfrac{\partial y^a}{\partial x^i} \right)$ 是满秩的，或者说秩为 m。因此，切空间

$T_x\mathscr{M}$ 在其前推下的像 $\Xi_*(T_x\mathscr{M})$ 也是一个 m 维线性空间，并可看作嵌入空间中的切空间 $T_{\Xi(x)}\mathbb{R}^n$ 的线性子空间。由 2.1.1.2 节中的讨论，线性空间 $\mathbb{R}^n$ 任一点处的切空间 $T_{\Xi(x)}\mathbb{R}^n$ 都同构于它自己，因此 $\Xi_*(T_x\mathscr{M})$ 也可视作欧氏空间 $\mathbb{R}^n$ 的一个 m 维线性子空间。欧氏空间 $\mathbb{R}^n$ 中有一个自然的内积 $y^\top z = \sum_{a=1}^{n} y^a z^a = \delta_{ab} y^a z^b$，即其黎曼度量矩阵为单位矩阵 (δ_{ab})，其中 δ_{ab} 为克罗内克张量的变种。这个内积可为 $\mathbb{R}^n$ 的线性子空间 $\Xi_*(T_x\mathscr{M})$ 定义一个内积。进一步，嵌入映射 Ξ 可以把这个内积引入到流形 $\mathscr{M}$ 的切空间中，从而定义一个黎曼结构。具体地，考虑切空间 $T_x\mathscr{M}$ 中的两个切向量 v, u，可定义它们的内积为 $\langle v, u\rangle := (\Xi_* v)^\top (\Xi_* u) = \delta_{ab} v^i \dfrac{\partial y^a}{\partial x^i} u^j \dfrac{\partial y^b}{\partial x^j}$，因此对应的黎曼结构具有坐标表达式 $\left(\delta_{ab} \dfrac{\partial y^a}{\partial x^i} \dfrac{\partial y^b}{\partial x^j}\right)$。这个黎曼结构也可看作是嵌入空间中的黎曼结构在映射 Ξ 下的拉回（pull-back）$\Xi^*\delta$。一个嵌入（对应的嵌入映射为 Ξ）被称为是一个等距嵌入（isometric embedding），如果来自嵌入空间的黎曼结构和流形 $\mathscr{M}$ 本身的黎曼结构是一致的。这意味着，$g_{ij} = \delta_{ab} \dfrac{\partial y^a}{\partial x^i} \dfrac{\partial y^b}{\partial x^j}$。以矩阵形式表示，这个关系可以写为 $G = J^\top J$，其中矩阵 $G_{ij} := g_{ij}$，而 $J_{ai} := \dfrac{\partial y^a}{\partial x^i}$。对于常见的定义为欧氏空间的子集的流形，例如超球面，其黎曼结构往往也由此欧氏空间确定，因此此欧氏空间就是它的等距嵌入空间。最后，由纳什嵌入定理（Nash embedding theorem）[138] 可知，任一黎曼流形都可以等距嵌入在一个欧氏空间中。

2.1.2.5　测度与积分

首先讨论无需黎曼结构的一般流形的情况。在 2.1.1.4 节中提到，流形上的测度可由外微分式所描述。本节首先为此给出一个直观的解释。流形 $\mathscr{M}$ 上一点 x 上的 m 个线性无关的切向量可构成一个平行多面体（parallelepiped），它可视作流形 $\mathscr{M}$ 在 x 处的一个体积微元，而一个测度（measure）则应为流形 $\mathscr{M}$ 每一点上的体积微元赋予一个非负值，作为其体积。由于体积微元的体积应与构成它的各向量成线性关系，并且可将这些向量的顺序作为这个体积的定向，使得任何两个向量交换顺序所得体积值应改变符号。因此在流形上的一个点 x 处，测度可以用切空间 $T_x\mathscr{M}$ 上的 m 重反对称线性函数来表示，也就是说，流形 $\mathscr{M}$ 上的一个测度可以用 $\mathscr{M}$ 上的一个 m 次外微分式表示。正式地，称（可定向的，orientable）流形 $\mathscr{M}$ 上的最高阶处处非零的外微分式为体积形式（volume form），即 $\mathscr{A}^m(\mathscr{M})$ 中处处非零的元素[①]。所有体积形式构成一个一维线性空间。给定体积形式 $\omega \in \mathscr{A}^m(\mathscr{M})$，考虑它所对应的测度给出的开子集 $\mathscr{I}$ 的体积。若 $\mathscr{I}$ 可与 $\mathbb{R}^m$ 的一个开子集通过映射 Φ 同

① 处处非零的光滑 m-形式 ω 不会改变定向（orientation），表现为在坐标系中 $\omega_{1\cdots m}$ 不会改变符号。因此无需要求 ω 处处为正，处处为负的情况可认为只是所取定向不同。

胚，则此体积由 $\mathbb{R}^m$ 上通常的勒贝格（Lebesgue）积分 $\int_{\Phi(\mathscr{I})} \omega_{1\cdots m}(x)\,\mathrm{d}x^1\wedge\cdots\wedge\mathrm{d}x^m$ 给出，其中 $\omega_{1\cdots m}$ 是体积形式 ω 将 $(\mathscr{I},\Phi)$ 视为坐标系的坐标表达式。可以证明，这个表达式是坐标不变的。对于一个坐标系无法覆盖 $\mathscr{I}$ 的情况，可将它分为若干坐标系覆盖，由于这个体积定义是坐标不变的，因此不同的坐标系划分方式会给出同一个体积值。在给定了体积形式的流形上便可定义一个与之绝对连续（absolutely continuous）的测度——或者称为分布——的密度函数（density function），即相对于此体积形式的拉东-尼科迪姆导数（Radon-Nikodym derivative）。最后，关于流形上的积分，最有名的结论便是斯托克斯定理（Stokes' theorem）：对于流形 $\mathscr{M}$ 上的一个 $(m-1)$-形式 $\omega\in\mathscr{A}^{m-1}(\mathscr{M})$ 和一个（第二类可数的）带边区域 $\mathscr{I}\subseteq\mathscr{M}$，$\int_{\mathscr{I}}\mathrm{d}\omega=\int_{\partial\mathscr{I}}\omega$，其中 $\partial\mathscr{I}$ 是 $\mathscr{I}$ 的边界。微积分中常见的格林公式、高斯定理和（狭义的）斯托克斯公式都是此定理的特例。

上述讨论是针对一般流形而言的，即不需要黎曼结构。而当流形 $\mathscr{M}$ 具有黎曼结构时，便可定义*黎曼体积形式*（Riemannian volume form）$\omega_g:=\sqrt{|G|}\,\mathrm{d}x^1\wedge\cdots\wedge\mathrm{d}x^m$，其中 $G=(g_{ij})$ 是黎曼度量矩阵，而 $|G|$ 则表示其行列式。黎曼体积形式 ω_g 是通过坐标表达式定义的，但可以证明，此坐标表达式具有坐标不变性，因而这个定义是不依赖于坐标系选取的，从而成为流形上得到合法定义的体积形式。由于它的坐标不变性，流形 $\mathscr{M}$ 上一个分布关于它的密度函数 p 便也是坐标不变的。在流形上，有时候也会考虑关于其某个特定的坐标系中的勒贝格测度的密度函数 p_L。它与关于黎曼体积形式的密度函数的关系为 $p_L=p\sqrt{|G|}$。另外，注意在一个黎曼流形 $\mathscr{M}$ 的嵌入空间 $\mathbb{R}^n$ 中，其勒贝格测度会为 $\Xi(\mathscr{M})$ 引入一个测度，即豪斯多夫测度（Hausdorff measure）。它也可以为流形 $\mathscr{M}$ 引入一个测度。注意对于等距嵌入的情况，这个测度正好就是流形 $\mathscr{M}$ 的黎曼体积形式。特别地，表示在 $\Xi(\mathscr{M})$ 上关于豪斯多夫测度的密度函数正好就是 $p\circ\Xi^{-1}$。

流形上的结构十分丰富，本节只介绍了一些基础的概念。下文各章中有的还会涉及更多的流形结构，这些概念将会在对应章节介绍。

2.2　一般 MCMC 动力学系统的完备表示形式

对 MCMC 方法最根本的要求是它可保持目标分布 p 在它的作用下不变，即 p 是其平稳分布（stationary distribution）。Ma 等人[120] 针对欧氏空间 $\mathbb{R}^m$ 中以分布 p 为平稳分布的一般动力学系统，给出了一个完备的表示形式（complete recipe）。这个表示形式以欧氏空间 $\mathbb{R}^m$ 中的一个扩散过程（diffusion process）来描述这样的动

力学系统：

$$\begin{aligned} \mathrm{d}x &= H(x)\,\mathrm{d}t + \sqrt{2D(x)}\,\mathrm{d}B_t(x), \\ H^i(x) &= \frac{1}{p(x)}\partial_j\Big(p(x)\big(D^{ij}(x) + Q^{ij}(x)\big)\Big), \end{aligned} \tag{2-2}$$

其中 $D_{m\times m}$ 是任意一个半正定（positive semi-definite）矩阵，称为扩散矩阵（diffusion matrix），$Q_{m\times m}$ 是一个任意反对称（skew-symmetric）矩阵，称为卷曲矩阵（curl matrix），而 $B_t(x)$ 表示 $\mathbb{R}^m$ 中的标准布朗运动（Brownian motion）。第一项 $H(x)\,\mathrm{d}t$ 代表一个确定性的漂移（drift），而第二项 $\sqrt{2D(x)}\,\mathrm{d}B_t(x)$ 则代表了随机性的扩散（diffusion）。当 D 是（严格）正定（positive definite）时，上式所描述的动力学系统将只有 p 这一个平稳分布。另外，这个表示形式是完备的，这是说任何能够以分布 p 为平稳分布的扩散过程都可以表示成这个形式。

这个表示形式给出了 MCMC 动力学系统的一个统一的表达形式，同时也方便了对 MCMC 动力学系统做统一的分析。在大规模贝叶斯推理任务中，随机梯度（stochastic gradient）这个在一个随机选取的子数据集上所得到的梯度 $(\partial_j \log p)$ 的有噪估计，对于数据方面的可扩展性具有关键作用。利用上述表示形式可以发现，当 $D \neq 0$ 时，MCMC 动力学系统是可以与随机梯度相容的，因为由随机梯度给漂移项带来的噪声的方差是扩散项所带来的噪声的方差的高阶小量[113,120]。另外对于许多 MCMC 方法，变量被取为所采变量 Z 的增广变量 $x = (Z, r)$，其中 r 是一个新引入的辅助变量。进而通过引入条件分布 $p(r|Z)$，可得增广变量的目标分布 $p(x) = p(Z)p(r|Z)$。这个做法可以促进对应的动力学系统探索样本空间中更加广阔的区域从而降低样本之间的自相关性并提高收敛效率[97,119,139]。

第 3 章　随机梯度测地线 MCMC 方法

本章将提出两个随机梯度 MCMC 方法，它们率先为处理没有全局坐标系的流形（例如超球面）上的推理任务引入了具有可扩展性的方法，即由使用随机梯度而带来的高效处理大规模数据的能力。为使用随机梯度，本章为这两个 MCMC 方法设计了新的合适的动力学系统，而为高效处理没有全局坐标系的流形结构，这两个 MCMC 方法考虑在流形的嵌入空间中利用测地流方程进行模拟。这个模拟方法同时也是二阶的，且不包含内循环，因而可以更准确和高效。本章也开发了基于这两个 MCMC 方法解决球面混合模型后验推理任务的方法。合成数据上的实验验证了两方法的正确性和可用性，而在球面混合模型上的真实数据实验则展示了它们处理大规模数据的高效性。

3.1　研究动机

基于动力学系统的马尔可夫链蒙特卡罗方法（dynamics-based Markov chain Monte Carlo，以下简称 MCMC 方法）是一类通过模拟一个动力学系统进行采样的方法。它们已成为贝叶斯推理方法的主力，其中一些知名的方法包括郎之万动力学系统方法（Langevin dynamics, LD）[58] 和哈密顿蒙特卡罗方法（Hamiltonian Monte Carlo, HMC）[96-97] 等。这类方法有许多针对不同任务的变种，其中的一类为处理隐变量在特定黎曼流形上的贝叶斯模型，考虑如何从黎曼流形上进行采样。测地线蒙特卡罗方法（geodesic Monte Carlo, GMC）[44] 就是这类方法中的突出代表。为了处理流形结构，它采用了流形嵌入技术，即在流形的欧氏等距嵌入空间中进行动力学系统的模拟（参见 2.1.2.4 节）。由于任一流形在其嵌入空间中均可全局表示，因而使用流形嵌入技术可以解除流形必须有全局坐标系这个限制，从而可适用于超球面（hypersphere）这样的流形。另外，由于黎曼流形的几何结构已经在等距嵌入的过程中体现，因此在嵌入空间中不需要显式表示诸如黎曼度量矩阵等结构，从而极大地简化了计算。而且常见的流形都是在其欧氏等距嵌入空间中定义的，因而考虑其嵌入空间就十分方便且自然。例如 $n-1$ 维超球面被定义为 n 维欧氏空间中单位模长向量的集合：$\mathbb{S}^{n-1} := \{\, x \in \mathbb{R}^n \mid \|x\| = 1 \,\}$，且其黎曼结构也是由 $\mathbb{R}^n$ 中继承的，因而此欧氏空间 $\mathbb{R}^n$ 自然就是 $\mathbb{S}^{n-1}$ 的等距嵌入空间。为了在嵌入空间中对动力学系统进行模拟，GMC 方法采用了测地线积分器（geodesic integrator），即使用流形的闭式测地流（geodesic flow）来更新流形上的点和向量。测地流可看作流形上的“匀速直线运动”，它是欧氏空间中点的匀速直线运动与向

量的平移的推广。使用这一技术要求流形具有闭式测地流，但对于一些常见的流形比如单纯形（simplex），超球面，斯蒂菲尔流形（Stiefel manifold）[42-43]，双曲流形（hyperbolic manifold）[38] 等，这个要求可以得到满足。另外 GMC 方法还可以用于从一个截断分布中进行采样[126]。

黎曼流形 MCMC 方法中另一个代表是受限哈密顿蒙特卡罗（constrained Hamiltonian Monte Carlo, CHMC）[125]。它可处理的流形是所有可用一个欧氏空间中的限制条件来定义的流形，因而比 GMC 方法具有更广泛的适用范围。但是 CHMC 方法的模拟过程中没有充分利用流形的几何结构（事实上更广泛的流形的几何结构可能会很复杂因而难以利用），所以 CHMC 方法难以使用测地线积分器，因而在其模拟过程中需要使用内循环，导致它不够高效。其他的黎曼流形 MCMC 方法，例如黎曼流形郎之万动力学系统方法（Riemannian manifold Langevin dynamics, RLD）和黎曼流形哈密顿蒙特卡罗方法（Riemannian manifold Hamiltonian Monte Carlo, RHMC）（二者皆出自 [127]）只可在流形的坐标空间中进行模拟，因而难以用于诸如超球面这样的没有全局坐标系的流形。具体来说，在坐标空间中进行模拟需要显式计算黎曼度量矩阵，并需要在模拟过程中不断地变更坐标系。更加麻烦的是，在坐标系的边缘，由于坐标系的问题，黎曼度量矩阵会变得奇异，从而导致实现中很容易出现数值问题。除此之外，RMHMC 方法也需要内循环。

虽然 GMC 方法具有显著优势，但它的可扩展性（scalability），即高效处理大规模数据集的能力，并不理想。针对提升 MCMC 方法的可扩展性这一需求，使用随机子数据集是一个有效的方法，亦即在每次需要计算梯度时，从原数据集中随机采出一个子数据集，并在这个子数据集上估计梯度。这个估计是真实梯度的一个有噪但无偏的估计，称作随机梯度（stochastic gradient）。在随机梯度 MCMC 方法方面，Welling 等人[60] 率先进行了探索，提出了随机梯度郎之万动力学系统方法（stochastic gradient Langevin dynamics, SGLD）。随后，Chen 等人[115] 将 HMC 方法拓展为可以使用随机梯度的版本，即随机梯度哈密顿蒙特卡罗方法（stochastic gradient Hamiltonian Monte Carlo, SGHMC）。他们发现，对于 HMC，必须在动力学系统中引入一个摩擦力项才能正确采样。Ding 等人[119] 提出了随机梯度诺泽-胡佛恒温器方法（stochastic gradient Nosé-Hoover thermostats, SGNHT），其中引入的恒温器变量（thermostats）可自动平衡摩擦力项和梯度噪声。Gan 等人[117] 通过使用多维恒温器变量，进一步将此方法拓展为多变量 SGNHT 方法（multivariate SGNHT, mSGNHT）。为统一表述这些随机梯度 MCMC 方法的动力学系统，Ma 等人[120] 为这些动力学系统开发了一个完备的一般化表示形式。针对在流形上使用随机梯度进行采样的任务，Patterson 等人[18] 和 Ma 等人[120] 分别开发了 RLD

和 RHMC 的随机梯度版本，称作随机梯度黎曼郎之万动力学系统方法（stochastic gradient Riemannian Langevin dynamics, SGRLD）和随机梯度黎曼哈密顿蒙特卡罗方法（stochastic gradient Riemannian Hamiltonian Monte Carlo, SGRHMC）。然而，这两种方法仍然需要流形具有全局坐标系，因而能够适用没有全局坐标系的流形（例如超球面）的可扩展的 MCMC 方法目前仍处于空白。

本章将提出两个新的随机梯度 MCMC 方法：随机梯度测地线蒙特卡罗方法（stochastic gradient geodesic Monte Carlo, SGGMC）和测地线随机梯度诺泽-胡佛恒温器方法（geodesic stochastic gradient Nosé-Hoover thermostats, gSGNHT）。这两个方法是首批可在没有全局坐标系的流形（例如超球面）上高效地进行大规模数据的后验采样的方法。在设计动力学系统方面，本章所考虑的动机希望得到可以使用随机梯度的动力学系统。但如 Chen 等人[115] 所强调的那样，当梯度中出现噪声时，原动力学系统需要恰当的改进才能正确采样。本章利用 Ma 等人[120] 所提出的动力学系统的完备表示形式来解决这一棘手的问题。在模拟动力学系统方面，所设计的 SGGMC 和 gSGNHT 的动力学系统适合使用二阶积分器（second-order integrator）[113] 进行针对离散时间的模拟，从而得到误差更小的可行算法。具体来说，一个积分器是 k 阶的，意味着它在第 L 步迭代后，由离散误差所产生的样本均值期望偏差及均方误差分别可由 $\mathcal{O}(L^{-k/(k+1)})$ 和 $\mathcal{O}(L^{-2k/(2k+1)})$ 的增长阶所控制[113]，因此更高的阶数意味着更高的近似精度。本章将使用对称分解积分器（symmetric splitting integrator, SSI）框架[113] 为所提的两个动力学系统设计二阶积分器进行模拟。在处理流形结构方面，本章参照 GMC 方法[44]，使用流形嵌入技术和测地线积分器，从而可以适用没有全局坐标系的流形并摆脱模拟过程中的内循环，因而比 SGRLD 和 SGRHMC 方法适用性更广，且比 CHMC 高效。由于可以使用随机梯度，这两个方法的可扩展性大大强于 GMC 方法。所提 SGGMC 和 gSGNHT 方法与已有方法的对比列于表 3.1 中。可以发现所提方法是率先同时实现表中所列四项优势的方法。另外，与 GMC 方法类似，所提方法也可以用于截断分布的高效后验采样[126]。

最后，本章用所提的 SGGMC 和 gSGNHT 方法成功地解决了球面混合模型（spherical admixture models, SAM）[22] 的大规模数据的后验推理问题。SAM 模型定义了一个层次化的生成过程来建模单位向量形式的数据（方向类数据），亦即处于超球面上的数据。为了更好地为这样的数据建模，SAM 模型将它的全局隐变量（即话题变量）也选择在了超球面上，因而它的后验推理问题成为了近似一个定义在超球面上的分布的问题。由于这种流形结构，加之模型的先验与似然不共轭，SAM 的后验推理问题十分具有挑战性。已有的 MCMC 方法或者不适用此问题，或者无法做到可扩展性，而变分推理方法由于近似能力所限，无法得到高质

表 3.1　一些 MCMC 方法的对比。

MCMC 方法	可扩展性	无需内循环	适用无全局坐标系的流形	积分器阶数
LD[58]	×	√	–	一阶
HMC[97]	×	√	–	二阶
GMC[44]	×	√	√	二阶
RLD[127]	×	√	×	一阶
RHMC[127]	×	×	×	二阶 †
CHMC[125]	×	×	√	二阶 †
SGLD[60]	√	√	–	一阶
SGHMC[115]/SGNHT[119]	√	√	–	一阶 ‡
SGRLD[18]/SGRHMC[120]	√	√	×	一阶
SGGMC/gSGNHT（所提方法）	√	√	√	二阶

–：该方法不支持流形上的采样；†：该方法的积分器不属于 SSI 框架；‡：SGHMC 和 mSGNHT 方法的 2 阶积分器版本已分别由 Chen 等人[113] 和 Li 等人[140] 开发。

量的结果。所提的 SGGMC 和 gSGNHT 方法则可在取得高质量结果的同时也能高效处理大规模数据。真实数据集上的实验结果表明，所提方法是目前 SAM 模型的后验推理问题上效率最高且结果最好的方法。

3.2　随机梯度测地线 MCMC 方法

本节正式开发 SGGMC 和 gSGNHT 方法。首先设计这两个方法的动力学系统，然后为它们开发在嵌入空间中的二阶积分器以实现算法。本章中所考虑的情形是为处于 m 维黎曼流形 $\mathcal{M}$ 上的隐变量 $Z \in \mathcal{M}$ 开发可从其后验分布中进行高效采样的方法。这个目标分布可通过其关于坐标空间中勒贝格测度的密度函数 p_L 或者其关于嵌入空间中的豪斯多夫测度的密度函数 p 来表示（参见 2.1.2.5 节）。

3.2.1　动力学系统的设计

本节利用 Ma 等人[120] 的动力学系统完备表示形式（参见 2.2 节）来开发动力学系统，以期所得动力学系统的平稳分布即为目标分布从而可以正确采样。注意到这个表示形式只适用于欧氏空间，本节考虑在流形的坐标空间中来开发动力学系统而不是嵌入空间 $\Xi(\mathscr{M})$（参见 2.1.2.4 节）。本节也因此使用目标分布在坐标空间中的密度函数 p_L。不过为摆脱对流形 $\mathscr{M}$ 具有全局坐标系的限制，本章将会在嵌入空间中对所得动力学系统进行模拟（参见本章 3.2.2 节）。

SGGMC 的动力学系统　SGGMC 方法为目标变量 $Z \in \mathscr{M}$ 引入动量（momentum）变量 $r \in \mathbb{R}^m$，并考虑增广变量 $x = (Z, r) \in \mathbb{R}^{2m}$ 的动力学系统。为增广变量 x 定义如下目标分布：$p_L(x) = p_L(Z)|G(Z)|^{-\frac{1}{2}} \exp\{-\frac{1}{2} r^\top G(Z)^{-1} r\}$，其中 G 是目标变量所在黎曼流形 $\mathscr{M}$ 在坐标空间中的黎曼度量矩阵（Riemannian metric matrix）。注意此增广目标分布关于目标变量 Z 的边缘分布正是目标分布 $p_L(Z)$。

此处引入的动量 r 这个辅助变量和增广变量的目标分布有一个经典力学中的来源，而这个来源也可将动量 r 解释为流形 $\mathscr{M}$ 上的余切向量（cotangent vector）。经典力学（可参见著作 [141-142]）中有拉格朗日形式和哈密顿形式这两种等价的形式。其中拉格朗日形式的核心是拉格朗日量（Lagrangian），它是粒子位置 Z 和速度 $\dot{Z}$ 的函数，定义为粒子动能（kinetic energy）与势能（potential energy）的差：$\mathfrak{L}(Z, \dot{Z}) = \frac{1}{2} \dot{Z}^\top G(Z) \dot{Z} + \log p_L(Z)$，这里 $\frac{1}{2} \dot{Z}^\top G(Z) \dot{Z}$ 是动能，而势能被定义为 $-\log p_L(Z)$。给定拉格朗日量之后，粒子的运动规律便可由拉格朗日方程描述。从黎曼流形的角度来看，速度可看作切向量 $\dot{Z} \in T_Z \mathscr{M}$，因而拉格朗日量可看作流形的切丛 $T\mathscr{M}$ 上的函数，其中动能项即为切向量范数的平方的一半。由拉格朗日量可以得到哈密顿量，进而可以通过哈密顿方程以哈密顿形式来描述粒子的运动规律。这是通过勒让德变换（Legendre transformation）得到的，而动量正是在这个过程中引出的，它被定义为 $r := \dfrac{\partial \mathfrak{L}}{\partial \dot{Z}} = G(Z)\dot{Z}$。回顾 2.1.2.2 节中所介绍的由流形的黎曼结构而引出的切空间与余切空间的同构，可发现若以 $\tilde{r} = \tilde{r}_i \, \mathrm{d}x^i \in T_Z^* \mathscr{M}$ 表示一个余切向量，其对应的切向量 $\tilde{r}^\# \in T_Z \mathscr{M}$ 可由 $g_Z(\tilde{r}^\#, v) = \tilde{r}[v], \forall v \in T_Z \mathscr{M}$ 唯一确定。由此可解得坐标表示为 $(\tilde{r}^\#)^j = g^{ij} \tilde{r}_i$，即 $\tilde{r} = G\tilde{r}^\#$（矩阵形式），且这个对应是双射。若取切向量 $\tilde{r}^\#$ 为速度 $\dot{Z}$，则对应的余切向量 $\tilde{r}$ 的向量表示即为 $r = G(Z)\dot{Z}$。因此可将动量解释为余切向量。得到动量之后，可将速度用动量表示为 $\dot{Z} = G^{-1}(Z) r$，便可进一步通过勒让德变换定义哈密顿量为：

$$\begin{aligned} \mathfrak{H}(x) :=& \left(r[\dot{Z}] - \mathfrak{L}(Z, \dot{Z}) \right) \Big|_{\dot{Z} = G^{-1}(Z) r} = \left(r^\top \dot{Z} - \mathfrak{L}(Z, \dot{Z}) \right) \Big|_{\dot{Z} = G^{-1}(Z) r} \\ &= \frac{1}{2} r^\top G(Z)^{-1} r - \log p_L(Z). \end{aligned}$$

哈密顿动力学系统的一个重要性质就是可保持未归一化密度函数为 $\exp\{-\mathfrak{H}(x)\}$ 的分布不变。这一分布正是上面所定义的增广变量 x 的目标分布 $p_L(x)$。但由于哈密顿动力学系统无法使用随机梯度，本节将设计一个新的也以 $p_L(x)$ 为平稳分布的动力学系统。

在 Ma 等人[120] 的完备表示形式（参见 2.2 节及式 (2-2)）下，为 SGGMC 的动力学系统定义其扩散矩阵（diffusion matrix）$D(x)$ 和卷曲矩阵（curl matrix）$Q(x)$

如下：

$$D(x)=\begin{pmatrix}0 & 0\\ 0 & J(Z)^{\top}CJ(Z)\end{pmatrix},\quad Q(x)=\begin{pmatrix}0 & -I\\ I & 0\end{pmatrix},$$

其中，C 是任一个 $n\times n$ 维对称正定矩阵，而 $n\times m$ 维满秩矩阵 $J(Z)$ 定义为：

$$J(Z)_{ij}=\frac{\partial \Xi^{i}(Z)}{\partial Z^{j}}. \tag{3-1}$$

继而根据式 (2-2)，所对应的动力学系统为：

$$\begin{cases}\mathrm{d}Z=G^{-1}r\,\mathrm{d}t,\\ \mathrm{d}r=\quad \nabla_Z\log p_L(Z)\,\mathrm{d}t-\dfrac{1}{2}\nabla_Z\log|G|\,\mathrm{d}t-J^{\top}CJG^{-1}r\,\mathrm{d}t\\ \qquad -\dfrac{1}{2}\nabla_Z\left[r^{\top}G^{-1}r\right]\mathrm{d}t+\mathcal{N}(0,2J^{\top}CJ\,\mathrm{d}t).\end{cases} \tag{3-2}$$

gSGNHT 的动力学系统　除了动量变量 $r\in\mathbb{R}^m$ 外，gSGNHT 方法为目标变量 Z 又引入一个称作恒温器（thermostats）的变量 $\xi\in\mathbb{R}$，并考虑增广变量 $x=(Z,r,\xi)\in\mathbb{R}^{2m+1}$。为增广变量 x 引入如下目标分布：$p_L(x)=p_L(Z)|G(Z)|^{-\frac{1}{2}}\exp\{-\frac{1}{2}r^{\top}G(Z)^{-1}r-\frac{m}{2}(\xi-C)^2\}$，其中 $C\in\mathbb{R}^{+}$ 是一个标量。此增广目标分布关于目标变量 Z 的边缘分布仍然是目标分布 $p_L(Z)$。为 gSGNHT 的动力学系统定义其扩散矩阵 $D(x)$ 和卷曲矩阵 $Q(x)$ 如下：

$$D(x)=\begin{pmatrix}0 & 0 & 0\\ 0 & CG(Z) & 0\\ 0 & 0 & 0\end{pmatrix},\quad Q(x)=\begin{pmatrix}0 & -I & 0\\ I & 0 & r/m\\ 0 & -r^{\top}/m & 0\end{pmatrix}.$$

根据式 (2-2)，可以得到 gSGNHT 的动力学系统为：

$$\begin{cases}\mathrm{d}Z=G^{-1}r\,\mathrm{d}t,\\ \mathrm{d}r=\nabla_Z\log p_L(Z)\,\mathrm{d}t-\dfrac{1}{2}\nabla_Z\log|G|\,\mathrm{d}t-\xi r\,\mathrm{d}t-\dfrac{1}{2}\nabla_Z\left[r^{\top}G^{-1}r\right]\mathrm{d}t+\mathcal{N}(0,2CG\,\mathrm{d}t),\\ \mathrm{d}\xi=(\dfrac{1}{m}r^{\top}G^{-1}r-1)\,\mathrm{d}t.\end{cases} \tag{3-3}$$

这两个动力学系统均为原创。它们的平稳分布对目标变量 Z 的边缘分布都是目标分布 $p_L(Z)$，因而都可以正确采样。它们分别是 SGHMC 和 SGNHT 方法在黎曼流形上的拓展。如此形式的动力学系统也适合开发二阶测地线积分器进行模拟，这是相较于 SGRHMC 方法的优势。

3.2.2　嵌入空间中的模拟

本节将为所开发的动力学系统设计积分器用于离散时间的数值模拟。本节使用对称分解积分器（symmetric splitting integrator, SSI）框架[113] 为它们设计积分器，这样可以保证所设计的积分器是二阶的。SSI 框架的想法是首先将动力学系统分解为闭式可解的若干部分，然后用这些闭式解交替模拟这些子动力学系统。GMC 方法的积分器虽然也属于 SSI 框架，但却不适合所提动力学系统，因为它无法处理扩散过程。所以本节需要为所提动力学系统重新开发积分器。但本节会采用 GMC 所使用的流形嵌入技术以及测地线积分器技术，从而可以解除流形必须有全局坐标系的限制，并摆脱内循环。本节将首先分解两个动力学系统，然后将各子动力学系统的闭式解表示在嵌入空间中，最后根据 SSI 框架得出对应的算法。

参考 2.1.2.4 节，考虑黎曼流形 $\mathscr{M}$ 的 n 维欧氏等距嵌入空间 $\mathbb{R}^n$（$n \geqslant m$），并记其嵌入映射为 $\Xi : \mathscr{M} \to \mathbb{R}^n$。通过此映射，可将流形上的点（或视作坐标系）$Z$ 表示为嵌入空间中的点 $y := \Xi(Z)$，并将流形上的动量（余切向量）r 表示为嵌入空间中的动量 $s := \Xi_*(r)$（Ξ_* 表示在映射 Ξ 下的前推）。

SGGMC 的积分器　SGGMC 的动力学系统（式 (3-2)）可被分解为：

$$\text{A:}\ \begin{cases} \mathrm{d}Z = G^{-1} r \,\mathrm{d}t, \\ \mathrm{d}r = -\dfrac{1}{2}\nabla_Z \left[r^\top G^{-1} r \right] \mathrm{d}t, \end{cases} \tag{3-4a}$$

$$\text{B:}\ \begin{cases} \mathrm{d}Z = 0, \\ \mathrm{d}r = -J^\top C J G^{-1} r \,\mathrm{d}t, \end{cases} \tag{3-4b}$$

$$\text{O:}\ \begin{cases} \mathrm{d}Z = 0, \\ \mathrm{d}r = \nabla_Z \log p_L(Z)\,\mathrm{d}t - \dfrac{1}{2}\nabla_Z \log |G(Z)|\,\mathrm{d}t + \mathcal{N}(0, 2J^\top C J\,\mathrm{d}t). \end{cases} \tag{3-4c}$$

子动力学系统 A 的闭式解正是流形上的测地流（geodesic flow）（Abraham 等人[143]，定理 3.7.1），这与 GMC 方法的情况类似。形象地来说，子动力学系统 A（式 (3-4a)）描述了不受任何力作用而自由地在流形上运动的粒子，而这种运动在黎曼流形领域正是由测地流来描述的。在欧氏空间中，这样的运动就是匀速直线运动：

$$\text{（欧氏空间测地流）A: } y(t) = y(0) + s(0)\,t, \quad s(t) = s(0),$$

而在超球面 $\mathbb{S}^{n-1}$ 上，这样的运动是沿着超球面上大圆（过球心的超平面与超球面

的交线）作旋转的运动：

$$\text{（超球面测地流）A: } \begin{cases} y(t) = y(0)\cos(\|s(0)\|\,t) + \big(s(0)/\|s(0)\|\big)\sin(\|s(0)\|\,t), \\ s(t) = -\|s(0)\|\,y(0)\sin(\|s(0)\|\,t) + s(0)\cos(\|s(0)\|\,t). \end{cases}$$

由于这个形式是在超球面 $\mathbb{S}^{n-1}$ 的等距嵌入空间 $\mathbb{R}^n$ 中写出的，因而不需要考虑超球面的坐标系的情况，所以虽然它没有全局坐标系但仍然可以用此方法进行模拟。这样的使用流形的测地流进行动力学模拟的方法即为测地线积分器。

子动力学系统 B 和 O 就与 GMC 不同了。下面将首先把这两个子动力学系统表达在嵌入空间中，再在嵌入空间中求得闭式解。注意到在式 (3-4b) 和式 (3-4c) 中 Z 不随时间变化，因而对应的嵌入空间中的表示 y 也是常量，即 $\mathrm{d}y = 0$。为了得到动量的变化规律在嵌入空间中的表示，这里首先需要推导出一些必需的结论。首先，在待采样流形 $\mathscr{M}$ 的坐标系中，由我们在上一节 3.2.1 中的推导，可知动量 r 与速度 $\dot{Z} = \dfrac{\mathrm{d}Z}{\mathrm{d}t}$ 的关系：$r = G(Z)\dot{Z}$。类似地，由于本节所考虑的嵌入空间是欧氏的，即黎曼度量矩阵是单位矩阵，因而在嵌入空间中动量 s（即余切向量对偶在切空间中的切向量）与速度 $\dot{y} := \dfrac{\mathrm{d}y}{\mathrm{d}t}$ 的关系为 $s = \dot{y}$，而嵌入空间中的速度 $\dot{y}$ 与坐标空间中的速度 $\dot{Z}$ 的关系为：$\dot{y} = \dfrac{\mathrm{d}\Xi(Z)}{\mathrm{d}t} = \left(\sum\limits_{j=1}^{m} \dfrac{\partial \Xi^i(Z)}{\partial Z^j}\dfrac{\mathrm{d}Z^j}{\mathrm{d}t}\right)_{i=1}^{n} = \left(\sum\limits_{j=1}^{m} J_{ij}\dfrac{\mathrm{d}Z^j}{\mathrm{d}t}\right)_{i=1}^{n} = J\dot{Z}$。根据类似的推导，可以得到 $\nabla_Z = J^\top \nabla_y$。值得注意的是，这些已知的关系可以推得 $s = \dot{y} = J\dot{Z} = JG^{-1}r$，这就将两个空间中的动量联系了起来。这个结果与直接使用上文中所提到的嵌入空间中动量的定义 $s := \Xi_*(r)$ 得到的二者的关系是一样的。这正是等距嵌入的好处。等距嵌入的另外一个性质是 $G(Z) = J^\top J$。最后，与在流形 $\mathscr{M}$ 的坐标系中方便使用勒贝格测度不同，在嵌入空间中的流形 $\Xi(\mathscr{M})$ 上使用豪斯多夫测度（Hausdorff measure）会更加自然。这个测度是 $\mathbb{R}^n$ 中的勒贝格测度在 $\Xi(\mathscr{M})$ 上的限制。记目标分布在嵌入空间中关于豪斯多夫测度的密度函数为 $p(y)$，它与在坐标空间中关于勒贝格测度的密度函数 $p_L(Z)$ 的关系为：$p(y) = p_L(Z)/\sqrt{|G(Z)|}$（参见 2.1.2.5 节）。

根据这些知识便可得到子动力学系统 B 和 O 在嵌入空间中的形式。将 $r = G(Z)\dot{Z}$，$\nabla_Z = J^\top \nabla_y$ 及 $p(y) = p_L(Z)/\sqrt{|G(Z)|}$ 代入式 (3-4b) 和式 (3-4c) 中，并注意到 Z 是常量，可得：

$$\text{B: } \begin{cases} \mathrm{d}y = 0, \\ G(Z)\,\mathrm{d}\dot{Z} = -J^\top C J\dot{Z}\,\mathrm{d}t, \end{cases} \quad \text{O: } \begin{cases} \mathrm{d}y = 0, \\ G(Z)\,\mathrm{d}\dot{Z} = J^\top \nabla_y \log p(y)\,\mathrm{d}t + J^\top \mathcal{N}(0, 2C\,\mathrm{d}t). \end{cases}$$

在上面两个动力学系统的动量方程上左乘 $J(Z)G(Z)^{-1}$ 可得：

$$\text{B: } \begin{cases} \mathrm{d}y = 0, \\ \mathrm{d}(J\dot{Z}) = -JG(Z)^{-1}J^{\top}CJ\dot{Z}\,\mathrm{d}t, \end{cases} \quad \text{O: } \begin{cases} \mathrm{d}y = 0, \\ \mathrm{d}(J\dot{Z}) = JG(Z)^{-1}J^{\top}\nabla_y \log p(y)\,\mathrm{d}t, \\ \qquad\qquad + JG(Z)^{-1}J^{\top}\mathcal{N}(0, 2C\,\mathrm{d}t). \end{cases}$$

再利用 $s = J\dot{Z}$ 及 $G = J^{\top}J$，可得：

$$\text{B: } \begin{cases} \mathrm{d}y = 0, \\ \mathrm{d}s = -J(J^{\top}J)^{-1}J^{\top}Cs\,\mathrm{d}t, \end{cases} \quad \text{O: } \begin{cases} \mathrm{d}y = 0, \\ \mathrm{d}s = J(J^{\top}J)^{-1}J^{\top}\Big(\nabla_y \log p(y)\,\mathrm{d}t \\ \qquad\qquad + \mathcal{N}(0, 2C\,\mathrm{d}t)\Big). \end{cases} \tag{3-5}$$

最后，对于矩阵 $J(Z)$，考虑用变量 y 作为其自变量，即 $J(y) := J(\Xi^{-1}(y))$。至此本节推导出了子动力学系统 B 和 O 在嵌入空间中的表达形式。

此表达形式可以进一步简化为更直观且易于计算的形式。首先，$J(J^{\top}J)^{-1}J^{\top}$ 这一项是将 $\mathbb{R}^n$ 中的向量朝向矩阵 J 的列空间（column space）的正交投影。矩阵 J 的列空间是 J 的各列所张成的线性空间，它可表达为 $\mathrm{Col}(J) := \{\, J\theta \mid \theta \in \mathbb{R}^m \,\}$（注意 J 的维数是 $n \times m$ 且 $n \geqslant m$）。由于此处所考虑的矩阵 J 是满秩的，因此它的列空间 $\mathrm{Col}(J)$ 是 $\mathbb{R}^n$ 的 m 维线性子空间，因而对于 $\mathbb{R}^n$ 中的任意一点 y，都可以进行如下正交分解：

$$y = y_{\parallel} + y_{\perp}, \quad y_{\parallel} \in \mathrm{Col}(J), y_{\perp} \in \big(\mathrm{Col}(J)\big)^{\perp}, \tag{3-6}$$

其中 $\big(\mathrm{Col}(J)\big)^{\perp} := \big\{\, v \in \mathbb{R}^n \;\big|\; \forall u \in \mathrm{Col}(J), u^{\top}v = 0 \,\big\}$ 是 $\mathrm{Col}(J)$ 在 $\mathbb{R}^n$ 中的正交补空间。由线性代数的知识，这个正交分解是唯一的。下面将求得这个唯一的 $y_{\parallel}$ 关于 y 及 J 的表达式。将 $y_{\parallel}$ 表示为 $J\theta_y$，其中 $\theta_y \in \mathbb{R}^m$。因此 $y_{\perp} = y - J\theta_y$。由于 $y_{\perp} \in \big(\mathrm{Col}(J)\big)^{\perp}$，因而由正交补空间的定义及列空间的表达形式可知，$\forall \theta \in \mathbb{R}^m, (J\theta)^{\top}y_{\perp} = \theta^{\top}J^{\top}(y - J\theta_y) = 0$。由 θ 的任意性，可知 $J^{\top}(y - J\theta_y) = 0$，即

$$\begin{aligned} \theta_y &= (J^{\top}J)^{-1}J^{\top}y, \\ y_{\parallel} &= J\theta_y = J(J^{\top}J)^{-1}J^{\top}y. \end{aligned} \tag{3-7}$$

因此 $J(J^{\top}J)^{-1}J^{\top}$ 即为将 $\mathbb{R}^n$ 中向量 y 投影到 J 的列空间 $\mathrm{Col}(J)$ 上的正交投影。此处将此正交投影记为 Λ。

其次，正交投影 Λ 可以表示为另外一种更加方便的形式。定义 $n \times (n-m)$ 维矩阵 P 为 $\mathbb{R}^n$ 的 $n-m$ 维线性子空间 $\big(\mathrm{Col}(J)\big)^{\perp}$ 的一组标准正交基按照列的方式

排列起来所得的矩阵。由定义有 $P^\top P = I_{n-m}$。利用矩阵 P，可以将正交补空间 $\big(\mathrm{Col}(J)\big)^\perp$ 表示为 $\big\{\, P\vartheta \;\big|\; \vartheta \in \mathbb{R}^{n-m} \,\big\}$。以此观点来看 $\mathbb{R}^n$ 中的正交分解（式 (3-6)），可以将 $y_\perp$ 表示为 $P\vartheta_y$，其中 $\vartheta_y \in \mathbb{R}^{n-m}$。由于 $y_\parallel = y - P\vartheta_y$ 是 $\mathrm{Col}(J)$ 中的元素，因此它与 $\big(\mathrm{Col}(J)\big)^\perp$ 中的任意元素 $P\vartheta$ 都正交，亦即 $(P\vartheta)^\top(y - P\vartheta_y) = 0$。由 ϑ 的任意性，可得 $P^\top(y - P\vartheta_y) = 0$，即

$$\begin{aligned}\vartheta_y &= (P^\top P)^{-1}P^\top y = P^\top y,\\ y_\parallel &= y - P\vartheta_y = (I_n - PP^\top)y.\end{aligned}$$

对比式 (3-7)，可发现正交投影 Λ 也可以使用 $I_n - PP^\top$ 来表达。这个表达形式的方便性和计算经济性将会在后面考虑超球面这个实例的时候加以具体说明。

值得一提的是，$J(y)$ 的列空间 $\mathrm{Col}(J(y))$ 正是流形 $\mathscr{M}$ 在嵌入空间中的表示 $\Xi(\mathscr{M})$ 在 y 处的切空间 $T_y\Xi(\mathscr{M})$。这是因为，根据矩阵 J 的定义（式 (3-1)），矩阵 $J(y)$ 的第 j 列是 $\left(\dfrac{\partial \Xi^i(Z)}{\partial Z^j}\right)_{i=1}^m$，其中 $Z = \Xi^{-1}(y)$。而另一方面，流形 $\mathscr{M}$ 在 Z 处的切向量 ∂_{Z^j} 在 Ξ 下的前推（push-forward）是 $\Xi_*(\partial_{Z^j}) := \big(\partial_{Z^j}[y^i \circ \Xi]\big)\,\partial_{y^i} = \dfrac{\partial \Xi^i(Z)}{\partial Z^j}\partial_{y^i}$，因此矩阵 $J(y)$ 的第 j 列正好就是切向量 ∂_{Z^j} 的前推在嵌入空间的切空间基底 $\{\partial_{y^i}\}_{i=1}^n$ 下的坐标。由于切向量集 $\{\partial_{Z^j}\}_{j=1}^m$ 所张成的线性空间就是流形 $\mathscr{M}$ 的切空间 $T_Z\mathscr{M}$，因此它们的前推 $\left\{\dfrac{\partial \Xi^i(Z)}{\partial Z^j}\partial_{y^i}\right\}_{j=1}^m$ 所张成的线性空间，或者等价地说矩阵 $J(y)$ 的列空间 $\mathrm{Col}(J(y))$，就是切空间 $T_Z\mathscr{M}$ 的前推 $\Xi_*(T_Z\mathscr{M})$。进一步，由等距嵌入的性质，亦即原流形 $\mathscr{M}$ 的几何结构在等距嵌入空间 $\mathbb{R}^n$ 中的体现是与 $\Xi(\mathscr{M})$ 在欧氏空间 $\mathbb{R}^n$ 中的几何结构相吻合的，可以发现切空间 $T_{\Xi^{-1}(y)}\mathscr{M}$ 的前推空间 $\Xi_*(T_Z\mathscr{M})$ 就是 $\Xi(\mathscr{M})$ 在 $\mathbb{R}^n$ 中的切空间 $T_y\Xi(\mathscr{M})$。因此可以说，矩阵 $J(y)$ 的列空间 $\mathrm{Col}(J(y))$ 就是 $T_y\Xi(\mathscr{M})$，而 $\Lambda(y)$ 就是在 $\mathbb{R}^n$ 中向 $T_y\Xi(\mathscr{M})$ 的正交投影。

最后，本节将利用这些知识来求解子动力学系统 B 和 O。通过将 $J(J^\top J)^{-1}J^\top$ 表达为 Λ，可将式 (3-5) 写为：

$$\text{B: }\begin{cases}\mathrm{d}y = 0,\\ \mathrm{d}s = \Lambda(y)Cs\,\mathrm{d}t,\end{cases}\qquad \text{O: }\begin{cases}\mathrm{d}y = 0,\\ \mathrm{d}s = \Lambda(y)\Big(\nabla_y \log p(y)\,\mathrm{d}t + \mathcal{N}(0, 2C\,\mathrm{d}t)\Big).\end{cases}$$

注意到这两个动力学系统中 y 都是常量，可得到它们的闭式解为：

$$\text{B: } \begin{cases} y(t) = y(0), \\ s(t) = \operatorname{expm}\big(-\Lambda\big(y(0)\big)Ct\big)\, s(0), \end{cases} \quad \text{O: } \begin{cases} y(t) = y(0), \\ s(t) = s(0) + \Lambda\big(y(0)\big)\Big(\nabla_y \log p\big(y(0)\big)t \\ \qquad\qquad + \mathcal{N}(0, 2Ct)\Big), \end{cases} \tag{3-8}$$

其中 $\operatorname{expm}(J) := \sum_{i=0}^{\infty} \frac{J^i}{i!}$ 表示矩阵的指数映射（exponential map）。对于 C 是标量的情况，可以进一步化简子动力学系统 B：

$$s(t) = \operatorname{expm}\big(-\Lambda\big(y(0)\big)Ct\big)s(0) = \sum_{i=0}^{\infty} \frac{\big(-\Lambda\big(y(0)\big)Ct\big)^i s(0)}{i!} = \sum_{i=0}^{\infty} \frac{(-Ct)^i \Lambda^i\big(y(0)\big)s(0)}{i!}.$$

而由于 $s(0) = \dot{y}(0)$ 已经在切空间 $T_{y(0)}\Xi(\mathscr{M})$ 中了，因此无论经 $\Lambda\big(y(0)\big)$ 投影多少次它还是不变，亦即对于任意自然数 i，$\Lambda^i\big(y(0)\big)s(0) = s(0)$。因此子动力学系统 B 可表示为：

$$s(t) = \sum_{i=0}^{\infty} \frac{(-Ct)^i s(0)}{i!} = s(0) \sum_{i=0}^{\infty} \frac{(-Ct)^i}{i!} = \exp(-Ct)s(0).$$

为更进一步说明子动力学系统 B 的行为，可对无穷小的 t 进行展开至一阶，得到：

$$s(t) \approx (1 - Ct)s(0).$$

这个正是 SGHMC 方法中为控制梯度噪声带来的影响而在动力学系统中添加的摩擦力项。本节的推导发现这个摩擦力项可以推广到流形上，只不过需要通过等距嵌入空间中的动量 s 来体现，而不是通常的坐标空间中的动量 r。对于 SGHMC 方法来说这两者是一样的，但对本节所考虑的情况这两者则会不同，例如 s 只能取在切空间内，而 r 则可以在坐标空间 $\mathbb{R}^m$ 的一个开子集内任意选取。

下面分析超球面这个具体的例子，通过两种方式得到正交投影 $\Lambda(y)$，来展示使用矩阵 P 来表达 $\Lambda(y)$（式 (3-8)）的自然与简洁。由于超球面 $\mathbb{S}^{n-1} := \{\, y \in \mathbb{R}^n \mid \|y\| = 1 \,\}$ 本身就在 $\mathbb{R}^n$ 中定义，且其结构也由 $\mathbb{R}^n$ 处继承，因此 $\mathbb{R}^n$ 就是 $\mathbb{S}^{n-1}$ 的等距嵌入空间，且等距嵌入映射为 $\Xi : \mathbb{S}^{n-1} \to \mathbb{R}^n, y \mapsto y$。

首先考虑通过矩阵 J 来计算正交投影 $\Lambda = J(J^\top J)^{-1}J^\top$。由于 J 是与坐标系的选取有关，需要先为超球面 $\mathbb{S}^{n-1}$ 选择一个局部坐标系。考虑它的上半球面 $\big(\mathbb{S}^{n-1}\big)^+ := \big\{\, y \in \mathbb{S}^{n-1} \,\big|\, y^n > 0 \,\big\}$ 这个局部。它可以被映射到 $\mathbb{R}^{n-1}$ 上，例如通过如下映射：

$$\Phi : \big(\mathbb{S}^{n-1}\big)^+ \to \mathbb{R}^{n-1}, (y^1, \cdots, y^{n-1}, y^n)^\top \mapsto (y^1, \cdots, y^{n-1})^\top.$$

此映射（在其像集上）可逆，其逆映射为：

$$\Phi^{-1}(Z^1,\cdots,Z^{n-1}) = \big(Z^1,\cdots,Z^{n-1},y^n(Z)\big)^\top,$$

其中 $y^n(Z) := \sqrt{1-\sum_{i=1}^{n-1}(Z^i)^2}$。易知此映射及其逆映射都是连续的，即此映射是一个同胚（homeomorphism），因而 $\big(\big(\mathbb{S}^{n-1}\big)^+,\Phi\big)$ 是超球面 $\mathbb{S}^{n-1}$ 的一个局部坐标系。在此坐标系下，由定义（式 (3-1)）可知矩阵 J 为：

$$J(Z) = \begin{pmatrix} I_{n-1} \\ -Z^\top / y^n(Z) \end{pmatrix},$$

且

$$J(Z)^\top J(Z) = I_{n-1} + \frac{ZZ^\top}{\big(y^n(Z)\big)^2}.$$

由谢尔曼-莫里森公式（Sherman-Morrison formula）[144] 可得：

$$\big(J(Z)^\top J(Z)\big)^{-1} = I_{n-1} - ZZ^\top.$$

根据由矩阵 J 所表达的正交投影 Λ（式 (3-7)），可得

$$\Lambda(Z) = J(J^\top J)^{-1}J^\top = \begin{pmatrix} I_{n-1}-ZZ^\top & -y^n(Z)\,Z \\ -y^n(Z)\,Z^\top & 1-\big(y^n(Z)\big)^2 \end{pmatrix} = I_n - \Phi^{-1}(Z)\,\Phi^{-1}(Z)^\top,$$

而若用 $y=\Xi(\Phi^{-1}(Z))$ 表达，则可得嵌入空间中的表达式：

$$\Lambda(y) = I_n - yy^\top, \quad y\in\Xi\big(\big(\mathbb{S}^{n-1}\big)^+\big).$$

由于在嵌入空间中的正交投影与流形 $\mathscr{M}$ 的局部坐标系无关，因此上式对于整个流形 $\mathbb{S}^{n-1}$ 都成立。

然后考虑利用正交补空间的标准正交基矩阵 P 来直接在嵌入空间中计算正交投影 $\Lambda(y)$。对于等距嵌入在 $\mathbb{R}^n$ 中的球面 $\mathbb{S}^{n-1}$ 来说，其在嵌入空间中的表示 $\Xi(\mathbb{S}^{n-1})$ 就是它自身 $\mathbb{S}^{n-1}$。因此它在点 y 处的切空间 $T_y\mathbb{S}^{n-1}$ 就是一个在点 y 处与球面 $\mathbb{S}^{n-1}$ 相切的 $n-1$ 维超平面。此平面垂直于向量 y，因此这个切空间的一维正交补空间就是向量 y 所张成的线性空间，即沿着向量 y 方向的直线。由于 y 本身就是归一的，因此 y 就是这个空间的标准正交基。根据由矩阵 P 所表达的正交投影 Λ（式 (3-8)），立即可得：

$$\Lambda(y) = I_n - yy^\top.$$

可见通过这种方式可以更加方便直观地得到正交投影 Λ 的表达式。

gSGNHT 的积分器　将 gSGNHT 的动力学系统（式 (3-3)）以类似的方式分解为：

$$\text{A:}\begin{cases}\mathrm{d}Z = G^{-1}r\,\mathrm{d}t,\\ \mathrm{d}r = -\dfrac{1}{2}\nabla_Z\left[r^\top G^{-1}r\right]\mathrm{d}t,\\ \mathrm{d}\xi = \left(\dfrac{1}{m}r^\top G^{-1}r - 1\right)\mathrm{d}t,\end{cases} \tag{3-9a}$$

$$\text{B:}\begin{cases}\mathrm{d}Z = 0,\\ \mathrm{d}r = -\xi r\,\mathrm{d}t,\\ \mathrm{d}\xi = 0,\end{cases} \tag{3-9b}$$

$$\text{O:}\begin{cases}\mathrm{d}Z = 0,\\ \mathrm{d}r = \nabla_Z \log p_L\,\mathrm{d}t - \dfrac{1}{2}\nabla_Z \log|G|\,\mathrm{d}t + \mathcal{N}(0, 2C G\,\mathrm{d}t),\\ \mathrm{d}\xi = 0.\end{cases} \tag{3-9c}$$

对于子动力学系统 A（式 (3-9a)），Z 与 r 的解与 SGGMC 相同，即测地流。对这个动力学系统的模拟即为测地线积分器。关于恒温器变量 ξ 的求解，首先可从式 (3-9a) 中 Z 与 r 的方程推出 $r^\top G^{-1}r$ 是常量：

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[r^\top G(Z)^{-1}r\right] = \nabla_Z\left[r^\top G(Z)^{-1}r\right]^\top \dot{Z} + 2\left[G(Z)^{-1}r\right]^\top \dot{r} = -2\dot{r}^\top\dot{Z} + 2\dot{Z}^\top\dot{r} = 0.$$

事实上，对于等距嵌入，有如下关系：

$$r^\top G^{-1}r = (G^{-1}r)^\top G(G^{-1}r) = \dot{Z}^\top(J^\top J)\dot{Z} = (J\dot{Z})^\top(J\dot{Z}) = s^\top s,$$

而 $\dfrac{1}{2}s^\top s$ 正是这个动力学系统中的动能，在子动力学系统 A 亦即不受外力的自由运动中是守恒的，因而这也能说明 $r^\top G^{-1}r$ 是常量。基于此，恒温器变量 ξ 的解为：

$$\xi(t) = \xi(0) + \left(\frac{1}{m}s(0)^\top s(0) - 1\right).$$

这个形式与 SGNHT 方法中 ξ 的动力学系统是对应的。而本节工作发现使用等距嵌入空间中的动量 s 可将这个形式推广到黎曼流形上。

子动力学系统 B（式 (3-9b)）可用类似上述 SGGMC 的方式求解：

$$s(t) = \exp\big(-\xi(0)t\big)\,s(0).$$

它对无穷小时间 t 的一阶展开复现了 SGNHT 中自适应地平衡摩擦力项和梯度噪声的过程，而本节工作将这个过程推广到了黎曼流形上。子动力学系统 O（式 (3-9c)）的解与 SGGMC 的相同。这两个子动力学系统中 ξ 都是常量，即 $\xi(t) = \xi(0)$。

算法 1 SGGMC 方法的采样过程

1: 随机初始化 $y^{(0)} \in \Xi(\mathscr{M})$；从标准高斯中采取 $s \sim \mathcal{N}(0, I)$ 并做投影 $s^{(0)} \leftarrow \Lambda(y^{(0)})s$；
2: **对** $k = 1, 2, \cdots,$ **执行操作:**
3: 　A：使用测地流为 (y, s) 赋值：$(y, s) \leftarrow \text{GeodFlow}_{\frac{\varepsilon_k}{2}}(y^{(k-1)}, s^{(k-1)})$；
4: 　B：$s \leftarrow \exp\left(-C\frac{\varepsilon_k}{2}\right)s$；
5: 　O：从原数据集 $\mathscr{D}$ 中随机采取一个固定大小的子数据集 $\widetilde{\mathscr{D}}$ 来估计随机梯度 $\tilde{\nabla}_y \log p(y)$，并作更新：
　　$s \leftarrow s + \Lambda(y)\left[\tilde{\nabla}_y \log p(y)\varepsilon_k + \mathcal{N}\big(0, (2C - \varepsilon_k\Sigma(y))\varepsilon_k\big)\right]$；
6: 　B：$s \leftarrow \exp\left(-C\frac{\varepsilon_k}{2}\right)s$；
7: 　A：使用测地流赋值：$(y^{(k)}, s^{(k)}) \leftarrow \text{GeodFlow}_{\frac{\varepsilon_k}{2}}(y, s)$。无 MH 测试。
8: **结束**

最终算法　现在考虑使用随机梯度进行模拟的情况。随机梯度只会影响两个动力学系统共有的子动力学系统 O。参照式 (1-1)，随机梯度可以表示为：

$$\tilde{\nabla}_y \log p(y) = \nabla_y \log p(y) + \mathcal{N}\big(0, \Sigma(y)\big),$$

其中 $\Sigma(y)$ 是随机梯度噪声的协方差矩阵。基于此，可以将子动力学系统 O 的解表示为：

$$s(t) = s(0) + \Lambda\big(y(0)\big)\left[-\tilde{\nabla}_y \log p\big(y(0)\big)t + \mathcal{N}\big(0, 2Ct - \Sigma\big(y(0)\big)\,t^2\big)\right]. \tag{3-10}$$

其中随机梯度噪声的协方差矩阵 $\Sigma(y)$ 可通过参照 Ahn 等人[145] 的做法，估计为经验费舍尔信息矩阵（empirical Fisher information matrix）。不过更加实用的方法是直接将它估计为零。这样做的理由在于，进行动力学系统的离散时间模拟时，所取的时间步长（step size）t 都是很小的，因而相较于动力学系统带来的随机扩散噪声的协方差 $2Ct$，随机梯度的噪声的协方差 Σt^2 是它的高阶小量，所以可以忽略。这个做法也由 Chen 等人[113] 的工作所支持。

最后，根据 SSI 框架，整个动力学系统的模拟可以通过交替地以 “ABOBA” 的模式分别用闭式解模拟每个子动力学系统。具体来说，对于选定的时间步长 ε，整个动力学系统的一步模拟包括先将子动力学系统 A 和 B 向前模拟 $\varepsilon/2$ 时间，再将子动力学系统 O 向前模拟 ε 时间，最后再将子动力学系统 B 和 A 向前模拟 $\varepsilon/2$ 时间。与其他的随机梯度 MCMC 方法类似，这里也为所提方法省略梅特罗波利斯-海斯廷斯取舍测试（Metropolis-Hastings rejection test，MH 测试），因为动力学系统可保证准确采样，而动力学系统的模拟的误差可以得到控制[113]。此外，每次 MH 测

算法 2 gSGNHT 方法的采样过程

1: 随机初始化 $y^{(0)} \in \Xi(\mathscr{M})$；从标准高斯中采取 $s \sim \mathcal{N}(0, I)$ 并做投影 $s^{(0)} \leftarrow \Lambda(y^{(0)})s$；做赋值 $\xi^{(0)} \leftarrow C$；

2: **对** $k = 1, 2, \cdots,$ **执行操作:**

3: A：使用测地流为 (y, s) 赋值：$(y, s) \leftarrow \text{GeodFlow}_{\frac{\varepsilon_k}{2}}(y^{(k-1)}, s^{(k-1)})$；做赋值 $\xi \leftarrow \xi^{(k)} + \left(\frac{1}{m} s^{(k-1)\top} s^{(k-1)} - 1\right) \frac{\varepsilon_k}{2}$；

4: B：$s \leftarrow \exp\left(-\xi \frac{\varepsilon_k}{2}\right) s$；

5: O：从原数据集 $\mathscr{D}$ 中随机采取一个固定大小的子数据集 $\widetilde{\mathscr{D}}$ 来估计随机梯度 $\tilde{\nabla}_y \log p(y)$，并作更新：
$s \leftarrow s + \Lambda(y)\left[\tilde{\nabla}_y \log p(y)\varepsilon_k + \mathcal{N}\big(0, (2C - \varepsilon_k \Sigma(y))\varepsilon_k\big)\right]$；

6: B：$s \leftarrow \exp\left(-\xi \frac{\varepsilon_k}{2}\right) s$；

7: A：使用测地流赋值：$(y^{(k)}, s^{(k)}) \leftarrow \text{GeodFlow}_{\frac{\varepsilon_k}{2}}(y, s)$；做赋值 $\xi^{(k)} \leftarrow \xi + \left(\frac{1}{m} s^{\top} s - 1\right) \frac{\varepsilon_k}{2}$。无 MH 测试。

8: **结束**

试需要遍历整个数据集，因而会抵消在动力学系统的模拟中使用随机梯度的好处，也就是说使用 MH 测试会丢失可扩展性。SGGMC 和 gSGNHT 方法的算法步骤分别列于算法 1 和算法 2 中（针对 C 选为标量的情况）。

对于 SGGMC 和 gSGNHT 方法，步长系列（step size scheme）$\{\varepsilon_k\}$ 推荐选择为一个固定的数字 ε。尽管一个缩减的步长系列（例如 Chen 等人[113] 中所提到的 $\varepsilon_k \propto k^{-\lambda}$ 系列，其中参数 $\lambda \in (0, 1)$）可以有更多的理论保证（例如即使随机梯度不是无偏的情况下，所采样本的均值也可以是渐进无偏的[113]），但这些好处可能在实际中很难体现出来，反而后期过小的步长会使方法在实际中收敛变慢。

算法中的参数，即步长 ε 和标量参数 C 可通过与 SGHMC[115] 类似的方式进行选择。通过引入单批学习率（per-batch learning rate）b 和动量系数（momentum coefficient）ς 这两个标量参数，可以进行如下设定：$\varepsilon = \sqrt{b/|\mathscr{D}|}$ and $C = \varsigma/\varepsilon$。通常 b 和 ς 在 0.1 和 0.01 附近取值。

3.3　球面混合模型的高效后验推理算法

本节考虑将所提的 SGGMC 和 gSGNHT 方法用于富有挑战性的球面混合模型（spherical admixture model, SAM）[22] 的后验推理任务上。SAM 模型是一个结构与隐式狄利克雷分配模型（latent Dirichlet allocation, LDA）[19] 类似的层次化贝叶斯

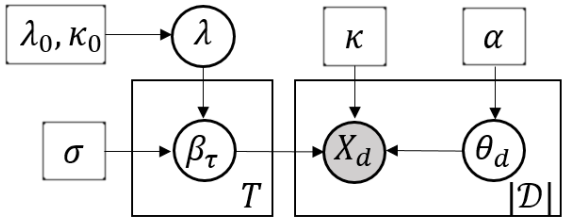


图 3.1　SAM 模型的结构。

话题模型，但它能够处理超球面数据（即每个数据点都在超球面 $\mathbb{S}^{n-1}$ 上的数据），例如由归一化词频-逆篇频（term frequency-inverse document frequency, tf-idf）特征所表示的文档数据。它使得层次化贝叶斯模型能够处理更加丰富的数据特征形式，并且与 LDA 模型相比，它能够直接地为单词的缺失而建模因而能够学到数据更有代表性的特征。

SAM 模型使用冯 · 米塞斯-费舍尔分布（von Mises-Fisher distribution, vMF）[146] 来为超球面上的随机变量建模。vMF 分布是超球面 $\mathbb{S}^{n-1}$ 上的单峰最大熵分布。它有两个参数：均值 $\lambda \in \mathbb{S}^{n-1}$ 和集中度 $\kappa \in \mathbb{R}^+$。它在超球面 $\mathbb{S}^{n-1}$ 的等距嵌入空间 $\mathbb{R}^n$ 里关于豪斯多夫测度的概率密度函数为：

$$\text{vMF}(y|\lambda,\kappa) = c_n(\kappa)\exp(\kappa\lambda^\top y), y \in \mathbb{S}^{n-1},$$

其中归一化系数 $c_n(\kappa) := \kappa^{n/2-1}/\big((2\pi)^{n/2}\mathcal{I}_{n/2-1}(\kappa)\big)$，而 $\mathcal{I}_k(\cdot)$ 是 k 阶第一类修正贝塞尔函数（modified Bessel function of the first kind in order k）。由于 vMF 分布的使用，本章之后部分的内容中所涉及的概率密度函数都是关于嵌入空间中的豪斯多夫测度而定义的，因而可以直接使用算法 1 和算法 2。

SAM 模型的结构如图 3.1 所示。模型中观测数据是 $X = \{X_d\}_{d=1}^{|\mathscr{D}|}$，其中每一个数据点 X_d 都是超球面 $\mathbb{S}^{n-1}$ 上的点。模型的隐变量包括话题变量（topic）$\beta = \{\beta_\tau\}_{\tau=1}^{T}(\beta_\tau \in \mathbb{S}^{n-1})$, 文档的话题配比变量（topic proportion）$\theta = \{\theta_d\}_{d=1}^{|\mathscr{D}|}$（$\theta_d$ 是 $T-1$ 维单纯形上的点），和整个数据集的均值 $\lambda \in \mathbb{S}^{n-1}$。模型超参数为 $(\lambda_0, \kappa_0, \sigma, \alpha, \kappa)$。模型的生成过程为：

- 采取数据集均值 $\lambda \sim \text{vMF}(\lambda|\lambda_0, \kappa_0)$；
- 对于每一个 $\tau = 1, \cdots, T$，采取话题 $\beta_\tau \sim \text{vMF}(\beta_\tau|\lambda, \sigma)$；
- 对于每一个 $d = 1, \cdots, |\mathscr{D}|$，采取话题配比 $\theta_d \sim \text{Dir}(\theta_d|\alpha)$ 和数据点 $X_d \sim \text{vMF}(X_d|\bar{X}(\beta, \theta_d), \kappa)$，其中 $\bar{X}(\beta, \theta_d) := \dfrac{\beta\theta_d}{\|\beta\theta_d\|}$（$\beta$ 为各话题按列排列起来的矩阵 $(\beta_1, \cdots, \beta_T)$）是各话题在球面上的近似加权平均。

依照此生成过程，可以写出数据和隐变量的联合分布：

$$p(X,\lambda,\beta,\theta)=\text{vMF}(\lambda|\lambda_0,\kappa_0)\prod_{\tau=1}^{T}\text{vMF}(\beta_\tau|\lambda,\sigma)\prod_{d=1}^{|\mathscr{D}|}\text{Dir}(\theta_d|\alpha)\text{vMF}(X_d|\bar{X}(\beta,\theta_d),\kappa).$$

对 SAM 模型进行贝叶斯推理即是估计话题隐变量的后验分布，即 $p(\beta|X)$。由于它的闭形式无法求得，人们需要使用各种近似方法对它进行估计。提出此模型的原论文[22] 中作者给出了一个变分推理方法（variational inference, VI），其中超球面 $\mathbb{S}^{n-1}$ 的限制是通过反复地对 $\mathbb{R}^n$ 中的向量进行归一化进行的。但这个方法基于平均场（mean-field）假设，这严重限制了它的近似能力和近似精度。考虑 MCMC 方法可以得到渐进准确的近似，但超球面 $\mathbb{S}^{n-1}$ 的限制为 MCMC 方法的实现带来了挑战，例如上述归一化的方法就很难直接用于 MCMC 的模拟中，因为归一化的操作可能会改变马尔可夫链的平稳分布。提出此模型的原论文[22] 中作者也尝试了一个简单的 MCMC 方法，但那个方法是基于随机游走 MH 测试的方法。由于缺少有针对性的动力学系统的引导，这个方法产生的样本的自相关性会非常高，因此收敛得极慢。实际中的表现也很不如意，甚至没有 VI 的结果好。之后便没有人再尝试使用 MCMC 方法解决 SAM 模型的后验推理问题了。由于后验分布定义在超球面这个没有全局坐标系的流形上，因而大多数的黎曼流形采样方法都很难适用于此任务，包括前面提到的 SGRLD 和 SGRHMC。目前只有 CHMC 和 GMC 两个方法可以用来解决这个任务，但它们都不是可扩展的方法。而所提的 SGGMC 和 gSGNHT 方法不仅可以适用于解决此任务，还具有可扩展性的优势来高效处理大规模数据集。

现在展示如何使用 SGGMC 和 gSGNHT 方法来直接从后验分布 $p(\beta|X)$ 中来采样。首先注意到，数据集均值 λ 这个隐变量可以解析地被积分掉。具体来说，考虑将 $p(X,\lambda,\beta,\theta)$ 中与 λ 有关的部分对 λ 进行积分：

$$\begin{aligned}&\int_{\mathbb{S}^{n-1}}\text{vMF}(\lambda|\lambda_0,\kappa_0)\prod_{\tau=1}^{T}\text{vMF}(\beta_\tau|\lambda,\sigma)\,\mathrm{d}\lambda\\&=c_n(\kappa_0)c_n(\sigma)^T\int_{\mathbb{S}^{n-1}}\exp\left(\bar{\lambda}(\beta)^\top\lambda\right)\,\mathrm{d}\lambda=c_n(\kappa_0)c_n(\sigma)^Tc_n(\left\|\bar{\lambda}(\beta)\right\|)^{-1},\end{aligned}$$

其中 $\bar{\lambda}(\beta):=\kappa_0\lambda_0+\sigma\sum_{\tau=1}^{T}\beta_\tau$。因此将 $p(X,\lambda,\beta,\theta)$ 对 λ 进行积分，得到剩下的变量 (X,β,θ) 的联合分布为：

$$p(X,\beta,\theta)=c_n(\kappa_0)c_n(\sigma)^Tc_n(\left\|\bar{\lambda}(\beta)\right\|)^{-1}\prod_{d=1}^{|\mathscr{D}|}\text{Dir}(\theta_d|\alpha)\text{vMF}(X_d|\bar{X}(\beta,\theta_d),\kappa).\quad(3\text{-}11)$$

要从目标分布 $p(\beta|X)$ 中采样，SGGMC 和 gSGNHT 只需要知道梯度 $\nabla_\beta \log p(\beta|X)$ 的一个随机估计即可。然而这仍然需要将局部隐变量 θ 积分掉。这个积分无法解析地计算，但所需要的梯度却可以通过 Du 等人[147] 所开发的双重随机梯度技术（doubly-stochastic gradient），使用 θ 的样本将它积分掉。具体来说，注意到所需梯度可以写为：

$$\begin{aligned}
&\nabla_\beta \log p(\beta|X) = \frac{1}{p(\beta|X)} \nabla_\beta \int p(\beta,\theta|X)\,\mathrm{d}\theta \\
&= \int \frac{\nabla_\beta p(\beta,\theta|X)}{p(\beta|X)}\,\mathrm{d}\theta = \int \frac{p(\beta,\theta|X)}{p(\beta|X)} \frac{\nabla_\beta p(\beta,\theta|X)}{p(\beta,\theta|X)}\,\mathrm{d}\theta \\
&= \mathbb{E}_{p(\theta|\beta,X)} \left[\nabla_\beta \log p(\beta,\theta|X)\right], \qquad (3\text{-}12)
\end{aligned}$$

其中 $\nabla_\beta \log p(\beta,\theta|X) = \nabla_\beta \log p(X,\beta,\theta)$ 是已知的（参见式 (3-11)），而对 $p(\theta|\beta,X)$ 的期望则可通过此分布的样本 $\{\theta^{(l)}\}_{l=1}^L$ 来估计：$\nabla_\beta \log p(\beta|X) \approx \frac{1}{L}\sum_{l=1}^{L} \nabla_\beta \log p(X,\beta,\theta^{(l)})$。由于样本 $\{\theta^{(l)}\}_{l=1}^L$ 处在 $T-1$ 维单纯形上，因而可以使用 GMC[44] 从它们的目标分布 $p(\theta|\beta,X)$ 中采样（单纯形上的测地流和正交投影参见 Byrne 等人[44] 的附录 A），这需要知道 $p(\theta|\beta,X)$ 的一个未归一化密度函数即可。由于 $p(\theta|\beta,X) \propto p(X,\beta,\theta)$（看作关于 θ 的函数）且 $p(X,\beta,\theta)$ 已知（参见式 (3-11)），因而这种估计方式是可行的。

最后，为了实现可扩展性，考虑在每次需要估计梯度时随机选取原大规模数据集 $\mathscr{D}$ 的一个小的子数据集 $|\widetilde{\mathscr{D}}|$（其大小 $|\widetilde{\mathscr{D}}|$ 是一个给定的固定值）并在它上面计算双重随机梯度。具体来说，如果将 $\widetilde{\mathscr{D}}$ 用被选中的数据点在原数据集 $\mathscr{D}$ 中的编号来表示，那么梯度的估计为：

$$\tilde{\nabla}_\beta \log p(\beta|X) = -\nabla_\beta \log c_n(\left\|\bar{\lambda}(\beta)\right\|) + \kappa \frac{|\mathscr{D}|}{L|\widetilde{\mathscr{D}}|} \sum_{l=1}^{L} \sum_{d\in\widetilde{\mathscr{D}}} X_d^\top \bar{X}(\beta,\theta_d^{(l)}). \qquad (3\text{-}13)$$

有了这个估计方法，就可以使用 SGGMC 和 gSGNHT 进行 SAM 模型的后验推理了。具体的算法步骤参见算法 3。

3.4　实验

此部分展示所提的 SGGMC 和 gSGNHT 方法在实际问题中准确性和效率方面的优势，包括在合成数据（synthetic data）和真实数据上的实验表现。合成数据上的实验中只考察 SGGMC 方法，因为 gSGNHT 方法的使用恒温器变量的优势已经在 Ding 等人[119] 的工作中展现。真实数据上的实验则会展示两种方法的优势。

算法 3 使用 SGGMC/gSGNHT 的 SAM 模型后验推理

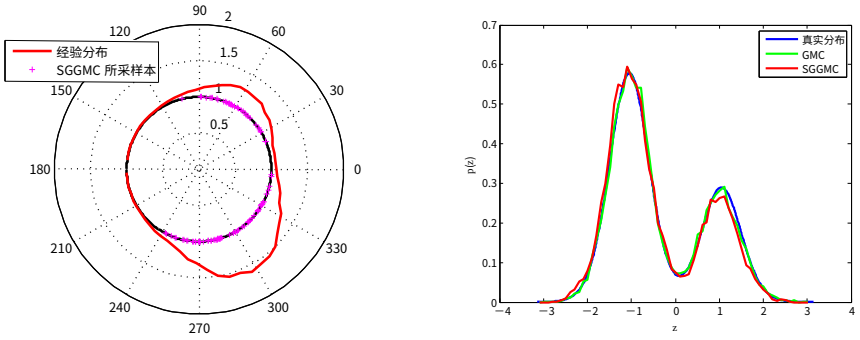
随机初始化话题变量 $\beta^{(0)}$。
对 $i = 1, 2, \cdots$ **执行操作:**
　从整个数据集 $\mathscr{D}$ 随机选取一个子数据集 $\widetilde{\mathscr{D}}$（以数据点在 $\mathscr{D}$ 中的编号表示）;
　对 $d \in \widetilde{\mathscr{D}}$ **执行操作:**
　　使用 GMC 方法从 $p(\theta_d|\beta^{(i-1)}, X_d)$ 中采取 L 个样本 $\{\theta_d^{(l)}\}_{l=1}^{L}$;
　结束
　使用 SGGMC（算法 1）或 gSGNHT（算法 2）从 $p(\beta|X)$ 中采取一个样本 $\beta^{(i)}$，其中随机梯度由式 (3-13) 计算。
结束

3.4.1　简单模拟实验

本节首先通过一个梯度噪声已知的理想情景实验来展示 SGGMC 的正确性和可用性。为方便展示结果，本节考虑从嵌入在 $\mathbb{R}^2$ 中的圆环（即一维球 $\mathbb{S}^1$）上进行采样的任务。实验中所选择的目标分布为 $p(Z) \propto \exp(5\theta_1^\top Z) + 2\exp(5\theta_2^\top Z)$，其中 $Z, \theta_1, \theta_2 \in \mathbb{S}^1$，且 $\theta_1 = -\theta_2 = \dfrac{\pi}{3}$（用与 $+x$ 轴方向的夹角表示）。用于 SGGMC 方法的随机梯度是通过给真实梯度加上高斯噪声 $\mathcal{N}(0, 1000I)$ 来制造的，并且在采样过程中使用这个已知噪声的方差作为式 (3-10) 中的随机梯度噪声协方差矩阵 Σ。GMC 和 SGGMC 方法均使用步长 $\varepsilon = 0.01$，而 SGGMC 方法使用动量系数 $\varsigma = 0.1$（参见 3.2.2 节最后）。角度空间（$\mathbb{S}^1$ 的一个局部坐标系）中的经验分布通过直方图（histogram）的方式来绘制，并取桶宽（bin size）为 0.1。

实验结果展示于图 3.2 中。左图在嵌入空间 $\mathbb{R}^2$ 中展示了 SGGMC 所采的 100 个样本以及它所采的 10,000 个样本的经验分布（empirical distribution），而右图在角度空间（$\mathbb{S}^1$ 的一个局部坐标系）中展示了 GMC 和 SGGMC 方法所采的 10,000 个样本的经验分布以及与真实分布的对比。从这些结果中可以发现，虽然梯度中有了噪声，但通过合适的动力学系统，SGGMC 仍然可以正确采样。

值得强调的是，尽管这个采样任务也可以通过在球坐标系（几乎是全局坐标系）中使用如 SGRLD 这样的可扩展的黎曼流形采样方法来完成，但这个做法过于繁杂，因为这种情况下需要计算坐标系中诸如黎曼度量矩阵等量，并且在坐标系的边界上必须要考虑一些特殊的操作，例如反弹等。数值不稳定性也会出现。这些复杂的情况会随着维度的增长而变得越来越明显。而所提方法是在嵌入空间中进行的，因此上述问题都可以得到避免，且能够优雅地拓展到高维情况下。



(a) 嵌入空间中 SGGMC 所采的样本　　(b) 角度空间中各分布的比较

图 3.2　简单模拟实验结果：(a) SGGMC 所采的样本及经验分布；(b) 经验分布与真实分布的比较。

3.4.2　合成数据实验

本节接下来将在一个简单的贝叶斯后验推理任务上考察 SGGMC 的正确性。这里考虑 Welling 等人[60] 在合成数据实验中所使用的模型的球面版本，即一维球面 $\mathbb{S}^1$ 上 vMF 分布的等权混合模型：

$$
\begin{aligned}
&p(Z_1) = \text{vMF}(Z_1|e_1,\kappa_1), \quad p(Z_2) = \text{vMF}(Z_2|e_1,\kappa_2),\\
&p(X|Z_1,Z_2) \propto \text{vMF}(X|Z_1,\kappa_X) + \text{vMF}(X|\lambda,\kappa_X),
\end{aligned}
$$

其中单位向量 $e_1 = (1,0)$，混合均值 $\lambda := \dfrac{Z_1+Z_2}{\|Z_1+Z_2\|}$。此模型的后验推理任务即为给定了数据集 $\mathscr{D} = \{X_i\}_{i=1}^{|\mathscr{D}|}$ 之后近似隐变量的后验分布 $p(Z_1,Z_2|\mathscr{D})$，其未归一化密度函数为：

$$
\begin{aligned}
&p(Z_1,Z_2|\mathscr{D}) \propto p(Z_1,Z_2,\mathscr{D})\\
\propto\ &\exp(\kappa_1 e_1^\top Z_1 + \kappa_2 e_1^\top Z_2)\prod_{i=1}^{|\mathscr{D}|}\left(\exp(\kappa_X Z_1^\top X_i) + \exp(\kappa_X \lambda (Z_1,Z_2)^\top X_i)\right).
\end{aligned}
$$

实验中，在模型方面，所选定的模型参数为 $\kappa_1 = \kappa_2 = \kappa_X = 20$，并使用 GMC 方法从似然分布 $p(X|Z_1 = Z_1^{(g)}, Z_2 = Z_2^{(g)})$ 中采取 100 个样本作为合成数据集 $\mathscr{D}$，其中用于生成数据的隐变量的固定值为 $Z_1^{(g)} = -\dfrac{\pi}{24}, Z_2^{(g)} = \dfrac{\pi}{8}$（以与 $+x$ 轴方向的夹角表示；上标 “(g)” 表示 “generate”，即 “生成”）。在采样推理算法方面，GMC 和 SGGMC 方法各自在 15,000 轮预热采样（burn-in）之后采取 25,000 个样本作为它们各自对后验分布的近似。GMC 方法使用步长 $\varepsilon = 1\times 10^{-3}$。SGGMC 方法参照 Ahn 等人[145] 的方法，使用经验费舍尔信息矩阵来估计式 (3-10) 中的梯度噪声

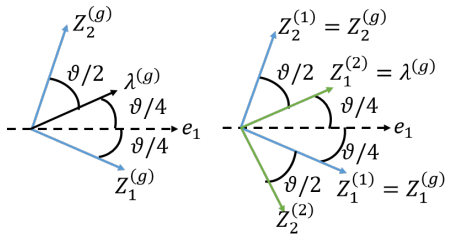


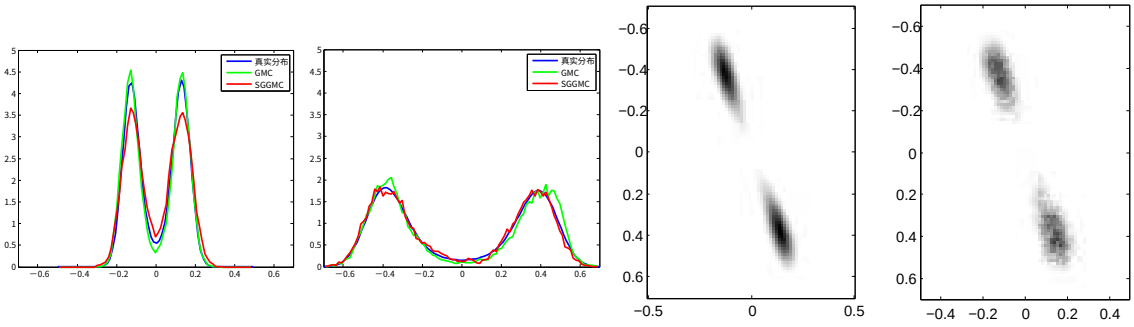
图 3.3　合成数据实验的设定（其中 $\vartheta = \frac{\pi}{6}$）。左图展示了生成数据时所用的隐变量的值 $Z_1^{(g)}, Z_2^{(g)}$，以及所生成的数据的一个峰值位置 $\lambda^{(g)}$（另一个峰值位置是 $Z_1^{(g)}$）；右图展示了弱先验下后验分布的两个峰值位置的近似：一个是 $(Z_1^{(1)}, Z_2^{(1)})$（蓝色箭头所示），另一个是 $(Z_1^{(2)}, Z_2^{(2)})$（绿色箭头所示）。

协方差矩阵 Σ [①]，并选取随机子数据集大小 $|\widetilde{\mathscr{D}}| = 10$，单批学习率 $b = 5 \times 10^{-4}$，以及动量系数 $\varsigma = 0.1$。

在展示实验结果之前，先来分析一下真实的后验分布应该具有什么样的特点。由于实验中数据是从密度函数正比于 $\text{vMF}(X|Z_1^{(g)}, \kappa_X) + \text{vMF}(X|\lambda^{(g)}, \kappa_y)$（其中 $\lambda^{(g)} := \frac{Z_1^{(g)} + Z_2^{(g)}}{\left\|Z_1^{(g)} + Z_2^{(g)}\right\|}$）的分布中所采的，而由 vMF 分布的单峰性，可以推测以这种方式得到的数据的分布会在 $Z_1^{(g)}$ 及 $\lambda^{(g)}$ 附近有两个峰值。而另一方面，抛开数据来说，由此模型的生成过程可以得知，数据的分布会在 Z_1 和 λ 附近有两个峰值。因此将这两个理论峰值位置与所拿到的数据的两个峰值位置进行匹配，可以大致（在弱先验的情况下）得知 Z_1 和 Z_2 后验的峰值位置。由于这个匹配有两种情况，因而 Z_1 与 Z_2 后验分布的峰值位置的近似也有两个：1) 令 $Z_1 = Z_1^{(g)}, \lambda = \lambda^{(g)}$，可解得：$Z_1^{(1)} := Z_1^{(g)}, Z_2^{(1)} := Z_2^{(g)}$，亦即生成数据时所使用的 Z_1 和 Z_2 的真实值；2) 令 $Z_1 = \lambda^{(g)}, \lambda = Z_1^{(g)}$，可解得：$Z_1^{(2)} := \lambda^{(g)}$，$Z_2^{(2)}$ 如图 3.3（右）所示。注意到后验分布的两个峰值位置的上述近似是关于 e_1 对称的，因而先验对它们的偏好是一样的。实验中的设定对应着 $\vartheta = \frac{\pi}{6}$，所以后验分布应在 $(Z_1, Z_2) = (Z_1^{(1)}, Z_2^{(1)}) = (-\frac{\pi}{24}, \frac{\pi}{8})$ 以及 $(Z_1, Z_2) = (Z_1^{(2)}, Z_2^{(2)}) = (\frac{\pi}{24}, -\frac{\pi}{8})$ 附近有两个峰值。这两个峰值位置的存在体现了后验分布中 Z_1 和 Z_2 之间的相关性，因而一个好的贝叶斯推理算法需要挖掘到这个相关性。

图 3.4(a-b) 在角度空间中展示了 Z_1 和 Z_2 的边缘后验分布的真实分布与 GMC 和 SGGMC 所采样本的经验分布，而图 3.4(c) 展示了 Z_1 和 Z_2 的联合后验分布的真实分布与 SGGMC 所采样本的经验分布。（这些图都是在角度空间中所作。由于从嵌入空间到角度空间的变换具有单位雅可比行列式（Jacobian determinant）（或

① 实验中发现，这个做法与直接将 Σ 取为 0 的效果没有明显区别。



(a) $p(Z_1|\mathscr{D})$　　(b) $p(Z_2|\mathscr{D})$　　(c) $p(Z_1, Z_2|\mathscr{D})$

图 3.4　合成数据实验结果：(a-b) 边缘后验分布 $p(Z_1|\mathscr{D})$ 与 $p(Z_2|\mathscr{D})$ 的真实分布以及 GMC 和 SGGMC 所得样本的经验分布；(c) 联合后验分布 $p(Z_1, Z_2|\mathscr{D})$ 的真实分布（左）与 SGGMC 所得样本的经验分布（右）。

者说，嵌入空间中的豪斯多夫测度对角度空间这个坐标空间中的勒贝格测度的拉东-尼科迪姆导数（Radon-Nikodym derivative）处处为 1），所以两个空间中的密度函数值相同。）可以发现，即使在使用子数据集（随机梯度）时，SGGMC 样本的分布也没有出现本质上的破坏。特别地，SGGMC 方法所得到的后验近似能够充分触及到上述分析中后验分布的两个峰值位置，因而充分挖掘到了 Z_1 与 Z_2 的相关性。

3.4.3　球面混合模型实验

现在考察所提的 SGGMC 和 gSGNHT 方法在真实任务上的表现和优势。实验场景选择为球面混合模型（spherical admixture model, SAM）处理真实文档数据的后验推理任务。使用 SGGMC 及 gSGNHT 方法完成此任务的具体方法参见 3.3 节及算法 3，其中式 (3-10) 中的随机梯度噪声协方差矩阵的估计 Σ 选取为零。本节把这两个使用随机子数据集（即使用随机梯度）的方法分别记为 SGGMC-batch 和 gSGNHT-batch。为了与只可使用全数据集的推理方法对比以及展示使用随机子数据集所带来的效率，本节也考虑使用全数据集进行训练的 SGGMC 和 gSGNHT 方法，并将它们分别记作 SGGMC-full 和 gSGNHT-full。

基准线方法　实验中选择用来对比效果的基准线方法包括：Reisinger 等人[22] 所提的平均场变分推理方法（variational inference, VI），根据 Hoffman 等人[30] 所做工作所改进的 VI 方法的随机梯度版本（stochastic variational inference, StoVI），以及 GMC 这个只可使用全数据集的不可扩展的 MCMC 方法。

对于 GMC 方法，有一个困难的地方需要额外说明。GMC 方法在其 MH 测试这个步骤中需要估计后验分布密度函数的对数 $\log p(\beta|X)$ 的值，但因为需要对局

部隐变量 θ 进行积分，所以即使是它的未归一化密度函数都无法准确计算。参考式 (3-12) 中对其梯度的估计方式 $\nabla_\beta \log p(\beta|X) = \mathbb{E}_{p(\theta|X,\beta)}\left[\nabla_\beta \log p(\beta,\theta|X)\right]$，一种可能的使用 $p(\theta|X,\beta)$ 样本 $\{\theta^{(l)}\}_{l=1}^{L}$ 的估计方式为：

$$\log p(\beta|X) \approx \frac{1}{L}\sum_{l=1}^{L} \log p(\beta,\theta^{(l)}|X) + \text{const.} \tag{3-14}$$

然而遗憾的是，这个估计是有偏的。此估计与真实值之间的差别可由下式看出：

$$\begin{aligned}
\log p(\beta|X) &= \mathbb{E}_{p(\theta|X,\beta)}\left[\log p(\beta|X)\right] \qquad \text{（这是因为 } \log p(\beta|X) \text{ 与 } \theta \text{ 无关）} \\
&= \mathbb{E}_{p(\theta|X,\beta)}\left[\log p(\beta,\theta|X) - \log p(\theta|X,\beta)\right] \\
&\approx \frac{1}{L}\sum_{l=1}^{L}\left(\log p(\beta,\theta^{(l)}|X) - \log p(\theta^{(l)}|X,\beta)\right),
\end{aligned}$$

即两者的偏差为 $\frac{1}{L}\sum_{l=1}^{L}\log p(\theta^{(l)}|X,\beta)$。对于这一偏差，虽然 $\log p(\theta|X,\beta)$ 在相差一个加性常数的意义下是可知的，但此常数只是对于 θ 是常数，它与 β 有关，而这里所考虑的正是关于 β 的函数。因而这一偏差仍然是无法计算的。尽管如此，式 (3-14) 似乎是估计 $\log p(\beta|X)$ 的唯一方式。它可以看作是在 β 的提议样本（proposal）与当前样本距离不远的情况下对 $\log p(\beta|X)$ 的一种近似，因为在这种情况下，偏差 $\frac{1}{L}\sum_{l=1}^{L}\log p(\theta^{(l)}|X,\beta)$ 可近似看作关于 β 的常数。此处把在 MH 测试中采用这种方式估计 $\log p(\beta|X)$ 的用于 SAM 模型后验推理任务的 GMC 方法称为 GMC-apprMH，其中 “apprMH” 是 “approximate MH test” 即 “近似 MH 测试” 的缩写。GMC-apprMH 方法中的梯度用与 SGGMC 方法相同的方式来估计，只不过每次需要在整个数据集上进行计算而不是在一个随机子数据集上估计。

由于 GMC-apprMH 方法中 MH 测试的不准确性，本节也考虑另一个使用 GMC 来完成 SAM 模型后验推理任务的方法，即 GMC-bGibbs，其中 “bGibbs” 是 “blockwise Gibbs sampling” 即 “分块吉布斯采样” 的缩写。此方法采取分块吉布斯采样的思想，交替地从 $p(\beta|\theta,X)$ 以及 $p(\theta|\beta,X)$ 中进行采样。由于这两个分布的未归一化密度函数都是已知的（即将式 (3-11) 所给出的 $p(X,\beta,\theta)$ 分别看作 β 和 θ 的函数），因此它们样本是可以通过 GMC 方法来采取的。根据分块吉布斯采样的性质，这个方法是渐进准确的，但可能会收敛得较慢。此外，GMC-apprMH 方法与 SGGMC 和 gSGNHT 是处在一个框架下的，它们都会在每次采取一个 β 的样本时采取多个 θ 的样本，而 GMC-bGibbs 则只采取一个 θ 的样本。

数据集　本节选取一大一小两个文档数据集进行实验。在小数据集上，各 GMC 方法以及 VI 方法等不具有可扩展性的方法也可以在可行的时间内收敛，因而可以考察各方法的最终收敛结果，而在大数据集上可以展示所提方法的可扩展性优势。这两个数据集都是以文档的归一化词频-逆篇频（term frequency-inverse document frequency, tf-idf）特征来表示的。这个特征是由这些文档原本的词袋（bag of words）特征转化而得到的。具体地，原本的词袋特征提供了词频 $\mathrm{tf}(d,w)$，即词 w 在文档 d 中出现的次数。而词频-逆篇频特征 $\mathrm{tfidf}(d,w)$ 可由下式计算得到：

$$\mathrm{tfidf}(d,w) = \mathrm{tf}(d,w) \log\big(|\mathscr{D}|/(1+\mathrm{df}(w))\big),$$

其中 $\mathrm{df}(w)$ 是词 w 的篇频，即包含词 w 的文档数。最终文档 d 的特征是将词数维向量 $\big(\mathrm{tfidf}(d,w)\big)_w$ 进行 2-范数的归一化而得到的。

小数据集 20News-different 是标准文档数据集 20Newsgroups[①] 的一个子集。这个数据集最初是在 SAM 模型的原论文[22] 中被提出的，用来展示 SAM 模型优于 LDA 模型[19] 的文档特征提取能力。它包含了原数据集共 20 个类别的文档中的 *rec.sport.baseball*，*sci.space* 和 *alt.atheism* 这 3 个类别的文档，共有训练文档 1,666 篇，测试文档 1,107 篇。原数据集的词数为 61,188，本实验根据一个适当的篇频（介于 0.36% 和 11.77% 之间）选取了 5,000 词作为新词表。

大数据集 150K-Wikipedia 是本实验所构建的一个数据集。它是基于 Zhang 等人[148] 所使用的具有 660 万篇文档的 Wikipedia 数据集[②] 而构建的。本实验从原数据集中词数大于 20 的文档中随机挑选 15 万篇作为训练集，1 千篇作为测试集。这个训练集的大小是与 Patterson 等人[18] 展示可扩展性的实验中所使用的训练集大小是相同的。原数据集的词数为 7,702，本实验根据一个适当的篇频（介于 0.44% 和 5.99% 之间）选取了 3,000 词作为新词表。

这两个处理好的数据集可从网站 “http://ml.cs.tsinghua.edu.cn/~changliu/sggmcmc-sam/” 下载。

实验设定及实现细节　对 SAM 模型做后验推理时，无论采用哪种推理算法，模型的超参数都取为定值。在小数据集 20News-different 上，各超参数选定为：$\sigma = 1\times10^4$，$\kappa_0 = 1\times10^4$，$\kappa_1 = 3\times10^4$，$\alpha = 10$，话题数 $T = 20$，而在大数据集 150K-Wikipedia 上，这些超参数选定为：$\sigma = 6\times10^3$，$\kappa_0 = 6\times10^3$，$\kappa_1 = 2\times10^4$，$\alpha = 10$，话题数 $T = 50$。本实验中，SAM 模型的向量值超参数 λ_0 被选定为各数

① 可从网站 “http://www.qwone.com/~jason/20Newsgroups/” 下载其标准版本。本工作中使用的是其 Matlab/Octave 版本。

② 可从网站 “http://ml.cs.tsinghua.edu.cn/~aonan/datasets/wikipedia/” 下载。

据集训练集中的文档特征均值的归一化向量。可扩展方法（StoVI, SGGMC-batch, gSGNHT-batch）所使用的随机子数据集的大小 $|\widetilde{\mathscr{D}}|$ 在小数据集 20News-different 上选定为 50，而在大数据集 150K-Wikipedia 上选定为 100。

VI 方法是基于提出 SAM 模型的工作[22] 所提供的 MATLAB 代码实现的，而 StoVI 方法是基于这个代码改写的。各 MCMC 推理方法都是基于 C++ 代码实现的。如本章 3.3 节及算法 3 中所述，对于文档 d，各 MCMC 推理方法都需要从分布 $p(\theta_d|X_d,\beta)$ 中采取话题配比 θ_d 的样本，并且这可以通过 GMC 方法来实现。在使用 GMC 方法进行这一步骤时，实验中使用如下初始化操作：$\theta_d=(\beta^\top\beta)^{-1}\beta^\top X_d$。这个 θ_d 的初始化值是无信息先验 $\alpha=1$ 的情况下，分布 $p(\theta_d|X_d,\beta)$ 的峰值位置。GMC-apprMH 和 GMC-bGibbs 方法在每次采取一个话题 β 的样本时需要对整个训练数据集中的全部文档进行上述 θ_d 的采样操作，而 SGGMC 和 gSGNHT 方法则只需要在随机选出的子数据集上进行这个 θ_d 的采样操作即可。注意到对于不同的文档，采取其各自的话题配比 θ_d 的过程是相互独立的，因而可考虑并行化这个过程。实验中使用了 OpenMP[①]在它们的 C++ 代码中实现并行。由于对所有的 MCMC 方法，每次采取一个话题 β 的样本时都需要这个采取各文档的话题配比 θ_d 的过程，因而只要在运行时使用的线程数相同，这个并行化操作对各 MCMC 方法来说仍然是公平的。实验中选取这个相同的线程数为 32。所有的实验代码可从网站“http://ml.cs.tsinghua.edu.cn/~changliu/sggmcmc-sam/”下载。

推理效果衡量方法　由于 SAM 模型的后验推理任务是一个无监督学习的任务，因此使用对数困惑度（log-perplexity）来评判这些后验推理方法的效果。对数困惑度衡量的是所训练的模型——在此处的设定下即是各推理方法所得到的 SAM 模型的后验分布的近似——在面对测试数据时的“困惑程度”，或者说是与测试数据的相左程度。一个好的训练结果——此处认为是一个好的推理方法所得的结果——会与测试数据更加契合，因而会有较小的对数困惑度。它可通过所训练的模型在测试数据集 $\mathscr{D}_{\text{test}}$ 上的负对数似然的平均值来计算。具体来说，变分推理方法 VI 和 StoVI 得到的是后验分布的一个点估计 $\hat{\beta}$，对应的对数困惑度为：

$$\text{log-perp}=-\frac{1}{|\mathscr{D}_{test}|}\sum_{d\in\mathscr{D}_{\text{test}}}\log p(X_d|\hat{\beta}),$$

而采样方法得到的是后验分布的一组样本 $\{\beta^{(i)}\}_{i=1}^N$，对应的对数困惑度为：

$$\text{log-perp}=-\frac{1}{|\mathscr{D}_{test}|}\sum_{d\in\mathscr{D}_{\text{test}}}\log\big(\frac{1}{N}\sum_{i=1}^N p(X_d|\beta^{(i)})\big).$$

① 参见网站“http://openmp.org/”。

由于在这两者情况下都需要对 $p(X_d|\beta)$ 进行估计。这个量已被多次提到是无法准确计算的，但它仍然可以通过采样的方式来估计。注意到 $p(X_d|\beta) = \int p(X_d, \theta_d|\beta)\,\mathrm{d}\theta_d = \mathbb{E}_{p(\theta_d|\beta)}[p(X_d|\beta, \theta_d)]$，而 $p(\theta_d|\beta) = p(\theta_d) = \mathrm{Dir}(\theta_d|\alpha)$（即 θ_d 的先验分布）是可以快速而准确地采样的，且 $p(X_d|\beta, \theta_d)$ 是可以准确计算的（由 SAM 模型的生成过程给出），因此对于每个 β 的值，都可以先从狄利克雷分布 $\mathrm{Dir}(\theta_d|\alpha)$ 中准确采取一组样本 $\{\theta_d^{(l)}\}_{l=1}^{L}$，再计算 $p(X_d|\beta, \theta_d^{(l)})$ 的均值，作为 $p(X_d|\beta)$ 的估计。

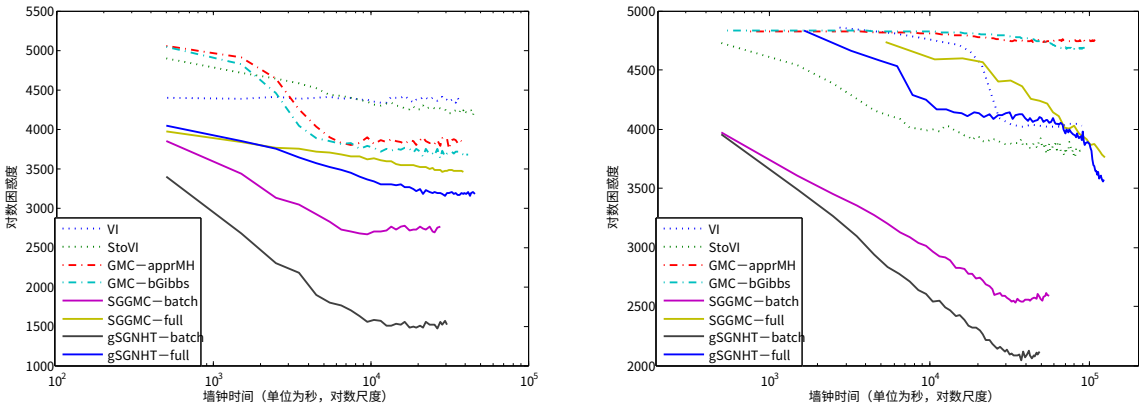
最后需要说明的是，这里计算的对数困惑度与著名的隐式狄利克雷分配模型（latent Dirichlet allocation, LDA）所计算的对数困惑度是不可比的，因为两个模型使用的是数据的不同形式的表示，并且使用了完全不同的分布来为数据建模。特别地，LDA 模型中使用了离散空间上的分布建模数据，而 SAM 模型使用了连续空间中的分布。由于前者在一个离散的点上给出的概率与后者在一个连续的点上给出的概率密度具有完全不同的含义，因此二者的数值不可比较。此外，这里希望考察的是针对 SAM 这个特定的模型的不同推理方法的效果比较，因此与 LDA 这个另外的模型所给出的结果无关。

避免数值溢出的技术处理　如本章 3.3 节中所述，SAM 模型使用了 vMF 分布对球面上的分布进行建模，而此分布的归一化系数 c_n 中涉及了第一类修正贝塞尔函数 $\mathcal{I}_k(x)$，其中 k 表示其阶数，而其自变量 x 是一个正实数。实验中发现，当阶数 k 很大时，$\mathcal{I}_k(x)$ 很容易要么趋于零要么趋于正无穷，从而带来数值问题。由于本实验中，阶数 $k = n/2 - 1$，其中 n 是词表中的词数，因而阶数 k 在上千的量级上，使得数值问题很容易发生。另一方面，为了避免没有意义的模型，SAM 模型中的超参数 σ，κ_0 及 κ_1 都被选取为一个相对大的值，而这会导致数值溢出问题几乎总是会发生。不过幸好在所提算法中，只有贝塞尔函数的对数 $\log \mathcal{I}_k(x)$ 是需要计算的量，而它的增长就会缓和很多。所以如果能够直接计算 $\log \mathcal{I}_k(x)$（而不是先计算 $\mathcal{I}_k(x)$ 再计算其对数），那么就能有效避免数值溢出的问题。

注意到

$$\log \mathcal{I}_k(x) = \log \left(\sum_{i=0}^{\infty} \frac{1}{i!\Gamma(i+k+1)} \left(\frac{x}{2}\right)^{2i+k} \right), \tag{3-15}$$

其中 $\Gamma(\cdot)$ 是伽马函数（Gamma function），因此可以借助对数和技术（log-sum trick）来直接计算其对数值。对数和技术是指在只知道 $a = \log A$ 以及 $b = \log B$（不失一般性地假设 $a \geqslant b$）的情况下，不通过计算 A 或 B 的值（它们可能会非常大以至



(a) 20News-different　　(b) 150K-Wikipedia

图 3.5　各推理方法所得的对数困惑度（log-perplexity）随推理所需的墙钟时间（wall-clock time）的变化曲线。(a) 20News-different 数据集上的结果；(b) 150K-Wikipedia 数据集上的结果。

于会导致数值溢出）来计算 $\log(A+B)$ 的值。由于

$$\log(A+B) = \log(\exp(a)+\exp(b)) = \log\big(\exp(a)\big(1+\exp(b-a)\big)\big)$$
$$= a + \log\big(1+\exp(b-a)\big),$$

且由于 $b-a \leqslant 0$ 因而 $1 < 1+\exp(b-a) \leqslant 2$，所以只要 a 和 b 可以正常表示，那么上式最后的形式中的每一项都不会造成数值溢出的问题，从而可以使计算过程更加稳定可靠。参照式 (3-15)，此技术可被不断地重复来计算 $\log\left(\sum_{i=0}^{l}\frac{1}{i!\Gamma(i+k+1)}\left(\frac{x}{2}\right)^{2i+k}\right), l = 0,1,2,\cdots$。由于每一个求和项 $\frac{1}{i!\Gamma(i+k+1)}\left(\frac{x}{2}\right)^{2i+k}$ 随 i 的增长会非常快地衰减（这是因为 $1/i!$ 及 $1/\Gamma(i+k+1)$ 都以快于指数的速度衰减），因而计算中无需取很大的 l 的值即可得到 $\mathcal{I}_k(x)$ 的很好的近似。在本实验的实现中，l 的选取是根据近似精度来确定的。对于一般的情况，l 的取值处于 10 左右。更多实现细节请参见本实验的代码（“http://ml.cs.tsinghua.edu.cn/~changliu/sggmcmc-sam/”）。

小数据集上的实验结果　图 3.5(a) 展示了各推理方法在小数据集 20News-different 上的表现。可以发现，所提 SGGMC 方法和 gSGNHT 方法都比其他方法表现更好。VI 方法收敛得很迅速，但其收敛后的结果无法进一步改善，这是因为它所做的平均场假设为它贴合真实后验分布的能力设下了一道鸿沟。StoVI 方法在这个小规模的数据集上收敛得比 VI 方法慢（因为此情况下计算准确梯度的代价并不高，而使用随机梯度这个有噪近似相比之下会更加影响收敛速度），但也同样受限于平均场

假设。所有 MCMC 方法（采样方法）最终都比 VI/StoVI 方法有更好的结果，这体现了 MCMC 方法的近似灵活性以及渐进准确性的优势。其中，两个 GMC 方法的表现类似，并且在一开始收敛得都比较慢，初期阶段的结果不如 VI/StoVI 方法，而所提 SGGMC/gSGNHT 方法则收敛得比这些方法都快，并且也取得了更好的结果。对于 SGGMC/gSGNHT 方法，使用更小的随机子数据集（即-batch 方法）的收敛速度要优于使用全数据集（即-full 方法），这得益于使用随机子数据集计算随机梯度的更小计算代价，以及可以平衡随机梯度噪声的恰当动力学系统。注意到使用了全数据集的 SGGMC-full/gSGNHT-full 方法仍然比两个 GMC 方法收敛更快，这可能是因为 SGGMC/gSGNHT 方法的动力学系统中的随机性可以帮助样本跳出局部最优点从而可以更快地找到更多的最优点以及全局最优点。另外，gSGNHT 方法的表现优于 SGGMC 方法，这体现了使用恒温器变量的好处。

大数据集上的实验结果　图 3.5(b) 展示了各推理方法在大数据集 150K-Wikipedia 上的表现。可以看出，所提方法 SGGMC/gSGNHT 相对基准线方法的优势在大数据集上变得更加明显，即收敛得更快且收敛结果更好。这表现出了所提方法的可扩展性。在此大数据集上 StoVI 方法的收敛速度快于 VI 方法，但两者最终的推理效果仍然被其平均场假设所限制。使用恒温器变量的好处以及使用随机子数据集的加速效果也同样可见，使用整个数据集的 SGGMC-full 和 gSGNHT-full 方法的效果也仍然优于两个 GMC 方法。这两个 GMC 方法是不可扩展的，它们在此大数据集的情况下收敛得非常慢，以至于在可行的时间内它们的推理效果甚至不如 VI/StoVI 方法。

3.5　本章小结与讨论

本章提出 SGGMC 和 gSGNHT 这两个随机梯度 MCMC 方法，用来针对大规模数据从定义在流形上的后验分布中高效采样。它们使用随机梯度来高效处理大规模数据，并使用流形嵌入技术解除了流形必须具有全局坐标系的限制。本章为它们设计了适合使用随机梯度的动力学系统，并开发了二阶测地线积分器来进行高效模拟。合成数据实验验证了它们的正确性和有效性，而它们在解决真实数据上 SAM 模型的后验推理任务中的表现明显地优于已有方法，展示了它们优秀的准确性和可扩展性。

所提方法仍然具有广阔的应用场景，包括使用 vMF 分布的模型（例如 vMF 混合模型[33,149-150] 以及狄利克雷过程（Dirichlet process）vMF 混合模型[151-153]）的高效后验采样，受限分布的高效采样[126]（例如截断高斯分布（truncated Gaussian

distribution)), 以及定义在斯蒂菲尔流形（Stiefel manifold）上分布的高效采样（例如贝叶斯矩阵补全任务[40]）等。所提方法的可扩展性在这些场景中都是一个重要的优势。

第 4 章　黎曼-斯坦因变分梯度下降方法

本章介绍黎曼-斯坦因变分梯度下降方法（Riemannian Stein variational gradient descent, RSVGD）。这个贝叶斯推理方法是斯坦因变分梯度下降方法（Stein variational gradient descent, SVGD）向黎曼流形情况的推广。这个推广具有两方面的好处：(a) 针对欧氏空间中的贝叶斯推理任务，RSVGD 方法可以使用信息几何（information geometry）技术从而比 SVGD 方法更加高效；(b) 针对黎曼流形上的贝叶斯推理任务，RSVGD 相比于此领域的现有方法，具有 SVGD 的独特优势，包括粒子高效性（particle efficiency）、迭代有效性（iteration-effectiveness）以及近似灵活性（approximation flexibility）。为能正确地推广到黎曼流形上，本章为 RSVGD 设计了原创的且具有技术复杂度的方法，来处理一般黎曼流形与欧氏空间本质上不同的特性。在推导 RSVGD 方法的过程中，本章也关注了与之相关的统计问题，提出了黎曼-斯坦因恒等式（Riemannian Stein's identity）以及黎曼-核化斯坦因差异量（Riemannian kernelized Stein discrepancy）。实验结果展示了 RSVGD 在欧氏空间推理任务上使用信息几何的能力所带来的胜于 SVGD 的高效性，以及在黎曼流形推理任务上胜于已有推理方法的粒子高效性，迭代有效性，以及近似灵活性。

4.1　研究动机

首先简要回顾贝叶斯推理的各类方法及其特点。贝叶斯推理是学习贝叶斯模型用以从数据中提取知识这一任务的核心。它的目标是要估计给定观测数据之后模型隐变量的后验分布，而这个后验分布通常十分复杂而没有闭式解，亦即它是不可行的（intractable）。变分推理方法（variational inference, VI）采用一个可行的（tractable）分布来近似后验。传统的 VI 方法通常会使用一个参数化分布族，或者说是一个统计模型（statistical model），来作为这个可得的分布。这类方法被称为基于模型的变分推理方法（model-based variational inference, ModVI）。这样一来，近似后验分布的任务便可转化为一个参数优化问题，进而可以使用各种成熟而高效的优化方法来求解。但由于所选的参数化分布族的覆盖范围（近似能力）终究有限（例如平均场（mean-field）形式的分布族无法描述变量之间的相关性），因而 ModVI 方法近似后验分布的精度始终会被一道无法跨越的鸿沟所限制。蒙特卡罗（Monte Carlo）方法，特别是其中应用非常广泛的马尔可夫链蒙特卡罗（Markov chain Monte Carlo, MCMC）方法，则希望直接从后验分布中采取样本来估计此分布。尽管它们具有渐进准确性（asymptotic accuracy）的保障，但其有限多样本的

效果在实际中通常难以保证，而且由于随机性的模拟以及样本之间的正自相关性（positive auto-correlation），它们的收敛相对较慢。

近来出现了一类新的 VI 方法，称为基于粒子的变分推理方法（particle-based variational inference, ParVI）。这类方法使用一组样本，或者被称为“粒子”（particle），来表示一个用来近似后验分布的变分分布（variational distribution），并且通过一个确定性的更新粒子的方法来最小化变分分布与后验分布的差距。与 MCMC 方法类似的是，ParVI 方法使用粒子这个非参数形式来表示变分分布具有很强的近似灵活性（approximation flexibility），这使得它们可以突破 ModVI 方法中遇到的近似能力的鸿沟从而取得愈发准确的近似。而胜于 MCMC 方法的是，ParVI 方法仍然是基于最小化分布之间的差距这一原则的，而由这个最优化原则得到的更新方式可使得这类方法具有迭代有效性（iteration-effectiveness），亦即每一步迭代（每一步更新）都可以保证得到更好的结果。虽然已有一些针对具体 MCMC 方法收敛到平稳分布的分析工作[58,104-105,109]，但一般 MCMC 方法的原则只是目标分布等于平稳分布，即若当前是平稳分布则之后也是此分布，而不一定保证从任一分布开始的演化过程会与平稳分布越发接近。此外，MCMC 方法通常需要一个较大的样本规模才能给出一个对目标分布的有效近似，而 ParVI 方法由于直接考虑有限个样本（粒子）的近似效果，因而它们可以通过更少的样本达到同样的近似效果，即粒子高效性（particle efficiency）。这个优势不仅可以节省存储推理结果的空间，还可以在处理后续任务（如预测等）中节省时间。研究现状介绍部分 1.2 节中的表 1.1 中已列出了这三类贝叶斯推理方法的比较，其中可以看出 ParVI 方法的优势。

斯坦因变分梯度下降方法（Stein variational gradient descent, SVGD）[81] 是 ParVI 方法中的一个杰出代表。SVGD 更新粒子的方式是通过在这组粒子上施加一个合适的确定性连续时间动力学系统（deterministic continuous-time dynamics）使得这组粒子所代表的分布可以朝向目标分布演化。SVGD 已经在实际问题中得到诸多应用，包括一些贝叶斯模型的推理任务[154-155] 以及强化学习任务[24-25]）。

贝叶斯推理领域中另外一个考量则是与黎曼流形的结合。这个思想的重要性可由两方面来体现：(a) 有一些模型的隐变量本身就处在一个特定黎曼流形上，例如球面混合模型（spherical admixture model, SAM）[22]，因而针对这些模型的贝叶斯推理的任务就变成了近似一个给定黎曼流形上的分布；(b) 常规贝叶斯模型的欧氏隐变量可看作这个模型的所有似然分布所构成的黎曼流形的一个自然的坐标系（coordinate system），因而也可以考虑在这个分布流形上进行贝叶斯推理，借用这个分布流形更加本质的几何特征来提高推理效率，即信息几何（information geometry）的思想[45-46]。这两方面的考量近年来也取得了诸多进展。在情况 (a) 方

面, Bonnabel[122] 和 Zhang 等人[156] 开发了可扩展的（scalable）以及稳定的黎曼流形上的优化方法, 从而可以加强 ModVI 方法处理流形隐变量的效率和效果。Brubaker 等人[125] 和 Byrne 等人[44] 开发了更加高效处理流形隐变量的 MCMC 方法，而本文已在第 3 章中为这些方法实现了可扩展性。在情况 (b) 方面，Hoffman 等人[30] 在 ModVI 方法中使用了基于信息几何的自然梯度提高了推理效率，而 Girolami 等人[127] 和 Ma 等人[120] 开发了可以利用黎曼流形结构的 MCMC 方法从而可以使用信息几何提高效率，随后 Li 等人[157] 将这些 MCMC 方法应用在贝叶斯神经网络（Bayesian neural network）的后验推理任务上, 取得了效果上的提高。然而, 在 ParVI 领域几乎还没有考虑黎曼流形结构的工作。Gemici 等人[158] 尝试将标准化流拓展到黎曼流形上，但他们的方法无法用于没有全局坐标系的流形上，例如超球面。

本章提出黎曼-斯坦因变分梯度下降方法（Riemannian Stein variational gradient descent, RSVGD），这个将 SVGD 方法优雅地拓展到黎曼流形的方法。所提方法 RSVGD 既可用于黎曼流形上后验分布的近似（即情况 (a)），也可利用信息几何来提高欧氏空间上后验推理的效率（即情况 (b)）。RSVGD 继承了 SVGD 的显著优势，为流形隐变量的推理任务带来了具有诸如粒子高效性等好处的解决方法。另外从技术上来说，向黎曼流形进行拓展并不是一个直接可得的操作，因为一般的黎曼流形具有与欧氏空间完全不同的性质（例如参见毛球定理（[131]，定理 8.5.13）），使得通常的处理方法不再适合。进行上述拓展必须要设计新的技术来解决这些细节问题。具体地，本章首先对 SVGD 方法进行一定的抽象，将其看作欧氏空间中一个流（flow）所对应的动力学系统下分布的演化，并将这个过程拓展到黎曼流形上。然后，通过一个原创的方法求解最优的动力学系统，进而得到 RSVGD 的算法及其在流形的坐标空间中的表示。SVGD 针对这个问题所使用的方法在黎曼流形的情况下不再适用。接着，本章也推导出了 RSVGD 方法在流形的嵌入空间（embedded space）中的表示形式，这样一来，像超球面这样的没有全局坐标系的黎曼流形上的推理任务便也可通过 RSVGD 解决，特别是避免了这种情况下在坐标空间中会出现的数值不稳定的问题。最后，本章在开发新方法的过程中也得到了一些副产品：黎曼-斯坦因恒等式（Riemannian Stein's identity）和黎曼-核化斯坦因差异量（Riemannian kernelized Stein discrepancy），作为相应概念在黎曼流形上的拓展。为考察所提方法 RSVGD 在实际任务中的表现，本章将 RSVGD 应用于贝叶斯逻辑回归模型（Bayesian logistic regression）以及 SAM 模型的后验推理任务上，并分别与 SVGD 和黎曼流形上的推理方法进行对比。实验结果验证了所期待的更快的收敛速度以及粒子高效性等优势。

需要说明的是，本章所考虑的黎曼流形是粒子（即样本）所在的空间，或者

说是目标分布的支撑空间（support space），这不同于最近解释 SVGD 行为的一些工作[85,159] 中所提到的由一些特定概率分布所构成的、以再生核希尔伯特空间（reproducing kernel Hilbert space, RKHS）$\mathscr{H}$ 为切空间的黎曼流形 $\mathscr{P}_{\mathscr{H}}$。那些工作提出流形 $\mathscr{P}_{\mathscr{H}}$ 是为了将原本的 SVGD 解释为在流形 $\mathscr{P}_{\mathscr{H}}$ 上最小化与目标分布之间 KL 散度的过程，而没有对 SVGD 方法进行拓展。而本章工作将 SVGD 方法的适用范围推广到了一般的黎曼流形上，从而可以提高各场景下推理任务的效率。

4.2　背景知识

关于黎曼流形的结构和性质，可参见 2.1 节。为解决本章所关心的问题，此部分引入雷诺输运定理，并从适合流形情形的角度介绍斯坦因变分梯度下降方法。

4.2.1　雷诺输运定理

雷诺输运定理（Reynolds transport theorem）可将一个动力学系统与对应的分布演化规律联系起来。这个定理是对定积分求导规则的推广，也是流体力学的基础。令 $V \in \mathscr{T}(\mathscr{M})$ 表示黎曼流形 $\mathscr{M}$ 上的一个向量场，而 $F_{(\cdot)}(\cdot)$ 表示它的流（参见 2.1.1.3 节）。对于光滑函数 $f_{(\cdot)}(\cdot) : \mathbb{R} \times \mathscr{M} \to \mathbb{R}$ 以及 $\mathscr{M}$ 的任意一个开子集 $\mathscr{I} \subset \mathscr{M}$，雷诺输运定理给出如下结论：

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{F_t(\mathscr{I})} f_t \omega_g = \int_{F_t(\mathscr{I})} \left(\frac{\partial f_t}{\partial t} + \mathrm{div}(f_t V) \right) \omega_g,$$

其中 $\mathrm{div} : \mathscr{T}(\mathscr{M}) \to C^\infty(\mathscr{M})$ 是向量场的散度，而 ω_g 是黎曼体积形式（参见 2.1.2.5 节）。在坐标系中，散度具有表达式 $\mathrm{div}(V) = \partial_i(\sqrt{|G|}V^i)/\sqrt{|G|}$，其中 $m \times m$ 维矩阵 G 是此坐标系中的黎曼度量矩阵，它由此坐标系中黎曼度量的表示 (g_{ij}) 所构成的矩阵，而 $|G|$ 表示它的行列式（determinant）。关于此定理的更多细节可参见 Romano 的著作[133] 第 164 页，Frankel 的著作[160] 第 142 页，以及 Abraham 等人的著作[131] 第 469 页。

4.2.2　斯坦因变分梯度下降方法（SVGD）

本节将从一个较为抽象但易于推广到黎曼流形的角度来介绍一下斯坦因变分梯度下降方法（Stein variational gradient descent, SVGD）[81] 这个 ParVI 方法。SVGD 方法通过在粒子上施加一个恰当的动力学系统来不断地更新它们，使得这些粒子所代表的分布在 KL 散度的意义下向着目标分布演化。它所考虑的粒子空间是最常见的欧氏空间 $\mathscr{M} = \mathbb{R}^m$。将粒子所代表的在动力学系统 V 下不断演化的分布记

为 q_t，并记目标分布为 p。SVGD 的第一个关键结论是：

$$-\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(q_t||p) = \mathbb{E}_{q_t}[V^\top \nabla \log p + \nabla^\top V]. \tag{4-1}$$

这个量所衡量的正是 q_t 趋向 p 的速度。为最小化 KL 散度，人们自然是希望这个减小率越大越好，亦即 V 应最大化 $-\mathrm{KL}(q_t||p)$。与之相似的是，在最大化一个函数 f 时，所希望寻找的更新方向 v 正是可以最大化函数 $f(x)$ 的方向导数 $f'_v(x)$ 的方向：$v^*(x) := \max_{v:\|v\|=1} \cdot \operatorname{argmax} f'_v(x)$，其中“$\max \cdot \operatorname{argmax}$”表示最优函数值与最优变量的数乘，而这里所找到的更新方向 $v^*(x)$ 正是函数 f 的梯度 $\nabla f(x)$。类比这一观点，SVGD 工作的原文[81] 中将 $-\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(q_t||p)$ 称为沿着向量场 V 的*方向导数*（directional derivative），并将 $V^* := \max_{V:\|V\|=1} \cdot \operatorname{argmax} -\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(q_t||p)$ 称为 KL 散度的*泛函梯度*（functional gradient）。通过不断地计算 V^* 并用它来更新粒子，这些粒子便可越发准确地代表目标分布 p。

上述结果式 (4-1) 揭示了方向导数与动力学系统 V 之间的关系，接下来 SVGD 需要找到泛函梯度。注意到对于欧氏空间 $\mathbb{R}^m$，其在任意一点的切空间都等距同构于 $\mathbb{R}^m$，因此向量场 V 可以通过 m 个 $\mathbb{R}^m$ 上的光滑函数来描述。为得到闭式解，SVGD 将向量场 V 所在的空间限制在了直积空间 $\mathscr{H}^m$ 中，其中 $\mathscr{H}$ 是 $\mathbb{R}^m$ 上的一个核函数（kernel）K 的再生核希尔伯特空间（reproducing kernel Hilbert space, RKHS）（可参见 Aronszajn 的著作[161] 或 Steinwart 等人的著作[162] 第 4 章及定义 4.18）。这个 RKHS 空间 $\mathscr{H}$ 是 $\mathbb{R}^m$ 上一些特定函数所构成的希尔伯特空间，其最重要的性质是对于任意的 $x \in \mathbb{R}^m$，都有 $K(x,\cdot) \in \mathscr{H}$，并且 $\langle f(\cdot), K(x,\cdot)\rangle_{\mathscr{H}} = f(x), \forall f \in \mathscr{H}$。$\mathscr{H}^m$ 也可以看作是一个特殊的向量值 RKHS 空间[163]。这个选取方法可以看作是使用 m 个独立的来自 $\mathscr{H}$ 的函数来描述 $\mathbb{R}^m$ 上的向量场 V。将向量场 V 选在 $\mathscr{H}^m$ 中，SVGD 得到的泛函梯度为：

$$V^*(\cdot) = \mathbb{E}_{q_t(x)}[K(x,\cdot)\nabla \log p(x) + \nabla K(x,\cdot)]. \tag{4-2}$$

注意到粒子所代表的分布 q_t 只出现在期望中，因此对最优向量场 V^* 进行估计时只需要知道 q_t 的样本，即当下已有的粒子即可，而不需要假设 q_t 的形式。这也是使用 KL 散度来衡量两个分布之间差别的好处。有了 V^* 的计算方法，SVGD 便有了可行的算法。不过，本文下一章 5.3 节中将阐明，SVGD 其实也需要对 q_t 作一平滑性假设，只不过是在求解泛函梯度时将此平滑性要求转移到了向量场 V 上，并通过从向量值 RKHS 空间中选取 V 来实现。尽管如此，SVGD 使用非参数化粒子形式而带来的近似灵活性及其粒子高效性仍可使其成为值得推广和改进的方法。

4.3　黎曼-斯坦因变分梯度下降方法

本节将提出黎曼-斯坦因变分梯度下降方法（Riemannian Stein variational gradient descent, RSVGD）。它在 SVGD 方法之上会遇到由黎曼流形不同于欧氏空间的性质所带来的诸多技术问题，需要一些具有原创性的处理方法。本节将首先从目前已有的理论基础出发推导黎曼流形上的方向导数，然后设计一个具有原创性的求解泛函梯度的方法从而得到 RSVGD（在坐标空间中所表示）的算法，最后再推导出在流形嵌入空间中所表示的 RSVGD 的算法，并给出超球面上的特例用于处理具体实例中所考虑的场景。

4.3.1　方向导数

现在考虑一般黎曼流形 $\mathscr{M}$ 上的方向导数。这里首先推导出两个有用的结论，然后考察 KL 散度在黎曼流形上定义的合理性，最后得到 KL 散度关于向量场 V 方向导数的表达形式。

第一个有用的结论是连续性方程在黎曼流形上的推广。它揭示了在向量场 V 所引出的动力学系统（参见 2.1.1.3 节）的作用下，分布的演化规律与 V 的关系。它的证明需要使用黎曼流形上的雷诺输运定理（参见本章 4.2.1 节）。

引理 4.1 (**黎曼流形上的连续性方程**，continuity equation on Riemannian manifolds)：令 q_t 是在向量场 $V \in \mathscr{T}(\mathscr{M})$ 所引出的动力学系统下连续演化的概率分布。则下面的等式几乎处处成立：

$$\frac{\partial q_t}{\partial t} = -\operatorname{div}(q_t V) = -V[q_t] - q_t \operatorname{div}(V). \tag{4-3}$$

此等式在坐标系中的表达式为：

$$\frac{\partial q_t}{\partial t} = -V^i \partial_i q_t - q_t \partial_i V^i - q_t V^i \partial_i \log \sqrt{|G|}.$$

证明 令 $F_{(\cdot)}(\cdot)$ 为向量场 V 的流。对于任意紧致子集 $\mathscr{I} \subset \mathscr{M}$，考虑积分 $\int_{F_t(\mathscr{I})} q_t \omega_g$。由于任一在时刻 0 处于 $\mathscr{I}$ 中的粒子总会在时刻 t 处于 $F_t(\mathscr{I})$ 中并且反之亦然，上述积分，亦即时刻 t 时处于 $F_t(\mathscr{I})$ 中粒子的比例，等于时刻 0 时处于 $\mathscr{I}$ 中粒子的比例，因而这个积分是一个常量：$\dfrac{\mathrm{d}}{\mathrm{d}t}\displaystyle\int_{F_t(\mathscr{I})} q_t \omega_g = 0$。而另一方面，雷诺输运定理（参见本章 4.2.1 节）给出对于任意紧致子集 $\mathscr{I}$ 和时间 t，都有

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}\int_{F_t(\mathscr{I})} q_t \omega_g = \int_{F_t(\mathscr{I})} \left(\frac{\partial q_t}{\partial t} + \operatorname{div}(q_t V)\right) \omega_g,$$

几乎处处成立。由子集 $\mathscr{I}$ 的任意性可知，上式要成立，其被积函数必须几乎处处等于零。所以最终可以得到 $\dfrac{\partial q_t}{\partial t} + \operatorname{div}(q_t V) = 0$ 几乎处处成立，即引理中的结论。□

第二个有用的结论是在 $\mathscr{M}$ 上一个微分同胚（diffeomorphism，即可逆且光滑的 $\mathscr{M} \to \mathscr{M}$ 映射）的作用下一个分布的概率密度函数的变换规律。

引理 4.2 (微分同胚作用下的概率密度函数)：令 ϕ 是 $\mathscr{M}$ 上的一个保持定向（orientation-preserving）的微分同胚（diffeomorphism），x 是 $\mathscr{M}$ 上的一个随机变量，而 p 是 x 所服从的分布关于黎曼体积形式的概率密度函数。记 $\phi_{\#}p$ 为经 ϕ 变换后的随机变量 $\phi(x)$ 所服从的分布关于黎曼体积形式的概率密度函数（或称为分布 p 在映射 ϕ 下的前推（push-forward））。考虑 $\mathscr{M}$ 的任一局部坐标系 $(\mathscr{I}, \Phi)$，在其中有：

$$\phi_{\#}p = \frac{\left(p\sqrt{|G|}\right) \circ \phi^{-1}}{\sqrt{|G|}} \left|\operatorname{Jac} \phi^{-1}\right|, \tag{4-4}$$

其中 G 是此坐标系 $(\mathscr{I}, \Phi)$ 中的黎曼度量矩阵，$|G|$ 是其行列式，而 $\left|\operatorname{Jac} \phi^{-1}\right|$ 是映射 $\Phi \circ \phi^{-1} \circ \Phi^{-1} : \mathbb{R}^m \to \mathbb{R}^m$ 的雅可比行列式（Jacobian determinant）。此表达式是坐标不变的。

证明 令 $\mathscr{I}$ 是流形 $\mathscr{M}$ 的任一紧致子集，$(\mathscr{J}, \Phi)$（其中 $\mathscr{J} \subset \mathscr{I}$）是 $\mathscr{I}$ 的一个局部坐标系，且坐标形式为 $\{x^i\}_{i=1}^m$。一方面，由 $\phi_{\#}p$ 的定义，可知 $\operatorname{Prob}_p(\mathscr{I}) = \operatorname{Prob}_{\phi_{\#}p}(\phi(\mathscr{I}))$。而另一方面，可以调用流形上的全局变量替换定理（theorem of global change of variables on manifold，参见 Abraham 等人的著作[131] 定理 8.1.7），得到如下结果：

$$\begin{aligned}
&\operatorname{Prob}_p(\mathscr{I}) \\
=& \int_{\mathscr{I}} p\omega_g = \int_{\phi(\mathscr{I})} \phi^{-1*}(p\omega_g) = \int_{\phi(\mathscr{I})} (p \circ \phi^{-1}) \phi^{-1*}(\omega_g) \qquad (4\text{-}5) \\
=& \int_{\phi(\mathscr{I})} (p \circ \phi^{-1})(\sqrt{|G|} \circ \phi^{-1}) \left|\operatorname{Jac} \phi^{-1}\right| \mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^m \\
=& \int_{\phi(\mathscr{I})} \frac{\left(p\sqrt{|G|}\right) \circ \phi^{-1}}{\sqrt{|G|}} \left|\operatorname{Jac} \phi^{-1}\right| \omega_g \qquad (4\text{-}6) \\
=& \operatorname{Prob}_{\frac{\left(p\sqrt{|G|}\right) \circ \phi^{-1}}{\sqrt{|G|}} |\operatorname{Jac} \phi^{-1}|} (\phi(\mathscr{I})),
\end{aligned}$$

其中 $\phi^{-1*}(\cdot)$ 是微分同胚 ϕ^{-1} 在 $\mathscr{M}$ 的 m-形式上的拉回（pull-back）。结合这两个方面并注意到子集 $\mathscr{I}$ 的任意性，便可得到定理中给出的结果。 □

在正式推导 KL 散度的方向导数之前，首先考察 KL 散度在一般黎曼流形上定义的合理性。考虑黎曼流形 $\mathscr{M}$ 上两个分布，并将它们用关于某一体积形式 ω 的密度函数 q^ω 和 p^ω 来表示（参见 2.1.2.5 节）。定义这两个分布的 KL 散度为：

$$\mathrm{KL}(q||p) := \int_{\mathscr{M}} q^\omega \log(q^\omega/p^\omega)\omega.$$

此定义的合理性意味着它只与这两个分布有关，而与所选取的体积形式 ω 无关。为了说明这一点，可考察另一个体积形式 ω' 下此定义给出的结果是否与 ω 所给出的结果一样。由于对于流形 $\mathscr{M}$ 上任意一点 x 来说，$\omega(x)$ 和 $\omega'(x)$ 都在同一个 1 维线性空间（即点 x 处的 m 次外形式空间 $\wedge^k T_x^*\mathscr{M}$，参见 2.1.1.4 节），因此存在正实数 $c(x) \in \mathbb{R}^+$ 使得 $\omega'(x) = c(x)\omega(x)$。这种方式给出了一个 $\mathscr{M}$ 上的光滑函数 $c : \mathscr{M} \to \mathbb{R}^+$。由概率密度函数的定义，可知 $q^{\omega'} = q^\omega/c$。因此 $\int_{\mathscr{M}} q^{\omega'} \log(q^{\omega'}/p^{\omega'})\omega' = \int_{\mathscr{M}} \frac{q^\omega}{c} \log \frac{q^\omega/c}{p^\omega/c}\omega = \int_{\mathscr{M}} q^\omega \log(q^\omega/p^\omega)\omega$，即使用 ω 和 ω' 这两个不同的体积形式所给出的 KL 散度是一样的。这说明了此定义的合理性。

现在可以得到本章中第一个关键结论，即 KL 散度的方向导数关于向量场 V 的表达式。

定理 4.1 (方向导数，directional derivative)： 令 q_t 是在向量场 $V \in \mathscr{T}(\mathscr{M})$ 所引出的动力学系统下连续演化的概率分布，p 是一个固定的目标分布。则 KL 散度关于向量场 V 的方向导数可以表示为：

$$-\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(q_t||p) = \mathbb{E}_{q_t}[\mathrm{div}(pV)/p] = \mathbb{E}_{q_t}\big[V[\log p] + \mathrm{div}(V)\big].$$

证明 令 $F_{(\cdot)}(\cdot)$ 为向量场 V 的流。对于在向量场 V 所引出的动力学系统的作用下演化的分布 q_t，由其定义可知 $q_t = (F_t)_\# q_0$，其中概率分布的前推 $(F_t)_\#$ 的定义见引理 4.2。由于流具有性质 $F_{t_1+t_2} = F_{t_1} \circ F_{t_2} = F_{t_2} \circ F_{t_1}, \forall t_1, t_2 \in \mathbb{R}$，因此有 $q_{t_1+t_2} = (F_{t_1+t_2})_\# q_0 = (F_{t_1} \circ F_{t_2})_\# q_0 = (F_{t_1})_\#\big((F_{t_2})_\# q_0\big) = (F_{t_1})_\# q_{t_2}$。除 q_t 外，推导过程中还需要另外一个在 V 下演化的分布 p_t，其在某一时刻 t_0 时刚好等于目标分布 p。有了这些准备，便可开始最终的推导：

$$-\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=t_0} \mathrm{KL}(q_t||p) = -\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \int_{\mathscr{M}} q_{t_0+t} \log \frac{q_{t_0+t}}{p_{t_0}}\omega_g$$

（将 q_{t_0+t} 视为 $(F_t)_\# q_{t_0}$ 并应用式 (4-4)）

$$= -\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \int_{\mathscr{M}} \frac{\big(q_{t_0}\sqrt{|G|}\big) \circ F_t^{-1}}{\sqrt{|G|}} \big|\mathrm{Jac}\, F_t^{-1}\big|$$

$$\cdot \left(\log \frac{\left(q_{t_0}\sqrt{|G|}\right) \circ F_t^{-1}}{\sqrt{|G|}} + \log\left|\mathrm{Jac}\, F_t^{-1}\right| - \log p_{t_0} \right) \omega_g$$

（将 F_t^{-1} 作用在整个积分上并应用全局变量替换定理式 (4-5)）

$$\begin{aligned} = & -\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \int_{F_t^{-1}(\mathscr{M})} \left(\left[\frac{\left(q_{t_0}\sqrt{|G|}\right) \circ F_t^{-1}}{\sqrt{|G|}} \left|\mathrm{Jac}\, F_t^{-1}\right| \right.\right. \\ & \left.\left. \cdot \left(\log \frac{\left(q_{t_0}\sqrt{|G|}\right) \circ F_t^{-1}}{\sqrt{|G|}} + \log\left|\mathrm{Jac}\, F_t^{-1}\right| - \log p_{t_0} \right) \right] \circ F_t \right) F_t^*(\omega_g) \end{aligned}$$

（由于 F_t^{-1} 是 $\mathscr{M}$ 上的一个微分同胚，因而 $F_t^{-1}(\mathscr{M}) = \mathscr{M}$；另外 $\left|\mathrm{Jac}\, F_t^{-1}\right| \circ F_t = |\mathrm{Jac}\, F_t|^{-1}$；参见式 (4-6) 可得 ω_g 在 F_t 下的拉回 $F_t^*(\omega_g)$ 的表达式）

$$\begin{aligned} = & -\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \int_{\mathscr{M}} \frac{q_{t_0}\sqrt{|G|}}{\sqrt{|G|} \circ F_t} |\mathrm{Jac}\, F_t|^{-1} \cdot \left(\log \frac{q_{t_0}\sqrt{|G|}}{\sqrt{|G|} \circ F_t} \right. \\ & \left. - \log|\mathrm{Jac}\, F_t| - \log(p_{t_0} \circ F_t) \right) \cdot \frac{\sqrt{|G|} \circ F_t}{\sqrt{|G|}} |\mathrm{Jac}\, F_t| \omega_g \end{aligned}$$

（整理各项）

$$= -\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \int_{\mathscr{M}} q_{t_0} \left[\log q_{t_0} - \log \left(\frac{\left(p_{t_0}\sqrt{|G|}\right) \circ F_t}{\sqrt{|G|}} |\mathrm{Jac}\, F_t| \right) \right] \omega_g$$

（应用流的性质 $F_t = F_{-t}^{-1}$；将 p_{t_0-t} 视为 $(F_{-t})_{\#} p_{t_0}$ 并逆向应用式 (4-4)）

$$= -\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \int_{\mathscr{M}} q_{t_0} \left[\log q_{t_0} - \log p_{t_0-t} \right] \omega_g$$

（$\mathscr{M}$ 不随时间 t 而变化（否则一个在边界上的积分将会出现））

$$= \int_{\mathscr{M}} q_{t_0} \left. \frac{\partial}{\partial t} (\log p_{t_0-t}) \right|_{t=0} \omega_g$$

（由流的齐次性可知，$\tilde{F}_t(\cdot) := F_{-t}(\cdot)$ 正是向量场 $-V$ 的流；参考式 (4-3) 可知 $\left.\frac{\partial p_{t_0-t}}{\partial t}\right|_{t=0} = -\operatorname{div}(p_{t_0}(-V)) = \operatorname{div}(p_{t_0}V) = -\left.\frac{\partial p_{t_0+t}}{\partial t}\right|_{t=0}$）

$$= -\int_{\mathscr{M}} q_{t_0} \left. \frac{\partial}{\partial t} (\log p_{t_0+t}) \right|_{t=0} \omega_g$$

（参照式 (4-3)）

$$= \int_{\mathscr{M}} (q_{t_0}/p_{t_0}) \operatorname{div}(p_{t_0}V) \omega_g = \mathbb{E}_{q_{t_0}}[\operatorname{div}(p_{t_0}V)/p_{t_0}]$$

（散度的性质）

$$= \mathbb{E}_{q_{t_0}}\big[V[\log p_{t_0}] + \mathrm{div}(V)\big].$$

由 t_0 的任意性可将最后的结果表示为定理中的形式。 □

定理 4.1 是 SVGD 中的关键结论式 (4-1) 在一般黎曼流形上的拓展。参照 SVGD 中的术语，这里可将 $\mathcal{A}_p V := V[\log p] + \mathrm{div}(V)$ 称作一般化斯坦因算符（generalized Stein's operator）。

最后，作为对上述定理 4.1 的补充，这里对黎曼流形情况下斯坦因恒等式（Stein's identity）成立的条件进行讨论。斯坦因恒等式是指 $-\dfrac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(q_t||p)\Big|_{t=t_0} = 0$，其中 q_t 满足 $q_{t_0} = p$，而斯坦因类（Stein class）是指所有使得斯坦因恒等式成立的向量场 V 的集合。斯坦因恒等式的意义在于，当 q_t 达到最优即 $q_t = p$ 时，沿着斯坦因类中任一向量场 V 的方向导数 $-\dfrac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(q_t||p)$ 都等于零，这类似于优化领域中人们所熟知的梯度等于零的最优条件。这些概念在使用泛函梯度的模来衡量两个分布之间差异的方法，也即斯坦因差异量方法（Stein discrepancy）[164] 中都发挥着重要作用。

现在推导在一般的黎曼流形的情况下斯坦因恒等式成立的条件，亦即斯坦因类。记 $\partial\mathscr{M}$ 为 m 维流形 $\mathscr{M}$ 的闭包的边界。它要么是空集（$\mathscr{M}$ 无边界的情况，例如超球面），要么是一个 $m-1$ 维流形。由定理 4.1，斯坦因恒等式成立这个条件即为 $\mathbb{E}_p[\mathrm{div}(pV)/p] = 0$。使用高斯定理（Gauss' theorem，参见 Abraham 等人的著作[131] 定理 8.2.9）可将此式变形为：

$$\int_{\mathscr{M}} \mathrm{div}(pV)\omega_g = \int_{\partial\mathscr{M}} \mathrm{i}_{(pV)}\omega_g = \sum_{i=1}^{m}\int_{\partial\mathscr{M}} p\sqrt{|G|}(-1)^{i+1}V^i\,\mathrm{d}x^{\neg i},$$

其中 $\mathrm{i}_V : \mathscr{A}^k(\mathscr{M}) \to \mathscr{A}^{k-1}(\mathscr{M})$ 称为内向积（interior product）或缩并（contraction），定义为 $\big(\mathrm{i}_V\omega\big)(x)[v_1,\cdots,v_{k-1}] = \omega(x)[V(x),v_1,\cdots,v_{k-1}]$，而 $\mathrm{d}x^{\neg i} := \mathrm{d}x^1\wedge\cdots\wedge\mathrm{d}x^{i-1}\wedge\mathrm{d}x^{i+1}\wedge\cdots\wedge\mathrm{d}x^m$（其中 $1\leqslant i\leqslant m$）是 $\partial\mathscr{M}$ 上的一个体积形式，其中 "$\wedge$" 表示外向积（exterior product）或称楔积（wedge product）（参见 2.1.1.4 节）。

对于像超球面这样的流形，其边界 $\partial\mathscr{M}$ 是空集，因而上面的积分总是零。因此对这些流形来说，所有的向量场都可满足斯坦因恒等式，因而斯坦因类就是整个向量场空间 $\mathscr{T}(\mathscr{M})$。对于闭包具有边界的流形，其 $\partial\mathscr{M}$ 不是空集。由边界的性质，在边界 $\partial\mathscr{M}$ 上任意一点附近，都存在一个 $\mathscr{M}$ 的局部坐标系 $(\mathscr{I},\Phi)$，其坐标表示 $(y^1,\cdots,y^m)$ 满足 $\forall x\in\partial\mathscr{M}\cap\mathscr{I}, y^m(x) = 0$。对于这样的坐标系，可有结论 $\mathrm{d}y^m = 0$，并且还可以得到一个 $\partial\mathscr{M}$ 的局部坐标系 $(\partial\mathscr{M}\cap\mathscr{I},\Psi = (y^1,\cdots,y^{m-1}))$。

在这个 $\partial\mathscr{M}$ 的局部坐标系 $(\partial\mathscr{M}\cap\mathscr{I},\Psi)$ 中，斯坦因恒等式成立的条件就变为：

$$\int_{\partial\mathscr{M}} p\tilde{V}^m\sqrt{|\tilde{G}|}\,\mathrm{d}y^{\neg m}=\int_{\partial\mathscr{M}} p\tilde{V}^m\sqrt{|\tilde{G}|}\,\mathrm{d}y^1\wedge\cdots\wedge\mathrm{d}y^{m-1}=0, \tag{4-7}$$

其中 $\tilde{G}$ 是 $\mathscr{M}$ 的在其局部坐标系 $(\mathscr{I},\Phi)$ 中的黎曼度量矩阵，而 $\tilde{V}^m$ 是向量场 V 在 $\mathscr{M}$ 的局部坐标系 $(\mathscr{I},\Phi)$ 中的第 m 个分量。此时的斯坦因类即为所有满足式 (4-7) 的向量场的集合。注意这个集合是与坐标系 $(\mathscr{I},\Phi)$ 的选取是无关的。

对于 $\mathscr{M}$ 是欧氏空间 $\mathbb{R}^m$ 的一个开子集这个特例，考虑在其闭包边界 $\partial\mathscr{M}$ 上的任意一点 x 附近选取局部坐标系 $(\mathscr{I},\Phi)$ 满足 $y^m=0$ 并且自然基底（natural basis）$\left\{\tilde{\partial}_i\,\middle|\,\tilde{\partial}_i:=\dfrac{\partial}{\partial y^i},i=1,\cdots,m\right\}$ 是标准正交的。在此坐标系下，可有 $|\tilde{G}(x)|=1$，并且单位向量 $\tilde{\partial}_m$ 正交于向量组 $\{\tilde{\partial}_1,\cdots,\tilde{\partial}_{m-1}\}$ 所张成的 $m-1$ 维线性空间。由于 $(y^1,\cdots,y^{m-1})$ 是 $\partial\mathscr{M}$ 的一个局部坐标系，因而上述这个 $m-1$ 维线性空间就是 $\partial\mathscr{M}$ 在 x 处的切空间。与之正交的向量 $\tilde{\partial}_m$ 即为这个切空间的单位法向量。因此可将它表示为欧氏空间中常见的形式 $\vec{n}$，并且可以被更加形象地解释为 $\partial\mathscr{M}$ 在 x 处的法向量。进而，$\tilde{V}^m$ 就是向量场 V 在 $\tilde{\partial}_m$ 方向上的分量。此处也将它表示为欧氏空间中常见的形式 $\vec{V}\cdot\vec{n}$。进一步，可将 $\partial\mathscr{M}$ 上的黎曼体积形式 $\sqrt{|\tilde{G}|}\,\mathrm{d}y^{\neg m}=\sqrt{|\tilde{G}|}\,\mathrm{d}y^1\wedge\cdots\wedge\mathrm{d}y^{m-1}=\mathrm{d}y^1\wedge\cdots\wedge\mathrm{d}y^{m-1}$ 表示为欧氏空间中常见的形式 $\mathrm{d}S$，即“曲面 $\partial\mathscr{M}$ 上的面积元”。而在此情况下 $\partial\mathscr{M}$ 通常是 $\mathbb{R}^m$ 中若干闭合的 $m-1$ 维曲面的并集，因而在 $\partial\mathscr{M}$ 上的积分号可写作闭合积分 $\oint$。综合这些分析和处理，便可写出斯坦因恒等式成立的条件，即式 (4-7)，在欧氏空间这个特例下的表达形式为：

$$\oint_{\partial\mathscr{M}} p\vec{V}\cdot\vec{n}\,\mathrm{d}S=0.$$

这与已有的结论（例如 [81]）相符。而本节所推得的式 (4-7) 则将此条件推广到了一般的黎曼流形上。

4.3.2　泛函梯度

定理 4.1 给出了方向导数 $-\dfrac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(q_t||p)$ 与向量场 V 的显式关系。现在考虑最大化此方向导数以确定泛函梯度，亦即求解下面的优化问题：

$$V^*:=\max_{V\in\mathfrak{X},\|V\|_{\mathfrak{X}}=1}\cdot\operatorname{argmax}\mathcal{J}(V):=\mathbb{E}_q\big[V[\log p]+\operatorname{div}(V)\big], \tag{4-8}$$

其中 q 的下标 t 被省略掉了，因为求解最优的 V 是针对一个固定时刻而言的。注意到在上面的优化问题中向量场所在的集合被限定在了赋范空间 $\mathfrak{X}$ 上。这个空间

在理想情况下应该为流形 $\mathscr{M}$ 上的整个向量场空间 $\mathscr{T}(\mathscr{M})$，然而这样得到的解很难使用粒子——亦即分布 V 的样本——进行估计。因此这里希望选择 $\mathscr{T}(\mathscr{M})$ 的一个合适的子集 $\mathfrak{X}$ 使得 V^* 可以自然地使用粒子进行估计。

对 $\mathfrak{X}$ 的要求　在考虑如何选择具体的向量场子空间 $\mathfrak{X}$ 之前，此处首先列出一个合适的 $\mathfrak{X}$ 所应满足的要求。由于一个具体的 $\mathfrak{X}$ 对应着一个最优解 V^*，因而对 $\mathfrak{X}$ 的要求可以通过对 V^* 的要求来表示。

- 要求 1：V^* 是流形 $\mathscr{M}$ 上的一个合法的向量场；
- 要求 2：V^* 具有坐标不变性；
- 要求 3：V^* 具有闭形式，并且闭形式中分布 q 只以期望的形式出现。

要求 3 即是如上所述的考虑 $\mathscr{T}(\mathscr{M})$ 子集的动机。闭形式使得 V^* 可方便地计算，而若 q 在闭形式中只以期望的形式出现，那么在估计 V^* 时就可以不需要 q 的密度函数的显式表达式，而只需要其样本亦即粒子即可通过计算样本均值的方式来估计。SVGD 方法中所选取的向量场子空间 $\mathscr{H}^m$ 可满足此要求。

前两个要求来自于一般黎曼流形与欧氏空间本质上不同的性质，这些不同之处使得 SVGD 所使用的技术和考量不再适用于黎曼流形的情况。需要说明的是，如果将 $\mathfrak{X}$ 选为整个向量场空间 $\mathscr{T}(\mathscr{M})$，那么要求 1 和 2 都可自动满足。但这里考虑的是将 $\mathfrak{X}$ 选为一个子空间，因而可能会出现不满足这两个要求的情况，例如 SVGD 中所选择的 $\mathfrak{X} = \mathscr{H}^m$。

要求 1 的必要性很显然，因为本章的推导都是基于向量场。但需要强调的是，这个条件并不是很轻易就可满足的。例如在欧氏空间 $\mathbb{R}^m$ 的情况下，由 m 个相互独立的光滑函数所描述的向量场 $(f^1, \cdots, f^m)$ 即是合法的。但是在一般的 m 维黎曼流形上，任意坐标系中 m 个分量都是光滑的这一条件并不足够保证一个向量场的表示是合法的。偶数维超球面就是一个常见的例子，因为由毛球定理（hairy ball theorem；参见 Abraham 等人的著作[131] 定理 8.5.13），其上的向量场必有一个零点（关键点，critical point）。这个条件是坐标系中 m 个光滑函数这种表示所无法保证的。因而 SVGD 中的选择 $\mathfrak{X} = \mathscr{H}^m$ 无法用于一般的黎曼流形。

要求 2 的提出是为了避免最优解 V^* 的歧义性和任意性。坐标不变性是微分流形领域中一个重要的概念。由于流形都可以局部等同于一个欧氏空间，即其局部坐标系，因此流形上的概念总可以通过它在坐标系中的形式来表示。但通常流形上的一个概念是通过一个具有明确含义的原则来定义的，而坐标系的选择却具有一定的自由度和任意性，因此这个概念在不同坐标系中的表示形式应该给出相同的结果，或者说这个概念的坐标表示形式在坐标变换下应该保持不变。这种坐标表示形式被称作具有坐标不变性。若一个坐标表达式没有坐标不变性，那么在两

个坐标系 $(\mathscr{I},\Phi)$ 和 $(\mathscr{J},\Psi)$ 相交的区域上，两个坐标系将会给出两个不同的结果，这样就会造成在 $\mathscr{I}\cap\mathscr{J}$ 上这个坐标表达式所代表的概念具有歧义性。而如果考虑选取一个特定的坐标系使得这个没有坐标不变性的表达式给出一个确定的概念，那么这种做法则会带来这个概念的任意性。因为对于一般的流形来说，它的任何坐标系都是等价的，即具有相同的地位。不存在一个客观的方式来说明某一个坐标系会比其他坐标系特殊。因此选定一个“特定”的坐标系就具有任意性。另外，对于超球面这样没有全局坐标系的流形来说，这种做法也很难实现：选取一个坐标系是不够的，而选取多个坐标系又会在这些坐标系相交的区域产生歧义性。因此，具有坐标不变性的坐标表达式才能反映流形上的一个合理的概念。

一般黎曼流形上梯度的坐标表达式 $\operatorname{grad} f=g^{ij}\partial_i f\partial_j$ 是坐标不变的，而欧氏空间中常见的梯度表达式 $\nabla f:=\sum_i \partial_i f\partial_i$ 则不是坐标不变的。为说明这一论点，考虑在流形上某一点附近的两个局部坐标系 $(\mathscr{I},\{x^i\}_{i=1}^m)$ 和 $(\mathscr{J},\{y^a\}_{a=1}^m)$，并将两个坐标系中的黎曼度量和自然基底分别记为 $g_{ij},\tilde{g}_{ab}$ 以及 $\{\partial_i\}_{i=1}^m,\{\tilde{\partial}_a\}_{a=1}^m$。在两个坐标系相交的区域 $\mathscr{I}\cap\mathscr{J}$，$g^{ij}\partial_i f\partial_j=\tilde{g}^{ab}\tilde{\partial}_a f\tilde{\partial}_b$，即两个坐标系中的表达式给出了同样的结果，但是 $\sum_i \partial_i f\partial_i=\sum_a\left(\sum_{i,b}\frac{\partial y^a}{\partial x^i}\frac{\partial y^b}{\partial x^i}\tilde{\partial}_b f\right)\tilde{\partial}_a\neq\sum_a\tilde{\partial}_a f\tilde{\partial}_a$，即两个坐标系中的表达式给出了不同的结果。因而只有将梯度的表达式推广为 $g^{ij}\partial_i f\partial_j$ 才可正确地将梯度这个概念推广到一般的黎曼流形上。（当然这个坐标表达式是根据定义梯度的原则而推导出来的，参见 2.1.2.2 节；这里只是在强调一个合理的坐标表达式应具有坐标不变性，以及欧氏空间中的表达式不一定能够直接通过坐标表达式推广到一般的黎曼流形上。）类似地，向量场的散度 $\operatorname{div} V=\partial_i(\sqrt{|G|}V^i)/\sqrt{|G|}$ 具有坐标不变的坐标表达式，而欧氏空间中常见的散度表达式 $\partial_i V^i=\frac{\partial y^a}{\partial x^i}\tilde{\partial}_a\left(\frac{\partial x^i}{\partial y^b}\tilde{V}^b\right)\neq\tilde{\partial}_a\tilde{V}^a$ 则并不是坐标不变的。（不过，向量场在函数上的作用 $V[f]=V^i\partial_i f=\tilde{V}^a\tilde{\partial}_a f$ 是坐标不变的，并且这个形式也就是欧氏空间中常见的表达式。）特别要说明的是，SVGD 所得到的泛函梯度式 (4-2) 不是坐标不变的：

$$\begin{aligned}V^*_{\text{SVGD}}(\cdot):=&\sum_{i=1}^m\mathbb{E}_{q(x)}[K(x,\cdot)\partial_{x^i}\log p(x)+\partial_{x^i}K(x,\cdot)]\partial_{x^i}\\=&\sum_{a=1}^m\sum_{i=1}^m\frac{\partial y^a}{\partial x^i}(\cdot)\mathbb{E}_{q(x)}\Big[K(x,\cdot)\frac{\partial y^b}{\partial x^i}(x)\tilde{\partial}_b\log p(x)+\frac{\partial y^b}{\partial x^i}(x)\tilde{\partial}_b K(x,\cdot)\Big]\tilde{\partial}_a\\\neq&\sum_{a=1}^m\mathbb{E}_{q(x)}[K(x,\cdot)\tilde{\partial}_a\log p(x)+\tilde{\partial}_a K(x,\cdot)]\tilde{\partial}_a.\end{aligned}$$

因而 SVGD 中的选择 $\mathfrak{X}=\mathscr{H}^m$ 从这个角度来说也无法用于一般的黎曼流形。

在展示本章的解决方案之前，首先考虑一些其他可能的方法，并说明这些方

法也无法满足上述要求。可以注意到 $\mathfrak{X}$ 是向量场空间 $\mathscr{T}(\mathscr{M})$ 的子空间，而这个线性空间既可以看作是实数域 $\mathbb{R}$ 上的，也可以看作是函数域 $C^\infty(\mathscr{M})$ 上的。但无论何种视角，$\mathscr{T}(\mathscr{M})$ 都是无限维的，因而难以对它进行表示。首先，可以考虑使用每点一个 m 维向量的方式来表示 $\mathscr{M}$ 上的一个向量场，但这种方式很容易不满足上述要求 1 和要求 2。这两个要求都是涉及流形上全局的性质，而这种表示方式却很难控制整个向量场的全局特征。SVGD 的处理方式就属于这一类，但上面的分析表明这种方式不能用于一般的黎曼流形。其次，可以考虑使用一个的关键点上的切向量通过平行移动（parallel transport）给出其他点上的切向量的方式来表示一个向量场。但这种方式使用起来会十分繁琐，而且关键点的选取具有任意性，从而不满足要求 2。第三种可能的方法是将向量场看作从流形 $\mathscr{M}$ 到其切丛（tangent bundle）$T\mathscr{M}$ 的映射，然后使用一个向量值函数来表示此映射。但是切丛 $T\mathscr{M}$ 通常来说也是一个流形而不是线性空间，因而难以使用学习向量值函数的方法（例如 [163]）寻找最优向量场。

解决方案　本节首先介绍所提解决方案，然后说明它们满足上面三个要求。注意核函数和 RKHS 空间的理论仍然适用于黎曼流形上的情况（例如 Steinwart 等人的著作[162] 第 4 章），因为这些理论都建立在比黎曼流形更加广泛的度量空间之上。令 $K : \mathscr{M} \times \mathscr{M} \to \mathbb{R}$ 是流形 $\mathscr{M}$ 上的一个光滑核函数，并令 $\mathscr{H}$ 表示其 RKHS 空间。由 Steinwart 等人的著作[162] 引理 4.3 以及流形坐标映射的光滑性（它们是微分同胚）可知，K 在流形的任一坐标空间上也是一个光滑核函数。这里要求所选的核函数 K 满足如下条件：零函数是其 RKHS 空间 $\mathscr{H}$ 中唯一的常值函数。常见的核函数例如高斯核函数（Gaussian kernel）是满足这个要求的（参见 Steinwart 等人的著作[162] 推论 4.44）。

本节提出的解决方案选择下面的向量场子空间：

$$\mathfrak{X} = \{ \operatorname{grad} f \mid f \in \mathscr{H} \}, \tag{4-9}$$

其中光滑函数的梯度 $\operatorname{grad} : C^\infty(\mathscr{M}) \to \mathscr{T}(\mathscr{M})$ 总是一个合法的向量场。在坐标系中，$\operatorname{grad} f = g^{ij}\partial_i f \partial_j$，其中 g^{ij} 是黎曼度量矩阵的逆矩阵 G^{-1} 的第 (i, j) 个元素。下面这个结论说明这样的空间 $\mathfrak{X}$ 可以是一个内积空间，因而也是赋范空间：

引理 4.3：　在一个合适的内积下，式 (4-9) 所定义的空间 $\mathfrak{X}$ 等距同构（isometrically isomorphic）于 RKHS 空间 $\mathscr{H}$。因而在此内积下，$\mathfrak{X}$ 是一个希尔伯特空间。

证明 定义映射 $\iota : \mathscr{H} \to \mathfrak{X}, f \mapsto \operatorname{grad} f$。由 RKHS 空间的线性性，可以验证这个映射是线性的：$\forall a, b \in \mathbb{R}, f, h \in \mathscr{H}, \iota(af + bh) = a \operatorname{grad} f + b \operatorname{grad} h = a\iota(f) + b\iota(h)$。

另外，对于满足条件 $\iota(f) = \iota(h)$ 的 RKHS 中的任意两个函数 $f, h \in \mathscr{H}$，都有 $\operatorname{grad}(f-h) = 0$，因而 $f-h$ 是 $\mathscr{H}$ 中的常函数，而由本节对所考虑的核函数 K 的假设，这个常函数一定是零函数。于是可以得到 $f = h$，因而 ι 是单射。而由 $\mathfrak{X}$ 的定义，映射 ι 一定是满射。因而它是 $\mathfrak{X}$ 和 $\mathscr{H}$ 之间的一个同构（isomorphism）。

进一步，可以为线性空间 $\mathfrak{X}$ 定义如下内积：

$$\langle \cdot, \cdot \rangle_{\mathfrak{X}} : \langle V, U \rangle_{\mathfrak{X}} = \left\langle \iota^{-1}(V), \iota^{-1}(U) \right\rangle_{\mathscr{H}}, \forall V, U \in \mathfrak{X}.$$

这个定义给出了一个合法的内积，因为 ι 是线性的并且 $\langle \cdot, \cdot \rangle_{\mathscr{H}}$ 是一个内积。由此内积的构造方式可以得知，ι 是这两个空间 $\mathfrak{X}$ 和 $\mathscr{H}$ 之间的等距同构（isometric isomorphism）。由于 $\mathscr{H}$ 是一个希尔伯特空间，因此 $\mathfrak{X}$ 在此内积下也是一个希尔伯特空间。 □

接下来展示本章中第二个关键结论，即式 (4-8) 中的目标函数 $\mathcal{J}(V)$ 可以写为 $\mathfrak{X}$ 中一个内积的形式，进而可以得到泛函梯度。

定理 4.2 (泛函梯度，functional gradient)：对于式 (4-9) 及引理 4.3 中所定义的希尔伯特空间 $(\mathfrak{X}, \langle \cdot, \cdot \rangle_{\mathfrak{X}})$ 以及式 (4-8) 中定义的目标函数 $\mathcal{J}$，可有结论 $\mathcal{J}(V) = \langle V, \operatorname{grad} f^* \rangle_{\mathfrak{X}}$，其中

$$f^*(\cdot) = \mathbb{E}_{q(x)} \Big[\big(\operatorname{grad} K(x, \cdot) \big) [\log p(x)] + \Delta K(x, \cdot) \Big], \tag{4-10}$$

以及 $\Delta f := \operatorname{div}(\operatorname{grad} f) = \partial_i \left(\sqrt{|G|} g^{ij} \partial_j f \right) / \sqrt{|G|}$ 是贝尔特拉米-拉普拉斯算符（Beltrami-Laplace operator）。特别地，优化问题 (4-8) 的最优解为：

$$\begin{aligned} V^*(x') &:= \max_{V \in \mathfrak{X}, \|V\|_{\mathfrak{X}} = 1} \cdot \operatorname{argmax} \mathcal{J}(V) \\ &= \operatorname{grad} f^*(x') \qquad (4\text{-}11a) \\ &= g'^{ab} \partial'_b \mathbb{E}_q \left[\left(g^{ij} \partial_j \log(p\sqrt{|G|}) + \partial_j g^{ij} \right) \partial_i K(x, x') + g^{ij} \partial_i \partial_j K(x, x') \right], \qquad (4\text{-}11b) \end{aligned}$$

其中带有撇上标 “′” 的符号表示它是以 x' 为自变量的，而其他符号是以 x 为自变量的。

证明 对于任意 $V \in \mathfrak{X}$，令 $f = \iota^{-1}(V)$（其中等距同构 $\iota : \mathscr{H} \to \mathfrak{X}$ 在引理 4.3 的证明中给出），亦即 f 是 $\mathscr{H}$ 中唯一使得 $V = \operatorname{grad} f$ 成立的函数。因此 V 可以在坐标系中表示为 $V = g^{ij} \partial_i f \partial_j$，并且可将式 (4-8) 中的目标函数 $\mathcal{J}(V)$ 在坐标系中表示：

$$\mathcal{J}(V) := \mathbb{E}_q \left[V[\log p] + \operatorname{div}(V) \right]$$

$$
\begin{aligned}
&= \mathbb{E}_q\left[V^j \partial_j \log(p\sqrt{|G|}) + \partial_j V^j\right] \\
&= \mathbb{E}_q\left[g^{ij}\partial_i f \partial_j \log(p\sqrt{|G|}) + \partial_j(g^{ij}\partial_i f)\right] \\
&= \mathbb{E}_q\left[\left(g^{ij}\partial_j \log(p\sqrt{|G|}) + \partial_j g^{ij}\right)\partial_i f + g^{ij}\partial_i\partial_i f\right].
\end{aligned}
$$

接下来的推导使用 Zhou 等人[165] 所给出的结论：对于任意 $x \in \mathscr{M}$，都有 $\partial_i K(x,\cdot), \partial_i\partial_j K(x,\cdot) \in \mathscr{H}$，且对于任意 $f \in \mathscr{H}$，都有 $\langle f(\cdot), \partial_i K(x,\cdot)\rangle_{\mathscr{H}} = \partial_i f(x)$ 以及 $\langle f(\cdot), \partial_i\partial_j K(x,\cdot)\rangle_{\mathscr{H}} = \partial_i\partial_j f(x)$。这样一来，上面的式子可以改写为：

$$
\begin{aligned}
\mathcal{J}(V) &= \mathbb{E}_q\Big[\left(g^{ij}\partial_j \log(p\sqrt{|G|}) + \partial_j g^{ij}\right)\langle f(\cdot), \partial_i K(x,\cdot)\rangle_{\mathscr{H}} + g^{ij}\langle f(\cdot), \partial_i\partial_j K(x,\cdot)\rangle_{\mathscr{H}}\Big] \\
&= \mathbb{E}_q\Big[\Big\langle f(\cdot), \left(g^{ij}\partial_j \log(p\sqrt{|G|}) + \partial_j g^{ij}\right)\partial_i K(x,\cdot) + g^{ij}\partial_i\partial_j K(x,\cdot)\Big\rangle_{\mathscr{H}}\Big] \\
&= \Big\langle f(\cdot), \mathbb{E}_q\Big[\left(g^{ij}\partial_j \log(p\sqrt{|G|}) + \partial_j g^{ij}\right)\partial_i K(A,\cdot) + g^{ij}\partial_i\partial_j K(x,\cdot)\Big]\Big\rangle_{\mathscr{H}},
\end{aligned}
$$

其中所有的函数、微分和期望若非专门指明都是关于变量 x 的。为简化此结果，定义函数 f^* 如下：

$$
\begin{aligned}
f^*(\cdot) &:= \mathbb{E}_q\Big[\left(g^{ij}\partial_j \log(p\sqrt{|G|}) + \partial_j g^{ij}\right)\partial_i K(x,\cdot) + g^{ij}\partial_i\partial_j K(x,\cdot)\Big] \\
&= \mathbb{E}_q\Big[g^{ij}\partial_j \log(p\sqrt{|G|})\partial_i K(x,\cdot) + \partial_j\big(\sqrt{|G|}g^{ij}\partial_i K(x,\cdot)\big)/\sqrt{|G|}\Big] \\
&= \mathbb{E}_q\Big[g^{ij}\partial_j \log(p\sqrt{|G|})\partial_i K(x,\cdot) + \Delta K(x,\cdot)\Big] \\
&= \mathbb{E}_q\Big[\big(\operatorname{grad} K(x,\cdot)\big)[\log p(x)] + \Delta K(x,\cdot)\Big],
\end{aligned}
$$

即式 (4-10) 中所给出的函数。使用函数 f^*，可以将目标函数表示为：

$$
\mathcal{J}(V) = \langle f(\cdot), f^*(\cdot)\rangle_{\mathscr{H}}.
$$

而由于 $\mathscr{H}$ 与 $\mathfrak{X}$ 是等距同构的，因而上面在 $\mathscr{H}$ 中的内积也可以在 $\mathfrak{X}$ 中表示，即为

$$
\mathcal{J}(V) = \langle \operatorname{grad} f, \operatorname{grad} f^*\rangle_{\mathfrak{X}} = \langle V, V^*\rangle_{\mathfrak{X}}.
$$

而由线性代数中的结论可知，$\|V^*\|_{\mathfrak{X}} = \max\limits_{V\in\mathfrak{X},\|V\|_{\mathfrak{X}}=1}\langle V, V^*\rangle_{\mathfrak{X}}$，且 $\dfrac{V^*}{\|V^*\|_{\mathfrak{X}}} = \operatorname*{argmax}\limits_{V\in\mathfrak{X},\|V\|_{\mathfrak{X}}=1}\langle V, V^*\rangle_{\mathfrak{X}}$。因而优化问题 (4-8) 所定义的最优解 $\max\limits_{V\in\mathfrak{X},\|V\|_{\mathfrak{X}}=1}\cdot \operatorname*{argmax}\limits_{V\in\mathfrak{X},\|V\|_{\mathfrak{X}}=1}\langle V, V^*\rangle_{\mathfrak{X}}$ 正是 $\|V^*\|_{\mathfrak{X}}\dfrac{V^*}{\|V^*\|_{\mathfrak{X}}} = V^*$，即为泛函梯度。至此定理得证。 □

这里所给出的最优解，即泛函梯度 V^*，是满足上面提出的 3 个要求的。参见式 (4-10) 和式 (4-11a)，由于梯度 grad、散度 div、贝尔特拉米-拉普拉斯算符 Δ、期望

以及向量场在光滑函数上的作用 $V[f]$ 都是流形上具有坐标不变性的概念，因此 f^* 以及其梯度 V^* 也都是流形上坐标不变的概念，自然满足要求 2。而由于光滑函数的梯度总是一个合法的向量场，因此 V^* 满足要求 1。参见式 (4-10) 和式 (4-11b) 的形式，可以发现要求 3 也显然可以得到满足。

作为对上述定理 4.2 的一点补充，现考虑优化问题 (4-8) 的最优值：

$$\begin{aligned}\mathcal{J}(V^*) = \mathbb{E}_q \mathbb{E}_{q'}\Big[\big(\operatorname{grad}' \log p'\big)\big[(\operatorname{grad} \log p)[K]\big] + \Delta'\Delta K \\ + (\operatorname{grad}' \log p')[\Delta K] + (\operatorname{grad} \log p)[\Delta' K]\Big],\end{aligned}$$

其中 $K = K(x, x')$，以及所有带有撇上标 “$'$” 的符号以 x' 为自变量，其他符号以 x 为自变量。作为对核化斯坦因差异量（kernelized Stein discrepancy, KSD）[164,166] 在一般黎曼流形上的推广，这个最优值可以称作分布 q 与 p 之间的黎曼-核化斯坦因差异量（Riemannian kernelized Stein discrepancy, RKSD）。与 KSD 类似，它也是一种积分类概率度量（integral probability metric）[167]，并可用于衡量两个分布之间的差异。

式 (4-11b) 给出了泛函梯度 V^* 这个向量场的坐标表达式。利用此式即可在坐标系中对相应的动力学系统进行模拟。具体地，对于一组粒子 $\{x^{(l)}\}_{l=1}^L$，它们的更新方式为：

$$\begin{aligned}x^{(l')} \leftarrow & x^{(l')} + \frac{\varepsilon}{L}\Bigg\{ g^{ab}(x^{(l')})\partial_{(x^{(l')})^b} \\ & \sum_{l=1}^{L}\Bigg[\left(g^{ij}\left(x^{(l)}\right)\partial_{\left(x^{(l)}\right)^j}\log\left(p\left(x^{(l)}\right)\sqrt{\left|G\left(x^{(l)}\right)\right|}\right) + \partial_{\left(x^{(l)}\right)^j} g^{ij}\left(x^{(l)}\right)\right) \\ & \cdot \partial_{\left(x^{(l)}\right)^i} K\left(x^{(l)}, x^{(l')}\right) + g^{ij}\left(x^{(l)}\right)\partial_{\left(x^{(l)}\right)^i}\partial_{\left(x^{(l)}\right)^j} K\left(x^{(l)}, x^{(l')}\right)\Bigg]\Bigg\},\end{aligned} \tag{4-12}$$

其中 ε 是更新步长。这就是所提 RSVGD 方法在坐标空间中的更新算法。它可适用于具有全局坐标系的流形。特别地，对于欧氏空间上的贝叶斯推理问题，RSVGD 方法可使用信息几何技术[45-46] 来加快收敛，其中的黎曼度量矩阵 (g_{ij}) 被取为似然分布的费舍尔信息矩阵（Fisher information matrix），或者在其中进一步考虑先验分布的信息，将它取为费舍尔信息矩阵与先验分布密度函数对数的海森矩阵（Hessian matrix）之差[127]（注意 (g^{ij}) 是 (g_{ij}) 的逆矩阵）。

4.3.3　嵌入空间中的表达式

上面的推导和讨论给出了在坐标空间中实现 RSVGD 的算法。但这并不总是最方便的方式。出于与上一章相同的考虑，本章也希望可以在流形的嵌入空间中

实现 RSVGD 的算法。具体来说，例如考虑超球面 $\mathbb{S}^{n-1} := \{ x \in \mathbb{R}^n \mid \|x\| = 1 \}$ 或斯蒂菲尔流形（Stiefel manifold）$\mathbb{M}_{m,n} := \left\{ M \in \mathbb{R}^{m\times n} \,\middle|\, M^\top M = I_m \right\}$[43]。一方面，它们没有全局坐标系，因而若使用坐标空间中的算法，在实现过程中不仅需要不断地更换粒子所在的局部坐标系，而且还需要计算坐标系中 g^{ij}，$|G|$ 和 $\partial_i K$ 这样的量。对于斯蒂菲尔流形来说，找到它的一个坐标系甚至都是十分困难的。另一方面，它们是通过欧氏空间中的一个子集的方式来定义的，其几何结构也都由这个欧氏空间所导出，因而这个欧氏空间就是这些流形的一个自然的等距嵌入空间。因此，在流形的嵌入空间中实现算法可为解决这些流形上的推理任务带来极大便利。

正式地说，$\mathbb{R}^n$ 是 m 维流形 $\mathscr{M}$（$n \geqslant m$）的嵌入空间，是指存在一个从 $\mathscr{M}$ 到 $\mathbb{R}^n$ 的光滑单射 Ξ，称为嵌入映射。若 $\mathscr{M}$ 是一个黎曼流形，则可定义等距嵌入（isometric embedding），即嵌入映射 Ξ 是保持度量结构的。具体来说，等距嵌入要求在 $\mathscr{M}$ 的任一坐标系 $(\mathscr{I}, \Phi)$ 中，都有 $g_{ij} = \sum_{a=1}^{n} \dfrac{\partial y^a}{\partial x^i}\dfrac{\partial y^a}{\partial x^j}$，其中 $\dfrac{\partial y^a}{\partial x^i} := \dfrac{\partial (\Xi \circ \Phi^{-1})^a}{\partial x^i}$。在嵌入空间中，$\Xi(\mathscr{M}) \subset \mathbb{R}^n$ 具有豪斯多夫测度（Hausdorff measure），它为流形 $\mathscr{M}$ 引入一个测度。这个测度刚好就是 $\mathscr{M}$ 的黎曼体积形式所引出的测度。详细介绍可参见 2.1.2.4 节。

下面的结论给出了在一般流形的等距嵌入空间中泛函梯度 V^* 的表达式：

命题 4.1 (一般黎曼流形的等距嵌入空间中的泛函梯度)：设黎曼流形 $\mathscr{M}$ 通过嵌入映射 Ξ 等距嵌入在 $\mathbb{R}^n$ 中，并记 $\mathbb{R}^n$ 的一组单位正交基为 $\{y^a\}_{a=1}^n$。令流形 $\mathscr{M}$ 上所有函数和算符都通过映射 $\Xi^{-1} : \Xi(\mathscr{M}) \to \mathscr{M}$（因 Ξ 是单射故此映射可合法定义）从 $\Xi(\mathscr{M})$ 上取值。定义矩阵 $J_{n\times m} : J_{ai} = \dfrac{\partial y^a}{\partial x^i}$，即嵌入映射 Ξ 的雅可比矩阵。定义矩阵 $P(y)_{n\times(n-m)}$ 为切空间 $T_y\Xi(\mathscr{M})$（作为 $\mathbb{R}^n$ 中的 m 维线性子空间）的正交补空间的一组标准正交基按列排列起来的矩阵。则在此等距嵌入空间中，泛函梯度可表示为 $V^*(y') = (I_n - P'P'^\top)\nabla' f^*(y')$，其中

$$
\begin{aligned}
f^*(y') = \mathbb{E}_q\Big[&\Big(\nabla \log\left(p\sqrt{|G|}\right)\Big)^\top \Big(I_n - PP^\top\Big)(\nabla K) + \nabla^\top \nabla K \\
&- \operatorname{tr}\Big(P^\top(\nabla\nabla^\top K)P\Big) + \Big((J^\top\nabla)^\top (G^{-1}J^\top)\Big)(\nabla K)\Big],
\end{aligned}
\tag{4-13}
$$

其中 $\nabla = (\partial_{y^1}, \cdots, \partial_{y^n})^\top$，$\operatorname{tr}(\cdot)$ 表示矩阵的迹（trace），带有撇上标 “$'$” 的符号表示以 y' 为自变量，K 同时以 y 和 y' 为自变量，而其他符号以 y 为自变量。

证明 首先推导出一个有用的结论。令 $f : \Xi(\mathscr{M}) \to \mathbb{R}$ 是嵌入空间中的一个光滑函数。通过与 Ξ^{-1} 复合，它可被视作流形 $\mathscr{M}$ 上的一个光滑函数。由求导的链式法

则，可以推得：

$$\partial_i f = \partial_a f \frac{\partial y^a}{\partial x^i} = J^\top \nabla f,$$

其中矩阵 $J_{n\times m}$ 及 ∇ 如命题中所述。另外对于等距嵌入，可有性质 $G = J^\top J$。

由式 (4-10) 可知，$f^*(y') = \mathbb{E}_q\big[f_1 + f_2\big]$，其中 $f_1 = (\operatorname{grad} K)[\log p]$，而 $f_2 = \Delta K$。接下来进行如下变形：

$$\begin{aligned}
f_1 &= g^{ij}(\partial_i \log p)(\partial_j K) \\
&= g^{ij}\frac{\partial y^a}{\partial x^i}(\partial_a \log p)\frac{\partial y^b}{\partial x^j}(\partial_b K) \\
&= (\nabla \log p)^\top (JG^{-1}J^\top)\nabla K, \\
f_2 &= g^{ij}(\partial_i K)(\partial_j \log\sqrt{|G|}) + \partial_i(g^{ij}\partial_j K) \\
&= (\nabla \log\sqrt{|G|})^\top (JG^{-1}J^\top)\nabla K + \frac{\partial y^a}{\partial x^i}\partial_a(g^{ij}\frac{\partial y^b}{\partial x^j}\partial_b K) \\
&= (\nabla \log\sqrt{|G|})^\top (JG^{-1}J^\top)\nabla K + (J^\top\nabla)^\top(G^{-1}J^\top\nabla K) \\
&= (\nabla \log\sqrt{|G|})^\top (JG^{-1}J^\top)\nabla K + \Big((J^\top\nabla)^\top(G^{-1}J^\top)\Big)\nabla K \\
&\quad + \operatorname{tr}\Big((\nabla\nabla^\top K)(JG^{-1}J^\top)\Big).
\end{aligned}$$

为进一步简化此表达式，注意到 $JG^{-1}J^\top = J(J^\top J)^{-1}J^\top$ 是 $\mathbb{R}^n$ 中向 J 的列空间作正交投影的算符，而 $J(y)$ 的列空间正是 $\Xi(\mathscr{M})$ 在 y 处的切空间，它是 $\mathbb{R}^n$ 的一个 m 维线性子空间。而这个正交投影也可以通过使用矩阵 P 表示为 $I_n - PP^\top$。具体细节可以参看上一章中的 3.2.2 节。使用矩阵 P 表示的优势在于，它与 $\mathscr{M}$ 坐标系的选择无关，这样即使 $\mathscr{M}$ 没有全局坐标系，它也能够方便地在嵌入空间中进行表示。另外，矩阵 P 通常可以更简单直接地得到，并且使用 P 的表达式计算起来会方便很多（参见 3.2.2 节的说明）。通过引入矩阵 P，可以继续对 f^* 变形：

$$\begin{aligned}
f_1 + f_2 &= (\nabla \log p\sqrt{|G|})^\top (JG^{-1}J^\top)\nabla K + \Big((J^\top\nabla)^\top(G^{-1}J^\top)\Big)\nabla K \\
&\quad + \operatorname{tr}\Big((\nabla\nabla^\top K)(JG^{-1}J^\top)\Big) \\
&= (\nabla \log p\sqrt{|G|})^\top (I_n - PP^\top)\nabla K + \Big((J^\top\nabla)^\top(G^{-1}J^\top)\Big)\nabla K \\
&\quad + \operatorname{tr}\Big((\nabla\nabla^\top K) - (\nabla\nabla^\top K)PP^\top\Big) \\
&= (\nabla \log p\sqrt{|G|})^\top (I_n - PP^\top)\nabla K + \Big((J^\top\nabla)^\top(G^{-1}J^\top)\Big)\nabla K \\
&\quad + \nabla^\top\nabla K - \operatorname{tr}\Big(P^\top(\nabla\nabla^\top K)P\Big).
\end{aligned}$$

□

最后，$V^*(y') = \operatorname{grad}' f^*(y') = g'^{ij}\partial'_i f^*(y')\partial'_j = g'^{ij}\dfrac{\partial y^a}{\partial x^i}(y')\partial'_a f^*(y')\dfrac{\partial y^b}{\partial x^j}(y')\partial'_b =$

$J'G'^{-1}J'\nabla' f^*(y') = (I_n - P'P'^{\top})\nabla' f(y')$，这便完成了命题结论的推导。

注意 P 是与 $\mathscr{M}$ 的坐标系选取无关的，而 J 和 G 则有关，但最终结果 V^* 是与坐标系选取无关的。在嵌入空间中模拟 V^* 所引出的动力学系统与在坐标空间中的情况十分不同，因为这需要在一个受限的空间 $\Xi(\mathscr{M})$ 上移动粒子。给定一组在嵌入空间中表示的粒子 $\{y^{(l)}\}_{l=1}^L$，此动力学系统的一个一阶近似模拟方法是：

$$y^{(l')} \leftarrow \mathrm{Exp}_{y^{(l')}}\big(\varepsilon V^*(y^{(l')})\big), \tag{4-14}$$

其中 $V^*(y^{(l')})$ 可由粒子 $\{y^{(l)}\}_{l=1}^L$ 通过式 (4-13) 来估计，而 Exp_y 是 $\Xi(\mathscr{M})$ 在点 y 处的指数映射（exponential map），它将点 y 处的一个切向量 $v \in T_y\Xi(\mathscr{M})$ 映射到与 v 相切的长度为 $\|v\|$ 的测地线（geodesic）的终点。指数映射可看作是线性空间中的加法在黎曼流形上的推广。在 $\mathbb{R}^n$ 中，指数映射会将切向量 v 映射到 v 方向长度为 $\|v\|$ 的线段的终点，也就是 $y+v$。在超球面 $\mathbb{S}^{n-1}$ 上，指数映射会将切向量 v 映射到与 v 相切的长度为 $\|v\|$ 的大圆（the great circle; orthodrome）圆弧的终点：

$$\text{（超球面上指数映射）}\qquad \mathrm{Exp}_y(v) = y\cos(\|v\|) + (v/\|v\|)\sin(\|v\|). \tag{4-15}$$

总之，式 (4-14) 是所提 RSVGD 方法在一般黎曼流形的嵌入空间中进行计算的算法。

特例：超球面　这一部分考虑将上面的命题 4.1，即一般流形上的泛函梯度 V^* 在嵌入空间中的表达式，特例化为超球面的情况 $\mathscr{M} = \mathbb{S}^{n-1}$。首先给出结论如下。

命题 4.2 (超球面的等距嵌入空间中的泛函梯度)：对于等距嵌入在 $\mathbb{R}^n$ 中的超球面 $\mathbb{S}^{n-1}$，可有结论 $V^*(y') = (I_n - y'y'^{\top})\nabla' f^*(y')$，其中

$$f^*(y') = \mathbb{E}_q\Big[\big(\nabla \log p\big)^{\top}(\nabla K) + \nabla^{\top}\nabla K - y^{\top}\big(\nabla\nabla^{\top}K\big)y - (y^{\top}\nabla \log p + n - 1)y^{\top}\nabla K\Big]. \tag{4-16}$$

证明 对于等距嵌入在 $\mathbb{R}^n$ 中的超球面 $\mathbb{S}^{n-1}$，其嵌入映射 Ξ 为单位映射。不失一般性，考虑 $\mathbb{S}^{n-1}$ 的上半球这个坐标系 $(\mathscr{I} := \{\, y \in \mathbb{R}^n \mid \|y\| = 1, y^n > 0 \,\}, \Phi : y \mapsto (y^1, \cdots, y^{n-1})^{\top} \in \mathbb{R}^{n-1})$。在此坐标系中，可有结论 $\Phi(\mathscr{I}) = \big\{\, x \in \mathbb{R}^{n-1} \;\big|\; \|x\| < 1 \,\big\}$ 以及 $\Xi(x) = (x^1, \cdots, x^{n-1}, \sqrt{1 - x^{\top}x})^{\top}$，因此

$$J = \begin{pmatrix} I_{n-1} \\ -\dfrac{x^{\top}}{\sqrt{1 - x^{\top}x}} \end{pmatrix},$$

并且 $G = I_{n-1} + \dfrac{xx^\top}{1-x^\top x}$，$G^{-1} = I_{n-1} - xx^\top$，$|G| = \dfrac{1}{1-x^\top x}$。$\Xi(\mathbb{S}^{n-1})$ 在 $y \in \mathbb{R}^n$ 处的切空间是一个与 $\mathbb{R}^n$ 中的向量 y 垂直的超平面，因而其正交补空间就是向量 y 所指方向的直线。因而可得 $P(y) = y$。将这些结果代入式 (4-13) 中，经过整理便可得到在超球面 $\mathbb{S}^{n-1}$ 上 V^* 的嵌入空间表达式式 (4-16)。　□

注意此超球面上的嵌入空间表达式式 (4-16) 与超球面 $\mathbb{S}^{n-1}$ 的坐标系完全无关。在超球面 $\mathbb{S}^{n-1}$ 的嵌入空间中使用所提 RSVGD 方法的算法可由式 (4-16) 及式 (4-15) 给出。

最后，进一步考虑 RSVGD 方法在超球面的 T 次乘积流形（product manifold）$(\mathbb{S}^{n-1})^T$ 上的表示。这个超球面乘积流形是球面混合模型（spherical admixture model, SAM）[22] 中话题这个全局隐变量所在的空间，因而解决 SAM 模型的后验推理问题，需要在此乘积空间上使用 RSVGD 方法。

首先来考虑一般的乘积流形 $(\mathscr{M})^T$。对其上任一点 $x = (x_{(1)}, \cdots, x_{(T)}) \in (\mathscr{M})^T$，它附近的一个局部坐标系可表示为 $\left(\bigotimes_{k=1}^{T} \mathscr{I}_{(k)}, \bigotimes_{k=1}^{T} \{x_{(k)}^{i_{(k)}}\}_{i_{(k)}=1}^{n-1}\right)$，其中每个 $\left(\mathscr{I}_{(k)}, \{x_{(k)}^{i_{(k)}}\}_{i_{(k)}=1}^{n-1}\right)$ 都是对应流形 $\mathscr{M}_{(k)}$ 上 $x_{(k)}$ 附近的局部坐标系。在乘积流形 $(\mathscr{M})^T$ 的上述坐标系中，自然基底为 $\left\{ \partial_{(k),i_{(k)}} \,\middle|\, k = 1, \cdots, T,\ i_{(k)} = 1, \cdots, n-1 \right\}$，而由乘积流形的定义，其切空间为各个因子流形切空间的乘积空间，其黎曼结构是各个因子流形切空间中内积的乘积：$g_{(k,l),i_{(k)},j_{(l)}} := \delta_{kl} g_{i_{(k)},j_{(l)}}$，其中 $\delta_{kl} = 1$ 当且仅当 $k = l$ 否则为零。由此构造方式，乘积流形 $(\mathscr{M})^T$ 上一个光滑函数 $f \in C^\infty((\mathscr{M})^T)$ 的梯度可以表示为 $\operatorname{grad} f = \sum_{k=1}^{T} g_{(k)}^{i_{(k)} j_{(k)}} \partial_{(k),i_{(k)}} f \partial_{(k),j_{(k)}}$，而其上一个向量场 $V = \sum_{k=1}^{T} V_{(k)}^{i_{(k)}} \partial_{(k),i_{(k)}} \in \mathscr{T}((\mathscr{M})^T)$ 的散度可以表示为 $\operatorname{div}(V) = \sum_{k=1}^{T} \left(\partial_{(k),i_{(k)}} V_{(k)}^{i_{(k)}} + V_{(k)}^{i_{(k)}} \partial_{(k),i_{(k)}} \log \sqrt{|G_{(k)}|} \right)$。光滑函数 f 的贝尔特拉米-拉普拉斯算符 Δf 也可以得到表示。

现在考虑超球面的 T 次乘积流形 $(\mathbb{S}^{n-1})^T$。为将它在其嵌入空间 $(\mathbb{R}^n)^T$ 中表示，可将第 k 个因子流形 $\mathbb{S}^{n-1}$ 等距嵌入在 $\mathbb{R}^n$ 中，并用 $\mathbb{R}^n$ 中的向量 $y_{(k)}$ 表示。对于 $(\mathbb{S}^{n-1})^T$ 上的点 $y = (y_{(1)}, \cdots, y_{(T)})$，可考虑乘积化的核函数 $K(y, y') = \prod_{k=1}^{T} K_{(k)}(y_{(k)}, y'_{(k)})$，其中 $K_{(k)}$ 是第 k 个因子流形 $\mathbb{S}^{n-1}$ 上的核函数。基于这些准备，可以将命题 4.2 拓展到超球面的乘积流形中：

命题 4.3 (超球面的乘积流形中的泛函梯度（嵌入空间表达式）)：对于等距嵌入在 $(\mathbb{R}^n)^T$ 中的超球面 $\mathbb{S}^{n-1}$ 的 T 次乘积流形 $(\mathbb{S}^{n-1})^T$，其上泛函梯度 $V^* =$

$(V^*_{(1)}, \cdot, V^*_{(T)})$ 可表示为 $V^*(y')_{(k')} = (I_n - y'_{(k')}{y'_{(k')}}^\top)\nabla'_{(k')} f^*(y')$，其中

$$f^*(y') = \mathbb{E}_q\Big[K(y,y')\sum_{k=1}^{T}\Big[\big(\nabla_{(k)}\log p\big)^\top\big(\nabla_{(k)}\log K_{(k)}\big) + \nabla^\top_{(k)}\nabla_{(k)}\log K_{(k)} - y^\top_{(k)}\big(\nabla_{(k)}\nabla^\top_{(k)}K_{(k)}\big)y_{(k)} + \big\|\nabla_{(k)}\log K_{(k)}\big\|^2 - \big(y^\top_{(k)}\nabla_{(k)}\log K_{(k)}\big)^2 - \big(y^\top_{(k)}\nabla_{(k)}\log p + n - 1\big)y^\top_{(k)}\nabla_{(k)}\log K_{(k)}\Big]\Big].$$

使用此命题中的结论即可应用 RSVGD 方法解决 SAM 模型的后验推理问题。

4.4　实验

由于实验结果表明，在 RSVGD 中简单地直接使用随机选取的子数据集来估计梯度（即随机梯度）不能取得很好的结果，因而本部分实验中将只考虑使用整个数据集（即准确梯度）的情况下考察 RSVGD 和各方法的表现。RSVGD 方法与随机子数据集（随机梯度）的相容性将留作后续工作。实验代码和数据可从网站“http://ml.cs.tsinghua.edu.cn/~changliu/rsvgd/”下载。

4.4.1　贝叶斯逻辑回归模型实验

为考察 RSVGD 方法在欧氏空间上的推理任务中（参见式 (4-12)）的实际表现，本实验选取贝叶斯逻辑回归模型（Bayesian logistic regression, BLR）的后验推理任务进行测试。BLR 模型首先从先验 $\theta \sim \mathcal{N}(0, \alpha I_m)$ 中生成隐变量 θ，再对于每一个数据点 X_d，从伯努利似然分布中生成此数据点的类别 $Y_d \sim \mathrm{Bern}(\sigma(\theta^\top X_d))$，其中 $\sigma(x) = 1/(1+e^{-x})$ 是 sigmoid 函数。BLR 模型的推理任务即给出对后验分布 $p(\theta|\{Y_d\},\{X_d\})$ 的估计。

使用 RSVGD 方法的细节　为应用 RSVGD 方法表达在坐标空间中的算法，首先在此计算式 (4-12) 中所需的量。由上述 BLR 模型的定义，可以得知：

$$\text{对数先验：}\ \log p_0(\theta) = -\frac{\theta^\top\theta}{2\alpha} + \text{const},$$

$$\text{对数似然：}\ \log p(\{Y_d\}|\theta,\{X_d\}) = \sum_{d=1}^{|\mathscr{D}|}\Big(Y_d\theta^\top X_d - \log(1+e^{\theta^\top X_d})\Big) + \text{const},$$

$$\text{对数后验：}\ \log p(\theta|\{Y_d\},\{X_d\}) = -\frac{\theta^\top\theta}{2\alpha} + \sum_{d=1}^{|\mathscr{D}|}\Big(Y_d\theta^\top X_d - \log(1+e^{\theta^\top X_d})\Big) + \text{const}.$$

由此可以推导出目标分布（即后验分布）密度函数的梯度为：

$$\nabla \log p(\theta|\{Y_d\},\{X_d\}) = -\frac{1}{\alpha}\theta + \sum_{d=1}^{|\mathscr{D}|}\big(Y_d - \sigma(\theta^\top X_d)\big)\,X_d.$$

如 4.3.2 节最后所述，使用信息几何的思想可以把黎曼度量矩阵取为费舍尔信息矩阵 $\mathcal{M}$ 与先验分布密度函数对数的海森矩阵之差[127]，由此可以计算对于 BLR 模型，其黎曼度量矩阵为：

$$\begin{aligned} G(\theta) &= \mathcal{M}\big(p(\{Y_d\}|\theta,\{X_d\})\big) - \nabla\nabla^\top \log p_0(\theta) \\ &= \mathbb{E}_{p(\{Y_d\}|\theta,\{X_d\})}\big[\big(\nabla \log p(\{Y_d\}|\theta,\{X_d\})\big)\big(\nabla \log p(\{Y_d\}|\theta,\{X_d\})\big)^\top\big] \\ &\quad - \nabla\nabla^\top \log p_0(\theta) \\ &= \sum_{d=1}^{|\mathscr{D}|} c_d X_d X_d^\top + \frac{1}{\alpha} I_m, \end{aligned}$$

其中 $c_d = \sigma(\theta^\top X_d)(1-\sigma(\theta^\top X_d))$。对于其逆矩阵 G^{-1}，可以直接数值计算 G 的逆，其时间复杂度为 $\mathcal{O}(m^3)$。除此之外，注意到还有一个计算此逆矩阵的方法，其时间复杂度为 $\mathcal{O}(m^2|\mathscr{D}|)$。这个方法首先选取 $G_0^{-1} = \alpha I_m$，然后迭代使用谢尔曼-莫里森公式（Sherman-Morrison formula）[144]：

$$G_d^{-1} = G_{d-1}^{-1} - \frac{c_d (G_{d-1}^{-1}X_d)(G_{d-1}^{-1}X_d)^\top}{1 + c_d X_d^\top G_{d-1}^{-1} X_d},$$

直到得到 $G_{|\mathscr{D}|}^{-1}$ 作为最终对逆矩阵的计算结果。对于小数据集，或者随机选取的子数据集，这种方式会具有优势。不过在本实验中，直接数值计算逆矩阵要更快一些。

下面来计算式 (4-12) 中剩下的量。首先注意到 $\partial_i G := \partial_{\theta_i} G = \sum_{d=1}^{|\mathscr{D}|} f_d X_{di} X_d X_d^\top$，其中 $f_d = \dfrac{1-e^{\theta^\top X_d}}{1+e^{\theta^\top X_d}} c_d$。另外，由于 $\partial_i G_{jk} = \sum_{d=1}^{|\mathscr{D}|} f_d X_{di} X_{dj} X_{dk}$，因此 $\partial_i G_{jk}$ 中的下标 i,j,k 是完全可以重排的。特别地，可有结论 $\partial_i G_{jk} = \partial_j G_{ik}$。由此可有：

$$\partial_i \log|G(\theta)| = \mathrm{tr}(G^{-1}\partial_i G) = \sum_{d=1}^{|\mathscr{D}|} f_d (X_d^\top G^{-1} X_d) X_{di},$$

以及

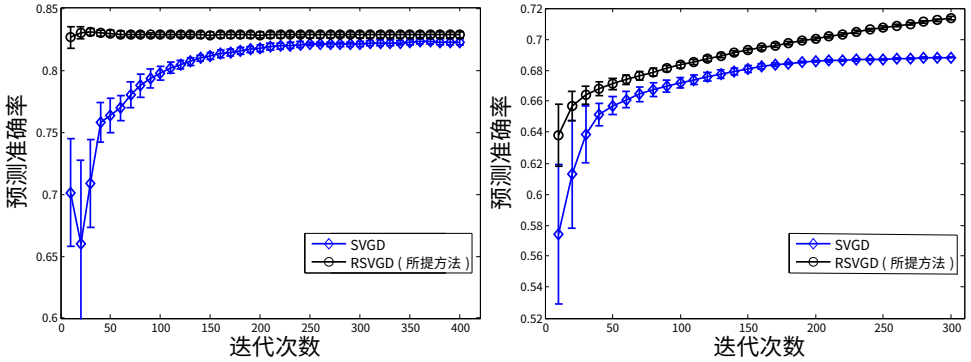
$$\sum_{j=1}^{m} \partial_j G_{ij}^{-1}(\theta) = -G_{(i,:)}^{-1} \sum_{j=1}^{m} (\partial_j G) G_{(:,j)}^{-1} = -\sum_{k=1}^{m} G_{(i,k)}^{-1} \sum_{j=1}^{m}\sum_{l=1}^{m} (\partial_j G)_{(k,l)} G_{(l,j)}^{-1}$$

$$= -\sum_{k=1}^{m} G^{-1}_{(i,k)} \sum_{j=1}^{m}\sum_{l=1}^{m} (\partial_k G)_{(j,l)} G^{-1}_{(l,j)} = -\sum_{k=1}^{m} G^{-1}_{(i,k)} \operatorname{tr}\left((\partial_k G) G^{-1}\right)$$
$$= -G^{-1}_{(i,:)} \nabla \log |G(\theta)|.$$

至此，式 (4-12) 中所有的量便都可以计算了。

核函数　由本章中引言（4.1 节）所介绍，在欧氏空间 $\mathbb{R}^m$ 上的推理任务中使用 RSVGD 相当于是在似然分布所构成的分布流形中模拟一个动力学系统，而这个欧氏空间就是这个分布流形的一个全局坐标系 $(\mathbb{R}^m, \Phi)$。为在分布流形上指定一个核函数，可以考虑 $\mathbb{R}^m$ 上的一个核函数，并将它与 Φ^{-1} 复合从而成为分布流形上的一个二元函数。由 Φ^{-1} 的双射性质，Steinwart 等人的著作[162] 引理 4.3 断言，这个二元函数就是分布流形上的一个核函数。因此只需选取 $\mathbb{R}^m$ 上的一个核函数即可。本实验中选择高斯核函数（Gaussian kernel）。由 Steinwart 等人的著作[162] 推论 4.44，高斯核函数的 RKHS 空间中零函数是唯一的常值函数，因而这个选择满足 RSVGD 方法对核函数的要求。更进一步，本实验在实现 SVGD 和 RSVGD 方法时将核函数取为具有不同带宽（bandwidth）的若干高斯核函数之和。由 Steinwart 等人的著作[162] 引理 4.5，这个和仍然是一个合法的核函数。实验中发现使用这个核函数比使用原本的 SVGD 方法中通过中位数方法确定带宽的单个高斯核函数的方法会取得更好的结果。

实验设定　本实验将所提的 RSVGD 方法与 SVGD 方法（均使用整个数据集估计梯度）进行对比，并以随训练迭代次数而变化的测试集上预测准确率来衡量。所使用的数据集包括 Splice19 数据集和 Covertype 数据集。Splice19 数据集（1,000 个训练样本，数据维度为 60）是 Mika 等人[168] 所编译整理的标准数据集中的一个，而 Covertype 数据集（581,012 个样本，数据维度为 54）是 SVGD 方法的原论文[81] 中使用的一个数据集。在 Splice19 数据集和 Covertype 数据集上，RSVGD 和 SVGD 两方法都分别重复了 20 次和 10 次，并取这些次的平均结果展示。参照 SVGD 原论文[81] 中的做法，本实验也在 Covertype 数据集上的每一次运行中都以 $80\% : 20\%$ 的比例随机地将数据集分为训练集和测试集。实验中固定了 BLR 模型的超参数 $\alpha = 0.01$，并对 RSVGD 和 SVGD 两方法都选择粒子数为 100。RSVGD 方法通过式 (4-12) 来更新粒子，并取步长 $\varepsilon = 0.3$。这个更新方式相当于优化中所说的梯度下降方法（gradient descent）。而 SVGD 方法使用其原论文中推荐的带有动量的 AdaGrad 方法来利用泛函梯度更新粒子，并取步长 $\varepsilon = 3 \times 10^{-3}$，遗忘率参数 0.9。对 SVGD 方法实验中也尝试了将带有动量的 AdaGrad 方法替换为梯度下降



(a) Splice19 数据集上的结果　　(b) Covertype 数据集上的结果

图 4.1　贝叶斯逻辑回归模型的后验推理任务中 RSVGD 和 SVGD 方法的推理结果（以在测试集上的预测准确率衡量）随训练迭代次数的变化曲线。

方法，但没有观察到更好的结果。

实验结果　从图 4.1 中可以看出，RSVGD 方法在两个数据集上都比 SVGD 方法具有更高的推理效率。这展示了 RSVGD 使用信息几何而带来的加快推理收敛速度的好处。尽管原本的 SVGD 方法中所使用的带有动量的 AdaGrad 方法可被解释为一个利用了目标函数几何性质的经验估计的优化方法[78]，但所提 RSVGD 方法提供了一个更加有原则性的方案，并且在实验中也取得了更好的效果。

4.4.2　球面混合模型实验

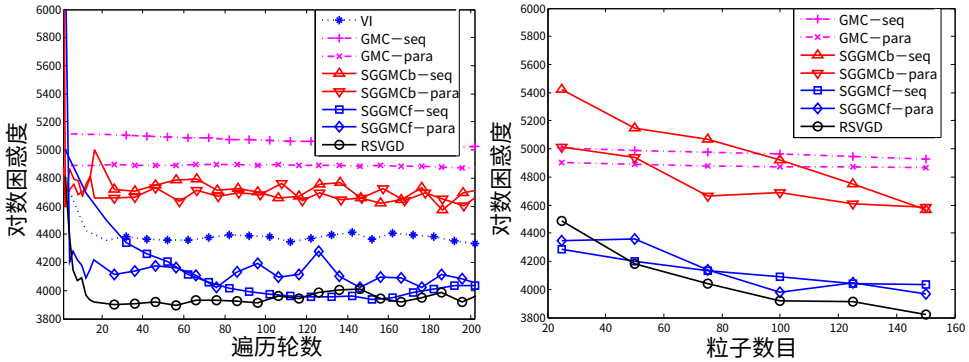
下面的实验将在球面混合模型（spherical admixture model, SAM）[22] 的后验推理任务上考察所提 RSVGD 方法在黎曼流形上的后验推理任务上的表现。SAM 模型是一个话题模型，它可以处理分布在超球面上的数据，例如由归一化词频-逆篇频（term frequency-inverse document frequency, tf-idf）特征所表示的文档数据。SAM 模型的后验推理任务是，在给定超球面 $\mathbb{S}^{n-1}$ 上的数据集 $X = \{X_d\}_d^{|\mathscr{D}|}$ 之后，估计话题这个全局隐变量 $\beta = \{\beta_\tau\}_{\tau=1}^T$ 的后验分布 $p(\beta|X)$，其中每一个数据点 X_d 和每一个话题 β_τ 都在超球面 $\mathbb{S}^{n-1}$ 上。后验分布密度函数的对数梯度 $\nabla \log p(\beta|X)$ 在嵌入空间中的表示可由上一章中的式 (3-12) 计算。详细信息可参考 3.3 节。注意此任务中的隐变量 $\beta = (\beta_1, \cdots, \beta_T)$ 是处在超球面的 T 次乘积流形 $(\mathbb{S}^{n-1})^T$ 上的，因此需要使用命题 4.3 中所给出的 RSVGD 的形式来实现算法。

核函数　与欧氏空间中所使用高斯核函数类似，在超球面 $\mathbb{S}^{n-1}$ 上可使用 vMF 核函数 $K(y, y') := \exp(\kappa y^\top y')$。这个 $\mathbb{S}^{n-1}$ 上的核函数是其等距嵌入空间 $\mathbb{R}^n$ 上的高斯核函数在它上面的限制，即对于超球面 $\mathbb{S}^{n-1}$ 上任意两点 y, y'，都有

$\exp(-\frac{\kappa}{2}\|y-y'\|^2) = \exp(-\kappa)\exp(\kappa y^\top y')$ 成立。因此，由 Steinwart 等人的著作[162] 引理 4.3 可知，vMF 核函数是超球面 $\mathbb{S}^{n-1}$ 上一个合法的核函数。另外，反正弦核函数 $K(y,y') := \arcsin(y^\top y')$ 也是超球面 $\mathbb{S}^{n-1}$ 上的一个核函数，这是因为反正弦函数的泰勒展开式（Taylor expansion）中各次项系数都是非负的，进而由 Steinwart 等人的著作[162] 引理 4.8 可知它是一个合法的核函数。不过不幸的是，这个核函数在实验中的表现并不如意，因此下面的实验使用的都是 vMF 核函数。有了超球面 $\mathbb{S}^{n-1}$ 上的核函数之后，便可根据 4.3.3 节中最后部分的做法，在其 T 次乘积流形上定义核函数：$K(y,y') := \prod_{k=1}^{T}\exp(\kappa y_{(k)}^\top y'_{(k)})$。最后，与上面的实验 4.4.1 节中的处理类似，本实验中也使用具有不同带宽（bandwidth）的若干 vMF 核函数之和作为最终使用的核函数。

实验设定　由于超球面这个流形的限制，诸多相关的推理方法都无法应用于此任务上，包括 SVGD 方法。SAM 模型的原工作[22] 中提出了一个基于平均场（mean-field）假设的变分推理方法（variational inference, VI）。这个方法是一种基于模型的变分推理方法（ModVI）。此外，在此任务上还有一些 MCMC 方法，例如测地线蒙特卡罗方法（geodesic Monte Carlo, GMC）[44]，以及在上一章所提出的具有可扩展性的随机梯度测地线蒙特卡罗方法（stochastic gradient geodesic Monte Carlo, SGGMC）和测地线随机梯度诺泽-胡佛恒温器方法（geodesic stochastic gradient Nosé-Hoover thermostats, gSGNHT）。本实验将选取 GMC 方法和 SGGMC 方法作为 MCMC 方法中的代表。这两个 MCMC 方法通常是通过模拟一条马尔可夫链的方式进行采样，即序列式（sequential, 简记为 “-seq”）采样，而为了与 ParVI 方法特别是所提 RSVGD 方法的模拟方式相匹配，本实验也为这两个 MCMC 方法实现了平行式（parallel，简记为 “-para”）采样，即同时平行地模拟多条马尔可夫链，每一次迭代所有链都各自更新一步自己的样本，并取所有链上的样本作为对目标分布的估计。另外对于 SGGMC 方法，由于它可以使用随机子数据集技术（或称随机梯度技术），因而可以同时考虑在一次迭代中使用全体数据集（full-batch，简记为 “f”）与使用随机子数据集（mini-batch，简记为 “b”）的 SGGMC 方法。SGGMCb 方法在实验中选取的随机子数据集大小为 50。在 ParVI 方法方面，本章中所提的 RSVGD 方法是第一个可以处理 SAM 模型后验推理任务的方法，因此 RSVGD 方法无法与其他 ParVI 方法对比，而只考虑与 VI 方法和 MCMC 方法进行对比。

为将 GMC，SGGMC 以及 RSVGD 等方法应用于 SAM 模型的推理任务中，可直接采用上一章 3.3 节中的框架，即算法 3。特别地，这些方法使用同样的梯度估计方法（除了标有 “-b” 亦即使用随机梯度的方法），因此它们之间的对比是公平



(a) 使用 100 个粒子（样本）的实验结果　(b) 遍历轮数为 200 时的实验结果

图 4.2　SAM 模型在 20News-different 数据集上的后验推理任务中各方法的推理效果。

的。这里采用上一章中 SAM 模型的实验部分（即 3.4.3 节）所使用的在 20News-different 数据集上的实验设定，例如模型超参数等。本实验也使用对数困惑度（log-perplexity）来衡量推理效果，它的值越小表明推理效果越好。不同的是，为了展示所提 RSVGD 方法的迭代有效性，实验结果将以各方法的推理效果随在数据集上的*遍历轮数*（epoch，即一个方法目前为止已经接触过的数据点的数目（重复接触过的数据点重复计数）占整个数据集大小的比例）的变化情况而表现。由于 SGGMCb 方法会在迭代中使用随机子数据集，因而考察随迭代次数的变化曲线对它是不公平的。而使用遍历轮数则可以公平比较。另外，为展示所提 RSVGD 方法的粒子高效性，本实验也会展示各方法推理效果随所用粒子（样本）数量的变化曲线。

实验结果　图 4.2(a) 展示了各推理方法推理效果随遍历轮数的变化曲线。从中可以发现，RSVGD 方法是随遍历轮数收敛最快的方法。这表明由于 RSVGD 方法是基于优化问题的且使用了确定性的更新方式，它在实际中具有迭代有效性的优势。VI 方法虽然收敛得也很快，但由于它为变分分布做了一个较强的平均场假设，因而最终无法取得一个很好的结果。与之对比，RSVGD 最终可取得更好的结果。这表明虽然同为变分推理方法，但由于 RSVGD 极大地放宽了对变分分布的假设，因而比 VI 方法具有更强的近似灵活性，从而取得更好的推理结果。这个结果与同样使用粒子（样本）的 SGGMCf 方法的结果是可比的。GMC 方法也在不断取得更好的结果，但图中所展示的遍历轮数范围（即 0 至 200 轮）仍然不足够。由于 SGGMC 方法所采样本之间仍然存在自相关性，因此在有限的粒子数目条件下，其结果没有 RSVGD 方法好。

图 4.2(b) 则展示了各方法使用不同粒子（样本）数时在遍历数据集 200 轮后的推理结果。可以发现在使用相同数目的粒子时，RSVGD 几乎总能取得最好的推

理效果，这展示了 RSVGD 的粒子高效性。在使用较少粒子（样本）时，SGGMCf 方法在 200 轮遍历后即可收敛，因而可以取得比 RSVGD 更好的结果。使用更多粒子（样本）时，样本之间的自相关性会更加明显地影响 SGGMCf 的收敛速度，使得这个遍历轮数后 SGGMCf 不足以收敛，从而被 RSVGD 超越。

4.5　本章小结与讨论

本章开发了黎曼-斯坦因变分梯度下降方法（Riemannian Stein variational gradient descent, RSVGD），作为 SVGD 方法[81] 在黎曼流形情况下的推广。本章将 SVGD 的思想做了抽象与推广，提出算法关键在于求出 KL 散度的方向导数和泛函梯度。本章推导出了黎曼流形上方向导数的表达式，而为了得到一个合法且有闭形式的泛函梯度，本章首先分析了一般黎曼流形对泛函梯度的一些要求，然后发现 SVGD 方法所用技术无法满足这些要求，最后提出解决方案，给出 RSVGD 方法。此方法在坐标空间中的表达式可由泛函梯度的坐标表示给出，而为了可以用于超球面这样没有全局坐标系的流形，本章也推导出了 RSVGD 在流形嵌入空间中的表达式。实验结果表明了在实际中，RSVGD 在欧氏空间上的推理任务中能够使用信息几何而带来的更快的收敛速度，以及在超球面上的推理任务中胜于已有可用方法的粒子高效性、迭代有效性以及近似灵活性。

此工作在未来可能的拓展方向包括深入挖掘所提的黎曼-核化斯坦因差异量（Riemannian kernelized Stein discrepancy, RKSD）在衡量两个分布之间差异上的性能。在黎曼流形的情况下，考虑流形几何结构的 RKSD 可以更好地反映此流形上两个分布之间的差别。探求可使用随机子数据集（随机梯度）的 RSVGD 方法是另一个有意义的研究方向。一方面，这种改进可使 RSVGD 变得更有可扩展性（scalability）从而可以快速处理大规模数据，而另一方面，如文中所说，简单直接地在 RSVGD 中使用随机梯度效果并不理想。将 RSVGD 方法应用于更多的任务中也是一个十分有前景的方向，例如欧氏空间上的推理任务如深度生成模型（deep generative models）和贝叶斯神经网络（Bayesian neural networks）的后验推理问题（其中黎曼度量矩阵可通过 Li 等人[157] 所提出的方法进行估计），以及黎曼流形上的推理任务如斯蒂菲尔流形上的贝叶斯矩阵补全方法（Bayesian matrix completion）[40]。

第 5 章　基于粒子的变分推理方法的分析与加速

基于粒子的变分推理方法（particle-based variational inference, ParVI）已在贝叶斯推理方法领域取得了令人瞩目的发展。它们所具有的近似灵活性和迭代有效性吸引了人们的注意。本章将从 ParVI 方法作为沃瑟斯坦空间（Wasserstein space）上的梯度流（gradient flow）这个视角进行深入探索，为 ParVI 方法进行一些理论分析，同时提出一些可行的新算法提高 ParVI 方法的实际表现。在理论方面，本章统一了现有 ParVI 方法使用有限多粒子进行模拟时所做的近似，发现这些近似实质上是一个必需的平滑操作，并可归结于平滑密度和平滑函数这两个等价的形式中。这个新的理解揭示了现有 ParVI 方法所做的假设以及它们之间的关系，同时也为开发新的 ParVI 方法提供了思路，例如本章中所开发的两个新的 ParVI 方法。在实用技术方面，本章为所有 ParVI 方法提出了一个加速框架以及一个带宽（bandwidth）选择方法。这些新技术都基于本章所建立的理论以及对沃瑟斯坦空间的几何性质的深入挖掘。实验结果表明，所提加速框架可使各 ParVI 方法获得更快的收敛速度，而所提带宽选择方法可使各 ParVI 方法所产生的粒子分布得更加准确而整齐。

5.1　研究动机

贝叶斯推理为建模数据的不确定性提供了强有力的工具。它的任务是在给定观测数据之后，估计一个贝叶斯模型隐变量的后验分布 p。这个分布的支撑空间（support space）$\mathscr{M}$ 就是此贝叶斯模型的隐空间。由于一般来说 p 的闭形式很难得到，因而便产生了各种近似方法。变分推理方法（variational inference, VI）通过在一个分布族中最小化与 p 之间的差距（通常由 KL 散度衡量）来给出一个可行的近似。这个分布族通常被选取为一个参数化分布族[30,62]，然后推理任务便可被转化为一个参数空间上的优化问题进而可以被有效求解。但是这个选择为用来近似 p 的分布（即变分分布，variational distribution）施加了一个较强的假设，限制了这类方法的近似灵活度，从而与目标分布 p 的接近程度会受到影响。马尔可夫链蒙特卡罗方法（Markov chain Monte Carlo, MCMC）[60,91,97,119] 则希望直接从 p 中进行采样。尽管它们具有渐进准确性（asymptotically accurate）的好处，但由于它们所采的样本之间具有自相关性，因而它们通常在实际中收敛得较慢。即使使用预热采样（burn-in），它们通常也需要一个较大的样本规模才能取得较好的结果[81]，而这则加重了处理后续任务的负担。

近来出现了一类新的贝叶斯推理方法，称为基于粒子的变分推理方法（particle-based variational inference, ParVI）。它们使用一组样本，或称粒子，来代表变分分布（类似 MCMC 方法），并通过最小化与 p 之间的 KL 散度来以一个确定性的方式更新这组粒子（类似 VI 方法）。与传统的 VI 方法对比，由于 ParVI 方法使用了非参数的粒子形式来表示变分分布，因而它们具有更强的近似灵活性。而与 MCMC 方法对比，由于它们直接考虑一组有限数目粒子的近似效果并考虑了粒子之间的相互作用，因而它们具有更强的粒子高效性，而基于优化 KL 散度的原理也使得它们可以收敛得更快。斯坦因变分梯度下降方法（Stein variational gradient descent, SVGD）[81] 是 ParVI 方法中的一个典型代表。它通过一个合适的向量场所引出的动力学系统来更新粒子，使得 KL 散度可以最快地被最小化。SVGD 所具有的 ParVI 方法的独特优势使得它在贝叶斯推理领域受到了高度关注。人们开发了它的很多变体[83-84,169] 并将它应用于诸多问题中[23-25,154,170-171]。

SVGD 方法随后被解释为使用有限多个粒子对一个由核函数所定义的分布流形 $\mathscr{P}_{\mathscr{H}}(\mathscr{M})$[85] 上 KL 散度的最速下降曲线族（steepest descending curves）进行模拟的方法。更加正式地，这个最速下降曲线族被称为*梯度流*（gradient flows）。受此启发，人们开始关注沃瑟斯坦空间（Wasserstein space）$\mathscr{P}_2(\mathscr{M})$[53-54] 上 KL 散度的梯度流（简称沃瑟斯坦梯度流），并考虑使用有限多个粒子来模拟此梯度流，进而开发了更多的 ParVI 方法。粒子优化方法（particle optimization method, PO）[55] 以及 w-SGLD 方法[56] 使用最小移动量框架（minimizing movement scheme, MMS）（参见 [61] 以及 Ambrosio 等人的著作[54] 定义 2.0.6）对沃瑟斯坦梯度流进行离散化处理，并通过一些近似使用有限多个粒子进行模拟。Blob 方法（最初称为 w-SGLD-B 方法）[56] 使用一个由向量场形式所表示的动力学系统来更新粒子。动力学系统的离散化处理是通过显式欧拉法来实现的，而它也做了一些近似从而可以使用有限多个粒子进行模拟。尽管这些 ParVI 方法已经有一些经验上的对比，但从理论上对它们所使用的有限多个粒子近似的分析却仍处空白。特别地，这些近似背后基于的假设以及这些具体近似方法之间的关系仍有待探索。另外还可以注意到，从优化领域的视角来看，现有的 ParVI 方法都是在模拟梯度流，却没有 ParVI 方法可以更充分地利用沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 的几何性质来对更加吸引人的 $\mathscr{P}_2(\mathscr{M})$ 上一阶加速方法进行模拟。此外，用于平滑操作的核函数（kernel）的带宽（bandwidth）对于 ParVI 方法的实际表现具有关键作用，而现在所使用的基于数值计算上的一个直觉的中位数方法（median method）[81] 不足以满足要求[169]。因此，ParVI 领域中还需要一个具有原则性的带宽选择方法。

本章中将深入挖掘沃瑟斯坦梯度流这个概念，以解决如上所述的 ParVI 领域

中的问题和需求。本章给出了一个 ParVI 方法的有限多个粒子近似的统一理论，并为所有 ParVI 方法提出了一个加速框架和一个具有原则性的带宽选择方法作为解决实际问题的可行方案。在理论方面，本章所给出的统一理论发现，各种 ParVI 方法本质上都是一个必需的平滑操作（smoothing treatment），并且可以由平滑密度（smoothing density）和平滑函数（smoothing functions）这两个等价的形式实现。本章揭示了现存 ParVI 方法所使用的有限多个粒子近似可归结在这两种平滑形式之中这一隐晦的事实，并由这两种形式的等价性将这些 ParVI 方法联系在了一起。而基于平滑操作的必要性，本章发现 ParVI 方法其实是做了对其变分分布的假设的。本章的理论也为开发新的 ParVI 方法给出了原则与启示，并据此设计开发了两个新的 ParVI 方法。在实际技术方面，所提加速框架是基于黎曼流形版本的[123-124] 涅斯捷洛夫加速方法（Nesterov's acceleration method）[77]。这个方法具有比梯度流模拟更快的收敛性的理论保证。为开发加速框架，本章对沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 的黎曼结构进行了具有原创性且十分深入的分析，得到了使用有限多个粒子对沃瑟斯坦空间上的指数映射（exponential map）与平行移动（parallel transport）进行计算的方法。这里需要强调的是，直接将涅斯捷洛夫加速方法应用于支撑空间 $\mathscr{M}$ 中的每个粒子上是缺乏理论合理性的，因为 KL 散度并不是在 $\mathscr{M}$ 上被最小化而是在 $\mathscr{P}_2(\mathscr{M})$ 上，因而对每一个粒子来说，它自己并不是在优化一个目标函数。为开发带宽选择方法，本章细致分析了平滑操作的目标并提出了带宽选择的一个原则，并根据此原则得到了一个可行的算法。实验结果表明，所提带宽选择方法可以比中位数方法[81] 得到质量更高的粒子，并且所提加速框架在实际中具有比原本的 ParVI 模拟方法更快的收敛速度。

相关工作　在理解 SVGD 方法方面，Liu[85] 首先给出了分布流形 $\mathscr{P}_{\mathscr{H}}(\mathscr{M})$ 上梯度流的解释。之后，Chen 等人[56] 试图将 SVGD 解释为沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的梯度流，但在他们的框架下这个解释无法成立。而本章中，在平滑操作这个观念下，SVGD 可被解释为使用有限多个粒子对沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上梯度流的近似。这样一来，所有的 ParVI 方法便有了一个统一的解释。分布流形 $\mathscr{P}_{\mathscr{H}}(\mathscr{M})$ 上梯度流的这个解释很难推广到其他 ParVI 方法上，并且这个空间的几何性质也无法为更进一步的分析和创新提供必要的工具。

在加速 ParVI 方法方面，PO 方法的算法形式与使用波利亚克动量方法（Polyak's momentum）[76] 的 SVGD 方法类似。但是，PO 方法并不是为加速而开发的，它所基于的原理并不能体现出加速的考虑，因此它作为 ParVI 方法的加速版本缺乏理论依据。另外，本章的实验也观察到 PO 方法没有所提的加速框架稳定，这与 Sutskever 等人[172] 的讨论类似。最近，Taghvaei 等人[57] 也考虑了为 ParVI 方

法加速。他们的工作使用的是 Wibosono 等人[173] 所开发的优化方法的变分形式开发了加速方法，而本章工作是基于黎曼流形上的涅斯捷洛夫加速方法以及沃瑟斯坦 $\mathscr{P}_2(\mathscr{M})$ 空间的几何性质。算法上，他们的方法使用了动量这组辅助变量来实现加速，而本章所提方法使用的是另一组辅助粒子。不过，这两个方法都需要处理使用有限多个粒子所需的近似，而本章是通过一个更加系统的方式来解决这个问题，并且具有多个变种，而他们的方法在本章的理论下只算作平滑密度这一类方法。此外，本章为沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 所建立的使用有限多个粒子计算指数映射（exponential map）和平行移动（parallel transport）的方法也为其他黎曼流形优化方法在沃瑟斯坦空间上的应用以进一步加强 ParVI 方法提供了直接的工具。

在将黎曼结构与 ParVI 方法结合方面，上一章所提的方法 RSVGD 考虑的是支撑空间 $\mathscr{M}$ 是一个黎曼流形的情况，包括使用信息几何技术以及解决黎曼流形上推理任务这两种情况。其思想是利用支撑空间 $\mathscr{M}$ 的几何结构，而本章中考虑的则是利用沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 其本身的几何结构。算法上，所提加速框架不需要知道模型的结构信息，并且计算起来更加快捷。Detommaso 等人[174] 的工作考虑了 KL 散度在分布流形 $\mathscr{P}_{\mathscr{H}}(\mathscr{M})$ 上的二阶微分信息以加速 SVGD，而所提加速框架考虑的则是具有可推广性的沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 因而可以应用于所有 ParVI 方法，并且所提方法仍然只需要一阶微分信息，因而计算更加快捷。

5.2　背景知识

本节将介绍沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 及其上的梯度流，并回顾已有的 ParVI 方法。本章中，为简化不必要的复杂性以及突出本章工作的贡献，将只考虑支撑空间为欧氏空间的情况，即 $\mathscr{M} = \mathbb{R}^m$。为准确描述沃瑟斯坦空间，这里将首先定义一些辅助概念。记 $\mathscr{C}_c^\infty(\mathscr{M})$ 为 $\mathscr{M}$ 上具有紧致支撑集的光滑向量值函数（值域为 $\mathbb{R}^m$）的集合，记 $C_c^\infty(\mathscr{M})$ 为这样的标量值函数（值域为 $\mathbb{R}$）的集合。记 $\mathscr{L}_q^2(\mathscr{M}) := \left\{ U \in \mathscr{T}(\mathbb{R}^m) \,\middle|\, \int \|U(x)\|_2^2 \,\mathrm{d}q < \infty \right\}$ 为 $\mathscr{M}$ 上关于分布 q 二次可积的向量值函数（将 $\mathbb{R}^m$ 上的向量场视为向量值函数 $\mathbb{R}^m \to \mathbb{R}^m$）所构成的希尔伯特空间（Hilbert space），其上内积定义为 $\langle U, V \rangle_{\mathscr{L}_q^2} := \int U(x) \cdot V(x) \,\mathrm{d}q$。记 $L_q^2(\mathscr{M})$ 为这样的标量值函数的集合。若这两个记号中没有指定分布 q，那么定义中所使用的测度将被取为勒贝格测度（Lebesgue measure）。定义一个分布 q 在一个 $\mathscr{M}$ 上的可测变换 $\mathcal{T} : \mathscr{M} \to \mathscr{M}$ 下的前推（push-forward）$\mathcal{T}_\# q$，为经过此变换 $\mathcal{T}$ 作用之后的满足分布 q 的随机变量所服从的分布。更加形式化地说，$\mathcal{T}_\# q$ 作为一个测度，为 $\mathscr{M}$ 上的任一可测集 $\mathscr{I}$ 给出的体积为 $\int_{\mathscr{I}} (\mathcal{T}_\# q)(x) \,\mathrm{d}x = \int_{\mathcal{T}^{-1}(\mathscr{I})} q(x) \,\mathrm{d}x$，其中 $\mathscr{I}$ 在可测变换 $\mathcal{T}$ 下的原像集 $\mathcal{T}^{-1}(\mathscr{I})$ 也是可测的。

5.2.1　作为黎曼流形的沃瑟斯坦空间

图 5.1 展示了本部分中涉及的概念。本文将定义在支撑空间 $\mathscr{M}$ 上所有概率分布所构成的集合记为 $\mathscr{P}(\mathscr{M})$。现在令 $\mathscr{M}$ 上定义了距离 $d(\cdot,\cdot)$。那么可以定义支撑空间 $\mathscr{M}$ 上的沃瑟斯坦空间（Wasserstein space）为：

$$\mathscr{P}_2(\mathscr{M}) := \left\{ q \in \mathscr{P}(\mathscr{M}) \;\middle|\; \exists x_0 \in \mathscr{M} \text{ s.t. } \mathbb{E}_q[d(x_0,x)^2] < +\infty \right\}.$$

在其上可定义著名的沃瑟斯坦距离（Wasserstein distance）（可参见 Villani 的著作[53] 定义 6.4）：

$$d_W(q,p) := \left(\inf_{\varrho \in \mathscr{R}(q,p)} \mathbb{E}_{\varrho(x,y)}[d(x,y)^2] \right)^{1/2},$$

其中 $\mathscr{R}(q,p) := \left\{ \varrho \in \mathscr{P}(\mathscr{M}\times\mathscr{M}) \;\middle|\; \int \varrho(x,y)\,\mathrm{d}x = p(y), \int \varrho(x,y)\,\mathrm{d}y = q(x) \right\}$ 是所有以分布 q 和 p 为边缘分布的乘积空间 $\mathscr{M}\times\mathscr{M}$ 上的联合分布的集合。有了此距离之后，沃瑟斯坦空间便成为了一个度量空间（metric space）。更进一步，人们发现了它的黎曼结构[52,175]。黎曼结构比度量空间的结构具有更多细节，因而可以定义更多的可以显式表示的量，例如梯度（可参见 Villani 的著作[53] 第 15 章）。为定义黎曼结构，首先要找到沃瑟斯坦空间上切向量（tangent vector）以及切空间（tangent spaces）的表示。为此，Villani 的著作[53] 定理 13.8 或 Ambrosio 等人的著作[54] 定理 8.3.1 及命题 8.4.5 给出结论，对于沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上任意一条光滑曲线 $(q_t)_t$，总存在支撑空间 $\mathscr{M}$ 上的一个几乎处处唯一的含时向量场 $V_t(x)$ 满足：（1）对于几乎处处 $t\in\mathbb{R}$，都有 $\partial_t q_t + \nabla\cdot(V_t q_t) = 0$ 成立，以及（2）$V_t \in \overline{\{\nabla\varphi \mid \varphi \in C_c^\infty\}}^{\mathscr{L}^2_{q_t}}$，其中的上划线表示取闭包（closure）操作。此向量场 $V_t(x)$ 的几乎处处唯一性使得它可以用来表示此光滑曲线 q_t 上的切向量，进而可以将它所属于的上述闭包空间当做沃瑟斯坦空间上的切空间，并记为 $T_{q_t}\mathscr{P}_2(\mathscr{M})$（可参见 Ambrosio 等人的著作[54] 定理 8.4.1）。注意到此切空间 $T_{q_t}\mathscr{P}_2(\mathscr{M})$ 是希尔伯特空间 $\mathscr{L}^2_{q_t}(\mathscr{M})$ 的闭子空间，因而可以在其上定义内积为 $\mathscr{L}^2_{q_t}(\mathscr{M})$ 的内积在 $T_{q_t}\mathscr{P}_2(\mathscr{M})$ 上的限制。这样沃瑟斯坦空间就有了一个黎曼结构。此黎曼结构的重要性质是，它所导出的距离（参见 2.1.2.5 节）正是沃瑟斯坦距离 d_W（即 Benamou-Brenier 公式[175]），因而这个黎曼结构与作为度量空间的沃瑟斯坦空间是相容的。需要说明的是，这个黎曼结构与信息几何[46] 中所考虑的费舍尔-姚度量（Fisher-Rao metric）这个黎曼结构是不同的。后者只能在参数分布族上定义。而沃瑟斯坦空间上的黎曼结构在参数分布族上的限制最近也被用于变分推理方法中[51]。

最后，$\mathscr{P}_2(\mathscr{M})$ 上切向量的这个 $\mathscr{M}$ 上向量场形式的表示也提供了一个简洁的

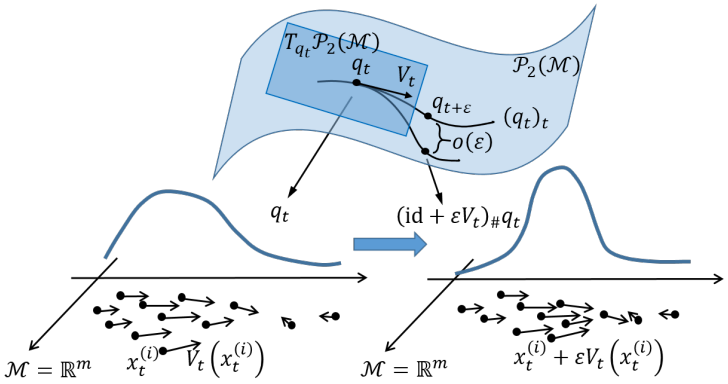


图 5.1　沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 及相关概念的展示。

模拟 $\mathscr{P}_2(\mathscr{M})$ 上光滑曲线 $(q_t)_t$（后称分布曲线）的方式。令 $V_t(x)$ 是分布曲线在 q_t 处的切向量，则前推分布 $(\mathrm{id}+\varepsilon V_t)_\# q_t$（其中 V_t 被视作 $\mathscr{M}$ 上的变换）是分布 $q_{t+\varepsilon}$ 的一个一阶近似（在沃瑟斯坦距离的意义下）（可参见 Ambrosio 等人的著作[54] 命题 8.4.6）。由前推分布的定义，这意味着若 $\{x^{(i)}\}_i$ 是分布 q_t 的一组样本，则 $\{x^{(i)}+\varepsilon V_t(x^{(i)})\}_i$ 可视为分布 $q_{t+\varepsilon}$ 的一组样本（对于绝对值很小的 ε 来说）。

5.2.2　沃瑟斯坦空间上的梯度流

直觉上来说，沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的一个函数 $\mathcal{F}$ 的梯度流就是 $\mathcal{F}$ 的 emph 最速下降曲线族$\{(q_t)_t\}$。它在度量空间上有多种数学上的具体定义方式（不一定是等价的定义）（可参考 Ambrosio 等人的著作[54] 定义 11.1.1），例如可定义为最小移动量框架（minimizing movement scheme, MMS）（可参见 Ambrosio 等人的著作[54] 定义 2.0.6）所定义的曲线族在步长趋于零时的极限曲线族。对于步长 ε，MMS 框架所定义的曲线族为：

$$q_{t+\varepsilon} = \operatorname*{argmin}_{q\in\mathscr{P}_2(\mathscr{M})} \mathcal{F}(q) + \frac{1}{2\varepsilon} d_W^2(q, q_t). \tag{5-1}$$

而若这个度量空间是一个黎曼流形，则这些不同的具体数学定义方式就都变成相互等价的了，并且这个合而为一的梯度流概念可以通过此流形的黎曼结构来定义（可参见 Villani 的著作[53] 命题 23.1 和注释 23.4，Ambrosio 等人的著作[54] 定理 11.1.6，以及 Erbar 等人的工作[176] 引理 2.7）。具体地，黎曼流形上梯度流的一条曲线 $(q_t)_t$ 满足其每一点处的切向量都是函数 $\mathcal{F}$ 在该点上的梯度（gradient），而梯度定义为：

$$\operatorname{grad}\mathcal{F}(q_t) := \max\cdot\operatorname*{argmax}_{V\in T_{q_t}\mathscr{P}_2(\mathscr{M}),\|V\|_{T_{q_t}\mathscr{P}_2(\mathscr{M})}=1} \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathcal{F}\Big((\mathrm{id}+\varepsilon V)_\# q_t\Big)\Big|_{\varepsilon=0}. \tag{5-2}$$

贝叶斯推理任务希望近似隐变量的后验分布 p。这可以通过在沃瑟斯坦空间上最小化 KL 散度（或称相对熵，relative entropy）来实现。对于分布 $p \in \mathscr{P}_2(\mathscr{M})$ 以及相对于 p 绝对连续（absolutely continuous）的分布 $q \in \mathscr{P}_2(\mathscr{M})$（因而可以定义 q 关于 p 的拉东-尼科迪姆导数（Radon-Nikodym derivative）q/p），关于 p 的 KL 散度定义为 $\mathrm{KL}_p(q) := \int_{\mathscr{M}} \log(q/p)\,\mathrm{d}q$。而若分布 q 关于分布 p 不绝对连续，则无法定义 q 关于 p 的拉东-尼科迪姆导数。通常将此种情况下的 KL 散度的值取作正无穷以保持其连续性。由于大多数贝叶斯模型的后验分布（取作 p）都是（关于 $\mathbb{R}^m$ 中的勒贝格测度（Lebesgue measure））绝对连续的，因此下面便只考虑 p 是绝对连续的情况。这种情况下，q 关于 p 的绝对连续性就成为了 q（关于 $\mathbb{R}^m$ 中的勒贝格测度）的绝对连续性。最小化 KL 散度的过程可通过模拟其梯度流来实现，因为梯度流具有最速下降曲线的特征。具体地，KL 散度的梯度流 $(q_t)_t$ 可通过此曲线在每一点处的切向量等于 KL 散度的梯度（可参见 Villani 的著作[53] 定理 23.18 或 Ambrosio 等人的著作[54] 例子 11.1.2）来刻画：

$$V_t^{\mathrm{GF}} := -\operatorname{grad} \mathrm{KL}_p(q_t) = \nabla \log p - \nabla \log q_t, \tag{5-3}$$

其中 q_t 应是绝对连续的。当 KL_p 在 $\mathscr{P}_2(\mathscr{M})$ 上测地 λ-凸（geodesically λ-convex），例如当密度函数 p 在 $\mathscr{M}$ 上 λ-对数凹（λ-log-concave）时（可参见 Villani 的著作[53] 定理 17.15 或 Ambrosio 等人的著作[54] 定理 9.4.11），其梯度流 $(q_t)_t$ 如人们所期待的那样，具有指数收敛性：$d_W(q_t, p) \leqslant e^{-\lambda t} d_W(q_0, p)$（可参见 Villani 的著作[53] 定理 23.25 及定理 24.7，以及 Ambrosio 等人的著作[54] 定理 11.1.4）。

注释 5.1：　郎之万动力学系统（Langevin dynamics）$\mathrm{d}x = \nabla \log p(x)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t(x)$（其中 B_t 是标准布朗运动（standard Brownian motion））也被发现是沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上关于分布 p 的 KL 散度的梯度流（可参见 Jordan 等人的著作[61]，他们从 MMS 框架的角度发现了这个结论）。从梯度流的角度来看，$\mathscr{M}$ 上的郎之万动力学系统与 $\mathscr{M}$ 上的确定性动力学系统 $\mathrm{d}x = V_t^{\mathrm{GF}}(x)\,\mathrm{d}t$ 会产生同样的分布曲线 $(q_t)_t$，而后者正是沃瑟斯坦梯度流所对应的动力学系统。这个结论也可以由福克-普朗克方程（Fokker-Planck equation）[177] 得出。

5.2.3　基于粒子的变分推理方法（ParVI）

本节介绍一些现有的基于粒子的变分推理方法（particle-based variational inference, ParVI）。斯坦因变分梯度下降方法（Stein variational gradient descent, SVGD）[81] 考虑使用 $\mathscr{M}$ 上的一个合适的向量场 V 来更新粒子：$x_{k+1}^{(i)} = x_k^{(i)} + \varepsilon V(x_k^{(i)})$，而这个向量场 V 是通过最大化 KL 散度的减小速度 $-\dfrac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathrm{KL}_p\big((\mathrm{id} + \varepsilon V)_{\#}q\big)\big|_{\varepsilon=0}$ 来选择

的，其中 q 是粒子 $\{x^{(i)}\}_i$ 所代表的分布。当向量场 V 从 $\mathscr{M}$ 上一个核函数（kernel）K 的向量值再生核希尔伯特空间（reproducing kernel Hilbert space, RKHS）$\mathscr{H}^m$（可参见 Steinwart 等人的著作[162] 定义 4.18）中选择时，其最优解可闭式表示：

$$V^{\text{SVGD}}(\cdot) := \mathbb{E}_{q(x)}[K(x,\cdot)\nabla \log p(x) + \nabla_x K(x,\cdot)]. \tag{5-4}$$

注意到求解 V^{SVGD} 的优化问题与刻画梯度的优化问题（式 (5-2)）形式十分类似，因而 Liu[85] 将 SVGD 解释为一个切空间取为 RKHS 空间 $\mathscr{H}^m$ 的分布流形 $\mathscr{P}_{\mathscr{H}}$ 上 KL 散度的梯度流。式 (5-4) 中的期望可通过在有限多个粒子上做平均而估计，这等价于将分布 $q(x)$ 取作经验分布（empirical distribution）$\hat{q}(x) := \dfrac{1}{N}\sum_{i=1}^{N}\delta_{x^{(i)}}(x)$，其中 $\delta_{x^{(i)}}(x)$ 是集中在 $x^{(i)}$ 这一点上的狄拉克测度（Dirac measure）。

其他 ParVI 方法则是通过对沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上 KL 散度的梯度流进行模拟而开发的。Blob 方法[56] 使用有限多个粒子来直接使用式 (5-3) 模拟梯度流。它将式 (5-3) 中需要进行估计的项 $U^{\text{GF}} := -\nabla \log q$ 重新表示为一个变分问题的最优解：

$$U^{\text{GF}} = \nabla\big(-\frac{\delta}{\delta q}\mathbb{E}_q[\log q]\big), \tag{5-5}$$

并通过与一个核函数 K 做卷积（convolution）来对其中的密度函数 q 进行部分地平滑：

$$\begin{aligned} U^{\text{Blob}} &= \nabla\big(-\frac{\delta}{\delta q}\mathbb{E}_q[\log(q * K)]\big) \\ &= -\nabla \log \tilde{q} - \nabla\big((q/\tilde{q}) * K\big), \end{aligned}$$

其中 “$*$” 表示卷积，而平滑密度函数定义为 $\tilde{q} := q * K$。在这个形式中，q 便可取作 $\hat{q}$。

粒子优化方法（particle optimization method, PO）[55] 使用 MMS 框架（式 (5-1)）来模拟梯度流，其中的沃瑟斯坦距离 d_W 通过求解一个对偶最优传输问题（dual optimal transport problem）来估计。最后推导出的粒子更新方法为：

$$x_k^{(i)} = x_{k-1}^{(i)} + \varepsilon(V^{\text{SVGD}}(x_{k-1}^{(i)}) + \mathcal{N}(0,\sigma^2 I)) + \lambda(x_{k-1}^{(i)} - x_{k-2}^{(i)}),$$

其中的 ε，σ，λ 都是方法参数。这个形式与结合了波利亚克动量方法（Polyak's momentum method）[76] 的 SVGD 方法类似。w-SGLD 方法[56] 通过求解熵正则化的（entropy-regularized）最优传输问题的原问题来估计 MMS 框架中的沃瑟斯坦距离 d_W。其算法与 PO 方法类似。

5.3　作为模拟沃瑟斯坦梯度流的 ParVI 方法

本节展示本章工作中关于 ParVI 方法主要的理论部分。此理论发现各种 ParVI 方法都通过平滑操作（smoothing）来模拟沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的梯度流，而这个平滑操作可分类为平滑密度（smoothing density）以及平滑函数（smoothing functions）这两类形式。本节分析了现有的 ParVI 方法，将它们以平滑密度和平滑函数两种形式进行归类，分析平滑操作的必要性以及各平滑方法之间的等价性，最后受所提理论启发开发了两个新的 ParVI 方法。

5.3.1　SVGD 方法模拟沃瑟斯坦梯度流的解释

目前，SVGD 方法只被理解为是对分布流形 $\mathscr{P}_{\mathscr{H}}(\mathscr{M})$ 上 KL 散度的梯度流的模拟[85]。本节首先将 SVGD 解释为使用有限多个粒子对沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上 KL 散度的梯度流的模拟，这样所有现有的 ParVI 方法都可用这同一种方式来理解，从而方便进行深入而统一的分析。注意到 V^{GF} 是希尔伯特空间 $\mathscr{L}_q^2$ 中的一个元素，因此可以通过如下方式来确定这个向量场：

$$V^{\mathrm{GF}} = \max \cdot \operatorname*{argmax}_{V\in\mathscr{L}_q^2, \|V\|_{\mathscr{L}_q^2}=1} \langle V^{\mathrm{GF}}, V\rangle_{\mathscr{L}_q^2}. \tag{5-6}$$

接着可以发现，如果将上述优化问题的优化域 $\mathscr{L}_q^2$ 替换为核函数 K 的向量值 RKHS 空间 $\mathscr{H}^m$，那么这个优化问题可以得到闭式解，并且这个解刚好就是 SVGD 中所使用的向量场 V^{SVGD}。这个发现将沃瑟斯坦梯度流和 SVGD 联系了起来。

定理 5.1 (V^{SVGD} 是对 V^{GF} 的近似)：SVGD 方法中所使用的向量场 V^{SVGD}（参见式 (5-4)）是对沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上 KL 散度的梯度流 V^{GF} 的近似，其近似方式为将确定 V^{GF} 的优化问题 (5-6) 中的优化域 $\mathscr{L}_q^2$ 替换为向量值 RKHS 空间 $\mathscr{H}^m$，即：

$$V^{\mathrm{SVGD}} = \max \cdot \operatorname*{argmax}_{V\in\mathscr{H}^m, \|V\|_{\mathscr{H}^m}=1} \langle V^{\mathrm{GF}}, V\rangle_{\mathscr{L}_q^2}. \tag{5-7}$$

证明 对于任意 $V \in \mathscr{H}^m$，优化问题 (5-7) 中的目标函数可以表示为：

$$\begin{aligned}
&\langle V^{\mathrm{GF}}, V\rangle_{\mathscr{L}_q^2} \\
&= \mathbb{E}_q[(\nabla \log p - \nabla \log q)\cdot V] = \mathbb{E}_q[\nabla \log p \cdot V] - \int_{\mathscr{M}} \nabla q \cdot V \,\mathrm{d}x \\
&\overset{(*)}{=} \mathbb{E}_q[\nabla \log p \cdot V] + \int_{\mathscr{M}} q \nabla \cdot V \,\mathrm{d}x \\
&= \mathbb{E}_{q(x)}\left[\sum_{a=1}^{m}\Big(\partial_a \log p(x) V^a(x) + \partial_a V^a(x)\Big)\right]
\end{aligned}$$

$$
\begin{aligned}
&\overset{(\#)}{=} \mathbb{E}_{q(x)}\left[\sum_{a=1}^{m}\left(\partial_a \log p(x)\langle K(x,\cdot), V^a(\cdot)\rangle_{\mathscr{H}} + \langle \partial_a K(x,\cdot), V^a(\cdot)\rangle_{\mathscr{H}}\right)\right] \\
&= \mathbb{E}_{q(x)}\left[\langle K(x,\cdot)\nabla \log p(x), V(\cdot)\rangle_{\mathscr{H}^m} + \langle \nabla K(x,\cdot), V(\cdot)\rangle_{\mathscr{H}^m}\right] \\
&= \mathbb{E}_{q(x)}\left[\langle K(x,\cdot)\nabla \log p(x) + \nabla K(x,\cdot), V(\cdot)\rangle_{\mathscr{H}^m}\right] \\
&= \left\langle \mathbb{E}_{q(x)}\left[K(x,\cdot)\nabla \log p(x) + \nabla K(x,\cdot)\right], V(\cdot)\right\rangle_{\mathscr{H}^m} \\
&= \langle V^{\text{SVGD}}, V\rangle_{\mathscr{H}^m},
\end{aligned}
$$

其中标有 $(*)$ 的等号成立是由一个分布的弱导数（weak derivative）的定义所保证的（可参见 Nicolaescu 的著作[132] 定义 10.2.1），而标有 (#) 的等号成立是由核函数 K 的 RKHS 空间 $\mathscr{H}$ 中任一函数 f 的再生性质（reproducing property）所保证的，即 $\langle K(x,\cdot), f(\cdot)\rangle_{\mathscr{H}} = f(x)$（可参见 Steinwart 等人的著作[162] 第 4 章），以及 $\langle \partial_{x^a} K(x,\cdot), f(\cdot)\rangle_{\mathscr{H}} = \partial_{x^a} f(x)$（可参见 Zhou 的著作[165]）。 □

本章下文中将会提及，向量值 RKHS 空间 $\mathscr{H}^m$ 大致可以看作 $\mathscr{L}_q^2$ 空间的一个子空间，因而 SVGD 向量场 V^{SVGD} 可以看作是梯度流 V^{GF} 在向量值 RKHS 空间 $\mathscr{H}^m$ 上的投影。

值得注意的是，之前将 SVGD 解释为分布流形 $\mathscr{P}_{\mathscr{H}}(\mathscr{M})$ 上梯度流的观点[56,85] 并不能完全让人满意。特别地，分布流形 $\mathscr{P}_{\mathscr{H}}(\mathscr{M})$ 并不是一个合法定义了的黎曼流形。此流形的定义方式是切空间为向量值 RKHS 空间 $\mathscr{H}^m$ 的分布流形，但是在微分流形的领域中，人们只知道给定了一个流形之后，其切空间可由此流形的拓扑结构（topology）所确定[129]，却不知道是否唯一存在一个流形其切空间刚好是一个事先指定好了的线性空间。特别地，对于一个合法的流形，其上任一光滑曲线在曲线上任一点处的切向量应在此点处的切空间中有唯一的表示。沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 满足这个要求（可参见 Villani 的著作[53] 定理 13.8，或 Ambrosio 等人的著作[54] 定理 8.3.1 及命题 8.4.5），但对于分布流形 $\mathscr{P}_{\mathscr{H}}(\mathscr{M})$ 来说，目前还没有类似的保证。分布流形 $\mathscr{P}_{\mathscr{H}}(\mathscr{M})$ 也缺少一些应用中所需要的性质，例如其上的距离目前还没有显式的表达式[56]。此外，SVGD 也可被解释为一个弗拉索夫过程（Vlasov process）[56,85]，即有相互作用的粒子族的演化过程。这个观点解释了 SVGD 为何可以保持目标分布 p 不变，但除此之外还不能提供关于 SVGD 以及 ParVI 方法的更多信息。

5.3.2　ParVI 方法的平滑操作

由上一节的分析，所有现有的 ParVI 方法都可以解释为是对沃瑟斯坦空间梯度流的模拟。本节进一步发现，在此模拟过程中，ParVI 方法都使用了一个近似，

即一个必需的平滑操作（smoothing），并且可以分为平滑密度和平滑函数这两种形式。

平滑密度　注意到 Blob 方法通过在梯度流的变分形式 (5-5) 中将分布 q 替换为核函数平滑过的分布 $\tilde{q} := \hat{q} * K$ 这种方式实现使用有限多个粒子来对沃瑟斯坦梯度流进行近似。而 w-SGLD 方法[56] 在使用最优传输问题的原形式（primal form）求解沃瑟斯坦距离 d_W 时，为此优化问题的目标函数中引入了一个熵正则项（entropy regularizer）。这些做法一方面是为了方便优化问题的求解，而另一方面，这个熵正则项可起到避免所解得的分布过于集中在一些点上的效果，也就是为所解得的分布的密度函数做了一个平滑性的要求。因而这两个方法都可以看作是通过平滑密度的方式，使用有限多个粒子对沃瑟斯坦梯度流进行近似。

平滑函数　定理 5.1 表明 SVGD 方法是通过在梯度流的优化形式中将函数族 $\mathscr{L}_q^2$ 替换为函数族 $\mathscr{H}^m$ 来近似沃瑟斯坦梯度流的。进一步，本节发现 $\mathscr{H}^m$ 中的函数大致上就是经过核函数平滑过后的 $\mathscr{L}_q^2$ 中的函数。这个不是很显然但富有启发性的结论由下面的定理准确叙述：

定理 5.2 ($\mathscr{H}^m$ 是对 $\mathscr{L}_q^2$ 的平滑)：　对于欧氏空间 $\mathscr{M} = \mathbb{R}^m$，考虑其上的高斯核函数 K 以及其上绝对连续的分布 q。则核函数 K 的向量值 RKHS 空间 $\mathscr{H}^m$ 等距同构于闭包空间 $\mathscr{G} := \overline{\{\phi * K \mid \phi \in \mathscr{C}_c^\infty\}}^{\mathscr{L}_q^2}$。

证明 当分布 q（关于 $\mathscr{M} = \mathbb{R}^m$ 的勒贝格测度）绝对连续时，函数空间 $\mathscr{L}_q^2$ 与函数空间 $\mathscr{L}^2$ 具有相同的拓扑性质，因而下面所引用的针对函数空间 $\mathscr{L}^2$ 的结论可以直接应用于函数空间 $\mathscr{L}_q^2$ 中。注意到映射 $\phi \mapsto \phi * K, \mathscr{L}^2 \to \mathscr{L}^2$ 是连续的，所以闭包空间 $\mathscr{G}$ 可变形为：$\mathscr{G} := \overline{\{\phi * K \mid \phi \in \mathscr{C}_c^\infty\}}^{\mathscr{L}_q^2} = \left\{ \phi * K \,\middle|\, \phi \in \overline{\mathscr{C}_c^\infty}^{\mathscr{L}^2} \right\} = \left\{ \phi * K \,\middle|\, \phi \in \mathscr{L}^2 \right\} = \left\{ \phi * K \,\middle|\, \phi \in L^2 \right\}^m$，其中倒数第二个等式成立是由 Kováčik 等人的著作[178] 定理 2.11 所保证的。而另一方面，由 Steinwart 等人的著作[162] 命题 4.46 及定理 4.47，映射 $\phi \mapsto \phi * K$ 是经核函数 K 平滑后的 L^2 空间 $\left\{ \phi * K \,\middle|\, \phi \in L^2 \right\}$ 与核函数 K 的 RKHS 空间 $\mathscr{H}$ 之间的等距同构。因此可以得知，$\mathscr{G}$ 与 $\mathscr{H}^m$ 是等距同构的。　□

注意到 $\overline{\mathscr{C}_c^\infty}^{\mathscr{L}_q^2} = \mathscr{L}_q^2$（可参见 Kováčik 等人的著作[178] 定理 2.11），因此可以说函数空间 $\mathscr{C}_c^\infty$ 大致上就是函数空间 $\mathscr{L}_q^2$，因此闭包空间 $\mathscr{G}$ 大致上就是经过核函数平滑过后的 $\mathscr{L}_q^2$ 中的函数所构成的集合，而此定理将这个集合与 RKHS 空间 $\mathscr{H}^m$ 等同了起来。由此可以发现，SVGD 方法在近似梯度流时将 $\mathscr{L}_q^2$ 替换为 $\mathscr{H}^m$ 的做法本质上就是一种平滑函数的操作。

如 5.2.3 节所提到的，PO 方法[55] 在求解对偶最优传输问题时，将作为优化域的函数族限制为二次函数族。由于二次函数具有有限的尖锐性（sharpness）（因为二次函数的二阶导数不会变化），因此这个处理本质上也是一种平滑函数的操作。

等价性　现在来分析平滑密度和平滑函数这两种操作的等价性，从而使上面的分析具有更丰富的含义以及更强的重要性。注意到 SVGD 方法优化形式（式 (5-7)）中的目标函数为 $\langle V^{\mathrm{GF}}, V\rangle_{\mathscr{L}_q^2} = \mathbb{E}_q[V^{\mathrm{GF}} \cdot V]$。本节将这个形式作一个推广，即将 $V^{\mathrm{GF}} \cdot V$ 替换为一个线性映射 $\mathcal{L} : \mathscr{L}_q^2(\mathscr{M}) \to L_q^2(\mathscr{M})$ 在 V 上的作用，从而将目标函数表达为 $\mathbb{E}_q[\mathcal{L}(V)]$。由积分的可交换性以及 $\mathcal{L}$ 的线性性可知，

$$\mathbb{E}_{\tilde{q}}[\mathcal{L}(V)] = \mathbb{E}_{q*K}[\mathcal{L}(V)] = \mathbb{E}_q[\mathcal{L}(V) * K] = \mathbb{E}_q[\mathcal{L}(V * K)].$$

这表明平滑分布操作 $q * K$ 与平滑函数操作 $V * K$ 是等价的。这个等价性将两类 ParVI 方法联系了起来，使得对一类 ParVI 方法的分析和开发的技术（例如本章将要开发的加速框架和带宽选择方法）也可用于另一类 ParVI 方法。

平滑操作的必需性以及 ParVI 方法所做的假设　这里需要强调的是，ParVI 方法的平滑操作是由 KL 散度的梯度流的合法定义性（well-definedness）所要求的。由于 ParVI 方法都在模拟此梯度流，因此这个平滑操作是必需的。当 q 不是绝对连续，例如将 q 取作经验分布 $\hat{q}$ 时，KL 散度将会取值无穷，因而其梯度流在此时也无法合理定义。本节将进一步指出，这个必需的平滑操作即为 ParVI 方法对其变分分布 q 所做的假设，即假设变分分布 q 是绝对连续的。只不过这个假设既可以直接通过平滑密度来实现，也可以通过将模拟方法表示为一个优化问题然后平滑函数优化域并取 $q = \hat{q}$ 来实现。

这个发现可能对使用平滑密度操作的 ParVI 方法来说比较直接，但对于使用平滑函数操作的 ParVI 方法来说就不那么显然了。特别地，针对 SVGD 这个使用平滑函数操作的 ParVI 方法，本节直接对它进行专门的分析，并指出既不平滑密度（即取 $q = \hat{q}$）也不平滑函数（即取优化问题 (5-7) 中的优化域为 $\mathscr{L}_p^2$ ①）会导致不合理的结果产生。

定理 5.3 (SVGD 方法中平滑操作的必要性)：　对于 $q = \hat{q}$ 且 $V \in \mathscr{L}_p^2$ 的情况，优化问题 (5-7) 没有最优解。事实上，这种情况下目标函数的上确界是无穷大，这表明一个最大化目标函数的 V 的序列会变得越发病态。

① 不取为 $\mathscr{L}_q^2$ 的原因：首先，斯坦因恒等式（Stein's identity）[85] 这个 SVGD 最初所基于的理论的成立条件要求 $V \in \mathscr{L}_p^2$。另外，在取 $q = \hat{q}$ 的情况下，由于 $\hat{q}$ 不是绝对连续的，因而 $\mathscr{L}_{\hat{q}}^2$ 不是一个合适的函数希尔伯特空间。

证明 本证明中可能会重新定义一些正文中已使用的符号。由定理 5.1 的证明中的推导，优化问题 (5-7) 的目标函数 $\langle V^{\mathrm{GF}}, V\rangle_{\mathscr{L}_q^2}$ 可写为 $\mathbb{E}_q[\nabla \log p \cdot V + \nabla \cdot V]$。取 $q = \hat{q}$ 及 $V \in \mathscr{L}_p^2$，优化问题 (5-7) 可以写为

$$\sup_{V\in\mathscr{L}_p^2, \|V\|_{\mathscr{L}_p^2}=1} \sum_{i=1}^{N} \Big(\nabla \log p(x^{(i)}) \cdot V(x^{(i)}) + \nabla \cdot V(x^{(i)})\Big), \tag{5-8}$$

下面要寻找一个满足式 (5-8) 中约束的 $\{V_n\}$ 的序列使得目标函数可趋于无穷。

这里假设存在 $r_0 > 0$ 使得对于任意满足 $\left\|x - x^{(i)}\right\|_\infty < r_0$（其中 i 是集合 $\{1, 2, \cdots, N\}$ 中的某一个值）的 x 都有 $p(x) > 0$。这个假设是合理的，否则会有一个粒子 $x^{(i)}$ 不可能是 $p(x)$ 的样本。

将 $V(x)$ 记为 $(V^1(x), \cdots, V^m(x))^\top$，将 $\nabla f(x)$ 记为 $(\partial_1 f(x), \cdots, \partial_m f(x))^\top$，则目标函数可写为

$$\begin{aligned}
\mathcal{J}_V &= \sum_{i=1}^{N} \Big(\nabla \log p(x^{(i)}) \cdot V(x^{(i)}) + \nabla \cdot V(x^{(i)})\Big) \\
&= \sum_{i=1}^{N} \Big(\sum_{a=1}^{m} \partial_a[\log p(x^{(i)})] V^a(x^{(i)}) + \sum_{a=1}^{m} \partial_a[V^a(x^{(i)})]\Big) \\
&= \sum_{a=1}^{m}\sum_{i=1}^{N} \Big(\partial_a[\log p(x^{(i)})] V^a(x^{(i)}) + \partial_a[V^a(x^{(i)})]\Big).
\end{aligned} \tag{5-9}$$

对于任意 $V \in \mathscr{L}_p^2$ 满足 $\|V\| = 1$, 可以相应地定义一个函数 $\phi = (\phi^1, \cdots, \phi^m)^\top \in \mathscr{L}^2$ 满足 $\phi(x) = p(x)^{\frac{1}{2}} V(x)$，亦即 $\phi^a(x) = p(x)^{\frac{1}{2}} V^a(x)$（其中 $a \in \{1, 2, \cdots, m\}$ 表示分量下标），且

$$\|\phi\|_2^2 = \int_{\mathbb{R}^m} \phi^2 \,\mathrm{d}x = \int_{\mathbb{R}^m} \sum_{a=1}^{m} (\phi^a(x))^2 \,\mathrm{d}x = \int_{\mathbb{R}^m} \sum_{a=1}^{m} (V^a(x))^2 p(x) \,\mathrm{d}x = \|V\|^2 = 1.$$

将式 (5-9) 重新以 ϕ 表示：

$$\begin{aligned}
\mathcal{J}_\phi &= \sum_{a=1}^{m}\sum_{i=1}^{N} \Big(\partial_a[\log p(x^{(i)})] V^a(x^{(i)}) + \partial_a[V^a(x^{(i)})]\Big) \\
&= \sum_{a=1}^{m}\sum_{i=1}^{N} \Big(\partial_a[\log p(x^{(i)})] \phi^a(x^{(i)}) p(x^{(i)})^{-\frac{1}{2}} + \partial_a[\phi^a(x^{(i)}) p(x^{(i)})^{-\frac{1}{2}}]\Big) \\
&= \sum_{a=1}^{m}\sum_{i=1}^{N} \Big(\frac{1}{2} p(x^{(i)})^{-\frac{3}{2}} \partial_a[p(x^{(i)})] \phi^a(x^{(i)}) + p(x^{(i)})^{-\frac{1}{2}} \partial_a[\phi^a(x^{(i)})]\Big)
\end{aligned} \tag{5-10}$$

$$= \sum_{a=1}^{m}\sum_{i=1}^{N}\Big(A_a^{(i)}\phi^a(x^{(i)}) + B^{(i)}\partial_a[\phi^a(x^{(i)})]\Big),$$

其中 $A_a^{(i)} := \frac{1}{2}p(x^{(i)})^{-\frac{3}{2}}\partial_a[p(x^{(i)})]$，而 $B^{(i)} := p(x^{(i)})^{-\frac{1}{2}} > 0$。因此，可转而构造一个 ϕ 的序列 $\{\phi^n\}$ 来说明优化问题

$$\inf_{\phi\in\mathscr{L}^2,\|\phi\|=1}\sum_{a=1}^{m}\sum_{i=1}^{N}\Big(A_a^{(i)}\phi^a(x^{(i)}) + B^{(i)}\partial_a[\phi^a(x^{(i)})]\Big) \tag{5-11}$$

没有最优解，然后再通过 $\{\phi^n\}$ 为原优化问题 (5-8) 构造 V 的序列 $\{V_n\}$ 来说明问题。

定义如下函数序列

$$\chi_n(x) = \begin{cases} I_n^{-1/2}(1-x^2)^{n/2}, & \text{for } x\in[-1,1], \\ 0, & \text{otherwise,} \end{cases}$$

其中 $I_n := \int_{-1}^{1}(1-x^2)^n\,\mathrm{d}x = \sqrt{\pi}\frac{\Gamma(n+1)}{\Gamma(n+3/2)}$，而 $\Gamma(\cdot)$ 是伽马函数 (Gamma function)。所以有 $\int_{\mathbb{R}}\chi_n(x)^2\,\mathrm{d}x = 1$。注意到当 $x = -1/\sqrt{n}$ 时，

$$\begin{aligned}\chi_n'(x) &= -nI_n^{-\frac{1}{2}}x(1-x^2)^{\frac{n-2}{2}} \\ &= \pi^{-\frac{1}{4}}\sqrt{\frac{\Gamma(n+\frac{3}{2})}{\Gamma(n+1)}}\sqrt{n}(1-\frac{1}{n})^{\frac{n-2}{2}} \qquad (x=-\frac{1}{\sqrt{n}}) \\ &> \pi^{-\frac{1}{4}}\sqrt{n}(1-\frac{1}{n})^{\frac{n-2}{2}}, \qquad (\Gamma(n+\frac{3}{2}) > \Gamma(n+1))\end{aligned}$$

因此，

$$\lim_{n\to\infty}\chi_n'(-\frac{1}{\sqrt{n}}) > \lim_{n\to\infty}\pi^{-\frac{1}{4}}\sqrt{n}(1-\frac{1}{n})^{\frac{n-2}{2}} = \pi^{-\frac{1}{4}}e^{-\frac{1}{2}}\lim_{n\to\infty}\sqrt{n} = +\infty.$$

记 $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \cdots, x_m^{(i)})^\top \in \mathbb{R}^m, i = 1, \cdots, N$ 并定义

$$r_1 := \frac{1}{3}\min_{i\neq j}\big\|x^{(i)} - x^{(j)}\big\|_\infty = \frac{1}{3}\min_{a\in\{1,\cdots,m\},i\neq j}\big|x_a^{(i)} - x_a^{(j)}\big|.$$

由此，可以将 χ_n 拓展到 $\mathbb{R}^m$ 上，并将其记为 ξ_n：

$$\xi_n(x_1, x_2, \cdots, x_m) = r^{-m/2}\prod_{a=1}^{m}\chi_n(\frac{x_i}{r}),$$

其中 $r = \min\{r_0, r_1\}$，且 ξ_n 的支撑集为 $\text{supp}(\xi_n) = [-r, r]^m$。易知，$\int_{\mathbb{R}^m} \xi_n(x)^2 \,\mathrm{d}x = 1$，以及

$$\lim_{n\to\infty} \partial_a \xi_n(-\epsilon_n) = +\infty,\ a = 1, 2, \cdots, m,$$

其中 $\epsilon_n = \dfrac{r}{\sqrt{n}}(1, 1, \cdots, 1)^\top$。

这里选择 $\phi^a(x) = \dfrac{1}{Nm}\sum_{i=1}^{N} \psi_a^{(i)}$，其中 $\psi_a^{(i)}$ 定义为

$$\psi_a^{(i)}(x) := \begin{cases} \xi_n(x - x^{(i)} - \epsilon_n), & \text{if } A_a^{(i)} >= 0, \\ -\xi_n(x - x^{(i)} + \epsilon_n), & \text{if } A_a^{(i)} < 0. \end{cases}$$

由于 $\int_{\mathbb{R}^m} \psi_a^{(i)}(x)\psi_a^{(j)}(x) \,\mathrm{d}x = 0, \forall i \neq j$，可以知道这样的 ϕ^n 满足优化问题 (5-11) 中的限制条件。注意到对于任意 i, j，都有 $A_a^{(i)}\psi_a^{(j)}(x^{(i)}) \geqslant 0$，且

$$\partial_a \psi_a^{(j)}(x^{(i)}) = \begin{cases} +\infty, & \text{when } n \to \infty, \text{if } i = j, \\ 0, & \text{if } i \neq j, \end{cases}$$

因此当 $n \to \infty$ 时，式 (5-10) 中的 $\mathcal{J}_{\phi^n} \to +\infty$。因此这个序列 $\{\phi^n\}$ 即为定理中所需要的序列。

最后，本证明使用 $\{\phi^n\}$ 来构造 V 的序列。由于 $\text{supp}(\phi^n) \subset \text{supp}(p)$，因此可以定义 $V_n = \phi^n / \sqrt{p(x)}$。这个序列中每个 V_n 都满足优化问题 (5-8) 中所要求的限制条件，且目标函数 $\mathcal{J}_{V_n}$ 会随着 $n \to \infty$ 而趋于无穷。因此这就是此定理中希望找的序列。注意到作为 $\mathscr{L}_p^2$ 上的函数，$\mathcal{J}_V$ 不可能取得无穷大这个值，因此无穷大这个上确界是无法取得的，也就是说，优化问题 (5-8) 没有最优解。　□

SVGD 方法声称它没有对变分分布 q 的密度形式做任何假设，而只需要使用其样本。而本节发现，SVGD 方法其实只是把对密度 q 的假设转移到了 V 上。将 V 取在 RKHS 空间 $\mathscr{H}^m$ 中并非只为了得到 V 的最优闭式解，更重要的是，这个做法可以保证得到一个正确的向量场。可以看到，在做平滑假设这件事上没有“免费的午餐”。ParVI 方法必须要么平滑密度要么平滑函数。

5.3.3　基于平滑操作分析的新 ParVI 方法

上一节中所提出的 ParVI 方法通过平滑操作来使用有限多个粒子对梯度流做模拟这个理论上的理解可作为开发新 ParVI 方法的一个原则。下面将分别基于平滑密度和平滑函数来开发两个新的 ParVI 方法。

平滑密度的梯度流方法（GFSD）　首先考虑使用平滑分布 $\tilde{q} := \hat{q} * K$ 来近似 q 并使用梯度流的向量场形式（式 (5-3)）：

$$U^{\text{GFSD}} := -\nabla \log \tilde{q}.$$

对应的 ParVI 方法被称为平滑密度的梯度流方法（gradient flow with smoothed density，GFSD）。

平滑函数的梯度流方法（GFSF）　除了平滑函数以及 SVGD 所使用的平滑函数方法，本章还发现了另一个方式来使用优化形式来表示梯度流中不便模拟的部分 $U^{\text{GF}} := -\nabla \log q$。通过平滑函数操作，这种表示形式可建立一种新的 ParVI 方法。对于向量场 U^{GF}，可将它变形为 $qU^{\text{GF}} + \nabla q = 0$，然后将它看成一个以弱导数（weak derivative）（可参见 Nicolaescu 的著作[132] 定义 10.2.1）这个更加广泛的概念所表示的形式。这意味着 $\mathbb{E}_q[\phi \cdot U - \nabla \cdot \phi] = 0, \forall \phi \in \mathscr{C}_c^\infty$（此式也可通过分部积分得出），或者等价地，

$$U^{\text{GF}} = \operatorname*{argmin}_{U \in \mathscr{L}^2} \max_{\substack{\phi \in \mathscr{C}_c^\infty, \\ \|\phi\|_{\mathscr{L}_q^2} = 1}} \big(\mathbb{E}_q[\phi \cdot U - \nabla \cdot \phi]\big)^2.$$

取 $q = \hat{q}$，并使用核函数 K 对函数 $\phi \in \mathscr{C}_c^\infty$ 做平滑处理，由定理 5.2，这便等价于将 ϕ 取在向量值 RKHS 空间中 $\mathscr{H}^m$：

$$U^{\text{GFSF}} := \operatorname*{argmin}_{U \in \mathscr{L}^2} \max_{\substack{\phi \in \mathscr{H}^m, \\ \|\phi\|_{\mathscr{H}^m} = 1}} \big(\mathbb{E}_{\hat{q}}[\phi \cdot U - \nabla \cdot \phi]\big)^2. \tag{5-12}$$

参见下面的推导过程可以发现，上面的优化问题有闭式最优解。以矩阵形式表示，这个最优解为 $\hat{U}^{\text{GFSF}} = \hat{K}' \hat{K}^{-1}$，其中 $\hat{U}^{\text{GFSF}}_{:,i} := U^{\text{GFSF}}(x^{(i)})$，$\hat{K}_{ij} := K(x^{(i)}, x^{(j)})$，以及 $\hat{K}'_{:,i} := \sum_j \nabla_{x^{(j)}} K(x^{(j)}, x^{(i)})$。

推导　由于这里使用了平滑函数技术，因此可以在确定向量场 U^{GFSF} 的优化问题 (5-12) 中，取 q 为经验分布 $\hat{q} = \dfrac{1}{N} \sum_{i=1}^N \delta_{x^{(i)}}(x)$。这样一来，此优化问题即变为：

$$\min_{U \in \mathscr{L}^2} \max_{\substack{\phi \in \mathscr{H}^m, \\ \|\phi\|_{\mathscr{H}^m} = 1}} \Big(\sum_{i=1}^N \big(\phi(x^{(i)}) \cdot U^{(i)} - \nabla \cdot \phi(x^{(i)})\big) \Big)^2,$$

其中 $U^{(i)} := U(x^{(i)})$。对于向量值 RKHS 空间 $\mathscr{H}^m$ 中的函数 ϕ，可以使用 RKHS 空间中的再生性质（reproducing property）[165]：$\langle \phi^a(\cdot), K(x, \cdot) \rangle_{\mathscr{H}} = \phi^a(x)$ 以及

$\langle \phi^a(\cdot), \partial_{x_b} K(x,\cdot)\rangle_{\mathscr{H}} = \partial_{x_b}\phi^a(x)$，并由此将目标函数变形为：

$$\begin{aligned}
&\Big(\sum_a \sum_j \big(U_a^{(j)}\phi^a(x^{(j)}) - \partial_{x_a^{(j)}}\phi^a(x^{(j)})\big)\Big)^2 \\
=\ &\left(\sum_a \left\langle \sum_j \big(U_a^{(j)} K(x^{(j)},\cdot) - \partial_{x_a^{(j)}} K(x^{(j)},\cdot)\big), \phi^a(\cdot)\right\rangle_{\mathscr{H}}\right)^2 \\
=\ &\left\langle \sum_j \big(U^{(j)} K(x^{(j)},\cdot) - \nabla_{x^{(j)}} K(x^{(j)},\cdot)\big), \phi(\cdot)\right\rangle_{\mathscr{H}^m}^2 .
\end{aligned}$$

记 $\zeta := \sum_j \big(U^{(j)}K(x^{(j)},\cdot) - \nabla_{x^{(j)}}K(x^{(j)},\cdot)\big) \in \mathscr{H}^m$。那么关于 ϕ 取完最大值之后的目标函数，亦即关于 U 的目标函数为：$\|\zeta\|^2_{\mathscr{H}^m} = \sum_{i,j}\big(U^{(i)}U^{(j)}K(x^{(i)},x^{(j)}) - 2U^{(i)}\nabla_{x^{(j)}}K(x^{(j)},x^{(i)}) + \nabla_{x^{(i)}}\nabla_{x^{(j)}}K(x^{(i)},x^{(j)})\big) = \operatorname{tr}(\hat{U}\hat{K}\hat{U}^\top) - 2\operatorname{tr}(\hat{K}'\hat{U}^\top) + \text{const}$，其中 $\hat{U}_{:,i} := U^{(i)}$，而 $\hat{K}$ 和 $\hat{K}'$ 定义如上。为关于 $\hat{U}$ 最小化这个二次目标函数，对这个函数求关于 $\hat{U}$ 的导数并令之等于零，最终求得最优解为 $\hat{U}^{\text{GFSF}} = \hat{K}'\hat{K}^{-1}$。

(推导毕)

此外，这个用于确定 U^{GF} 的优化问题也可以通过标量值函数 $\varphi \in C_c^\infty$ 来表示，进而对 φ 做平滑处理即取 $\varphi \in \mathscr{H}$ 从而可以得到一个 ParVI 方法。参见下面的推导，这种做法将会得到相同的结果。

推导　记 φ 为 $\mathscr{M}$ 上的任一标量值函数。对于等式 $U(x) = -\nabla \log q(x)$，即 $U(x)q(x) + \nabla q(x) = 0$，可以使用标量值函数 φ 在弱导数的意义下表示它：

$$\mathbb{E}_{q(x)}[\varphi(x)U(x) - \nabla\varphi(x)] = 0, \forall \varphi \in C_c^\infty.$$

由于此处考虑平滑函数操作，因而可以将上式中的 q 取作经验分布 $\hat{q} = \frac{1}{N}\sum_{j=1}^N \delta_{x^{(j)}}(x)$，其中 $\{x^{(j)}\}_j$ 是分布 $q(x)$ 的一组样本。此时上面确定 $U(x)$ 的式子便可写作：

$$\sum_j \big(\varphi(x^{(j)})U^{(j)} - \nabla\varphi(x^{(j)})\big) = 0, \forall \varphi \in C_c^\infty,$$

其中 $U^{(j)} = U(x^{(j)})$。

下面考虑使用核函数 K 对函数 φ 进行平滑操作。由定理 5.2，这等价于将 φ 取自核函数 K 的 RKHS 空间 $\mathscr{H}$ 中。同时，将等于零的等式写作最小化平方值的

形式可以得到：

$$\min_{\hat{U}\in\mathbb{R}^{m\times N}} \max_{\substack{\varphi\in\mathscr{H},\\ \|\varphi\|_{\mathscr{H}}=1}} \mathcal{J}(\hat{U},\varphi) := \sum_{j,a}\left(\varphi(x^{(j)})\hat{U}_{aj} - \partial_{x_a^{(j)}}\varphi(x^{(j)})\right)^2,$$

其中 $\hat{U}_{aj} := u_a(x^{(j)})$。通过使用 RKHS 空间中的再生性质（reproducing property），可以将上述优化问题中的目标函数 $\mathcal{J}(\hat{U},\varphi)$ 重写为：

$$\begin{aligned}\mathcal{J}(\hat{U},\varphi) &= \sum_a \langle\varphi(\cdot),\zeta_a(\cdot)\rangle_{\mathscr{H}}^2,\\ \zeta_a(\cdot) &:= \sum_j\left(\hat{U}_{aj}K(x^{(j)},\cdot) - \partial_{x_a^{(j)}}K(x^{(j)},\cdot)\right).\end{aligned}$$

通过使用线性代数的操作，上述最优化问题可写作：

$$\max_{\varphi\in\mathscr{H},\|\varphi\|_{\mathscr{H}}=1} \mathcal{J}(\hat{U},\varphi) = \lambda_1(M(\hat{U})),$$

其中 $\lambda_1(M(\hat{U}))$ 表示矩阵 M 的最大本征值（eigenvalue），而矩阵 $M(\hat{U})$ 定义为 $M(\hat{U})_{ab} := \langle\zeta_a(\cdot),\zeta_b(\cdot)\rangle_{\mathscr{H}}$，即

$$M(\hat{U}) = \hat{U}\hat{K}\hat{U}^\top - (\hat{K}'\hat{U}^\top + \hat{U}\hat{K}'^\top) + \hat{K}'',$$

其中 $\hat{K}''_{ab} := \sum_{i,j}\partial_{x_a^{(i)}}\partial_{x_b^{(j)}}K(x^{(i)},x^{(j)})$。假设这组样本 $\{x^{(j)}\}_j$ 没有重复元素。那么矩阵 $\hat{K}$ 则是正定的（positive definite），因此可以对它进行乔里斯基分解（Cholesky decomposition）：$\hat{K} = CC^\top$，其中 C 是一个与 K 等阶的非奇异矩阵。此时矩阵 $M(\hat{U})$ 可表示为 $M(\hat{U}) = (\hat{U}C - \hat{K}'C^{-1\top})(\hat{U}C - \hat{K}'C^{-1\top})^\top + (\hat{K}'' - \hat{K}'\hat{K}^{-1}\hat{K}'^\top)$。而只要 $\hat{U}C \neq \hat{K}'C^{-1\top}$，其中的第一项就会是半正定的（positive semidefinite）且其最大特征值为正，进而有 $\lambda_1(M(U)) > \lambda_1(\hat{K}'' - \hat{K}'\hat{K}^{-1}\hat{K}'^\top)$ 这个与 $\hat{U}$ 无关的量。因此，为最小化 $\lambda_1(M(\hat{U}))$，一定需要 $\hat{U}C = \hat{K}'C^{-1\top}$，此时才有 $\lambda_1(M(U)) = \lambda_1(\hat{K}'' - \hat{K}'\hat{K}^{-1}\hat{K}'^\top)$。这个条件可写为 $\hat{U} = \hat{K}'(CC^\top)^{-1} = \hat{K}'\hat{K}^{-1}$，从而得到了与上面平滑向量值函数 $\phi\in\mathscr{H}^m$ 相同的结果。

（推导毕）

这个 ParVI 方法被称为平滑函数的梯度流方法（gradient flow with smoothed functions，GFSF）。注意到上面优化问题的目标函数符合上一节 5.3.2 中所分析的形式 $\mathbb{E}_q[\mathcal{L}(\phi)]$（其中 $\mathcal{L}$ 是线性变换），因此这里所提的 GFSF 方法所使用的平滑函数的操作是与平滑密度相等价的。GFSF 方法与 SVGD 方法在都写成矩阵形式时，可发现它们之间一个有趣的联系：$\hat{V}^{\text{GFSF}} = \hat{R} + \hat{K}'\hat{K}^{-1}$，而 $\hat{V}^{\text{SVGD}} = \hat{R}\hat{K} + \hat{K}'$，其

中 $\hat{R}_{:,i} := \nabla \log p(x^{(i)})$，因此 $\hat{V}^{\text{GFSF}} = \hat{V}^{\text{SVGD}} \hat{K}^{-1}$。另外，GFSF 对 $-\nabla \log q$ 的估计方式与 Li 等人[179] 所使用的方法一致。他们的推导中使用的是斯坦因恒等式（Stein's identity）在 RKHS 空间上根据直觉的直接推广，而这个推广的合法性仍需进一步考量。在实际应用中，GFSF 方法会在为矩阵 $\hat{K}$ 求逆之前为之加上一个小的对角矩阵以保证数值稳定性。这个操作在 Li 等人[179] 的工作中也被采用。

参见注释 5.1，所有的 ParVI 方法都在近似模拟与郎之万动力学系统（Langevin dynamics, LD）[59] 相同的过程。但 ParVI 方法通过使用平滑核函数直接考虑了粒子之间的相互作用，所以这些方法中的每一个粒子都知道其他粒子的位置，因而这组粒子可以共同为目标分布构建一个好的近似。因此，ParVI 方法相比使用随机性模拟的 LD 方法具有更好的粒子高效性（particle efficiency）。另外，在 LD 方法中已有一些工作[18,60] 使用了随机梯度进行模拟，以提高处理大规模数据的能力。对于这个做法，人们已经认识到由随机梯度带来的噪声是 LD 中布朗运动对应的噪声的高阶小量[113]，因而只要使用足够小的离散步长，使用随机梯度的做法便不会对模拟结果产生任何实质性的影响。而由 ParVI 方法与 LD 的上述对应关系，ParVI 方法便也可以使用随机梯度进行模拟，从而提高可扩展性。

5.4　沃瑟斯坦空间上的一阶加速方法

上一节已经在沃瑟斯坦梯度流的视角下为 ParVI 方法建立了一个统一的理论框架。但对梯度流的直接模拟对应着优化领域中的梯度下降方法（gradient descent），但除此之外，优化领域中所使用的例如涅斯捷洛夫加速方法（Nesterov's acceleration method）[77] 则可以取得比梯度下降方法更快的收敛速度。而涅斯捷洛夫加速方法在黎曼流形上的推广也取得了进展，例如黎曼加速梯度方法（Riemannian accelerated gradient, RAG）[123] 以及黎曼-涅斯捷洛夫方法（Riemannian Nesterov's method, RNes）[124]。本节希望可以利用这些技术来提高 ParVI 方法的表现。不过，利用这些技术需要得到沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 更多的黎曼几何结构。

5.4.1　沃瑟斯坦空间上的指数映射和平行移动

利用 RAG 和 RNes 方法都需要知道沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的*指数映射*（exponential map），指数映射的逆映射（逆指数映射），以及*平行移动*（parallel transport）[①]。如图 5.2 所示，沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的指数映射 $\text{Exp}_q : T_q\mathscr{P}_2(\mathscr{M}) \to \mathscr{P}_2(\mathscr{M})$ 是将点 q 沿着给定方向的测地线（geodesic；可视作“直线”在黎曼流形

① 更加具体地，是此黎曼流形上列维-奇维塔联络（Levi-Civita connection）下的指数映射与平行移动。由于列维-奇维塔联络完全由黎曼结构确定，因此本节中所考虑的指数映射与平行移动也完全由黎曼结构确定。

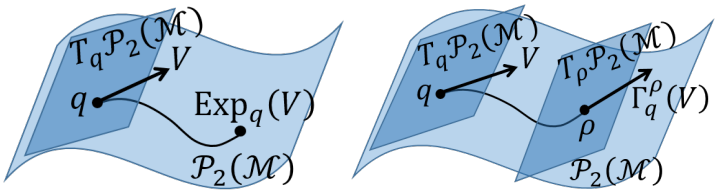


图 5.2　指数映射（左）与平行移动（右）两概念的图示。

上的推广）移动到另一点的操作，而平行移动 $\Gamma_q^\rho : T_q\mathscr{P}_2(\mathscr{M}) \to T_\rho\mathscr{P}_2(\mathscr{M})$ 则是将点 q 处的切向量移动到点 ρ 使之成为点 ρ 处的一个切向量，并且要求在移动过程中保持一个特定意义下的平行。本节将研究沃瑟斯坦空间上的这些概念，并给出使用有限多个粒子的可行的估计方式。

指数映射　对于绝对连续的分布 q，其指数映射可表示为 $\mathrm{Exp}_q(V) = (\mathrm{id} + V)_{\#}q$（可参见 Villani 的著作[53] 推论 7.22，Ambrosio 等人的著作[54] 命题 8.4.6，或 Erbar 等人的著作[176] 命题 2.1）。而由分布的前推的定义，由这个表达式可以很容易得到使用有限多个粒子实现指数映射的方法：若 $\{x^{(i)}\}_i$ 是分布 q 的一组样本，则 $\{x^{(i)} + V(x^{(i)})\}_i$ 是分布 $\mathrm{Exp}_q(V)$ 的一组样本。

逆指数映射　此处将首先推导沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的逆指数映射的准确表达式，然后再为其开发可使用有限多个粒子的计算方法。给定沃瑟斯坦空间上两个分布 $q, \rho \in \mathscr{P}_2(\mathscr{M})$，逆指数映射 $\mathrm{Exp}_q^{-1}(\rho)$ 定义为从 q 到 ρ 的 $\mathscr{P}_2(\mathscr{M})$ 上的测地线 $q_{t\in[0,1]}$ 在点 q 处的切向量。当 q 绝对连续时，从分布 q 到分布 ρ 的最优传输映射（optimal transport map）$\mathcal{T}_q^\rho$ 总是存在的（可参见 Villani 的著作[53] 定理 10.38）。这个绝对连续的要求对 ParVI 方法所考虑的情况是适用的，因为上一节 5.3 的分析结果表明 ParVI 方法都需要平滑操作，而不论哪种平滑方式，都等价于平滑密度，因此 ParVI 方法中所考虑的变分分布 q 是绝对连续的。在此情况下，上述测地线 $q_{t\in[0,1]}$ 可使用最优传输映射表示为 $q_t = \left((1-t)\,\mathrm{id} + t\mathcal{T}_q^\rho\right)_{\#}q$（可参见 Ambrosio 等人的著作[54] 定理 7.2.2），且此测地线在 q 处（即在 $t = 0$ 时）的切向量可以表示为 $\lim_{t\to 0}\frac{1}{t}(\mathcal{T}_q^{q_t} - \mathrm{id})$（视作 $\mathscr{M} = \mathbb{R}^m$ 上的一个向量场）（可参见 Ambrosio 等人的著作[54] 命题 8.4.6）。而由最优传输映射的唯一性，可有 $\mathcal{T}_q^{q_t} = (1-t)\,\mathrm{id} + t\mathcal{T}_q^\rho$，这使得上面的极限变为 $\mathcal{T}_q^\rho - \mathrm{id}$。因而可有 $\mathrm{Exp}_q^{-1}(\rho) = \mathcal{T}_q^\rho - \mathrm{id}$。

为使用有限多个粒子来估计逆指数映射，可以使用分布 q 的一组样本 $\{x^{(i)}\}_{i=1}^N$ 和分布 ρ 的一组样本 $\{y^{(i)}\}_{i=1}^N$ 来计算这两组样本之间的离散最优传输映射，并以此映射作为对两分布间最优传输映射 $\mathcal{T}_q^\rho$ 的近似。然而，计算两组样本之间的离散最优传输映射仍然是一个计算代价很大的任务[180]。虽然也有一些计算代价

更小的近似解法，例如著名的 Sinkhorn 方法[181] 及其改进版本[182]，但它也需要 $\mathcal{O}(N^2)$ 级别的经验计算复杂度，而且实验也发现它所产生的结果过于不稳定。下面考虑一种计算更加便捷且实验中的表现也更加稳定的近似方法。此近似需要假设分布 q 和 p 的这两组样本是成对近邻的（pairwisely close）：$d(x^{(i)}, y^{(i)}) \ll \min\big\{\min_{j\neq i} d(x^{(i)}, x^{(j)}), \min_{j\neq i} d(y^{(i)}, y^{(j)})\big\}$。这个成对近邻的条件意味着对于任意的 $i \neq j$，都有 $\dfrac{d(x^{(i)}, x^{(j)})}{d(x^{(j)}, y^{(j)})} \gg 1$。而另一方面，由三角不等式（triangle inequality）可知，$d(x^{(i)}, y^{(j)}) \geqslant \big|d(x^{(i)}, x^{(j)}) - d(x^{(j)}, y^{(j)})\big|$，即 $\dfrac{d(x^{(i)}, y^{(j)})}{d(x^{(j)}, y^{(j)})} \geqslant \left|\dfrac{d(x^{(i)}, x^{(j)})}{d(x^{(j)}, y^{(j)})} - 1\right|$。综合两者可以得到 $\dfrac{d(x^{(i)}, y^{(j)})}{d(x^{(j)}, y^{(j)})} \gg 1$，并进一步交换指标 i 和 j 得到 $d(x^{(i)}, y^{(i)}) \ll \min_{j\neq i} d(x^{(i)}, y^{(j)})$。这个结果意味着，映射 $x^{(i)} \mapsto y^{(i)}, \forall i$ 是从 $\{x^{(i)}\}_i$ 传输到 $\{y^{(i)}\}_i$ 的过程中产生的传输成本最小的映射。更加详细地说，在上述映射的基础上，考虑从点 $x^{(i)}$ 向点 $y^{(j)}$（其中 $j \neq i$）增加一个单位传输量。这个传输量将会使总的传输成本增加 $d(x^{(i)}, y^{(j)}) - d(x^{(i)}, y^{(i)}) + d(x^{(j)}, y^{(i)}) - d(x^{(j)}, y^{(j)})$，而这个量总是正的。因此这种情况下可以合理地将最优传输映射 $\mathcal{T}_q^\rho$ 近似为离散传输映射 $\mathcal{T}_q^\rho(x^{(i)}) \approx y^{(i)}$，进而得到计算逆指数映射的方法：

命题 5.1 (逆指数映射)：对于分布 q 与分布 ρ 的成对近邻样本 $\{x^{(i)}\}_i$ 与 $\{y^{(i)}\}_i$，可有结论 $\big(\mathrm{Exp}_q^{-1}(\rho)\big)(x^{(i)}) \approx y^{(i)} - x^{(i)}$。

此结论的推导过程已由上文给出。

平行移动　首先注意到，关于沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的平行移动已有一些正式的研究[183-184]，但这些结果并不适合使用有限多个粒子来估计。本节将使用希尔德之梯（Schild's ladder）[185-186] 这个平行移动的一阶近似方法来计算沃瑟斯坦空间上的平行移动。由于希尔德之梯方法只涉及到指数映射及其逆映射，而沃瑟斯坦空间上的这两个映射的有限多个粒子的估计方法已经由上面给出，因此最终可以得到使用有限多个粒子的平行移动估计方法。

命题 5.2 (平行移动)：对于分布 q 与分布 ρ 的成对近邻样本 $\{x^{(i)}\}_i$ 与 $\{y^{(i)}\}_i$，可有结论 $\big(\Gamma_q^\rho(V)\big)(y^{(i)}) \approx V(x^{(i)}), \forall V \in T_q\mathscr{P}_2(\mathscr{M})$。

证明 本证明首先在沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 的背景下介绍希尔德之梯方法（Schild's ladder method）[185-186] 来估计将 q 处的切向量 $V \in T_q\mathscr{P}_2(\mathscr{M})$ 平行移动到 ρ 处的切向量 $\Gamma_q^\rho(V)$。如图 5.3 所示，给定分布 q，ρ 以及 q 处的切向量 $V \in T_q\mathscr{P}_2(\mathscr{M})$，估计 $\Gamma_q^\rho(V)$ 的步骤为：

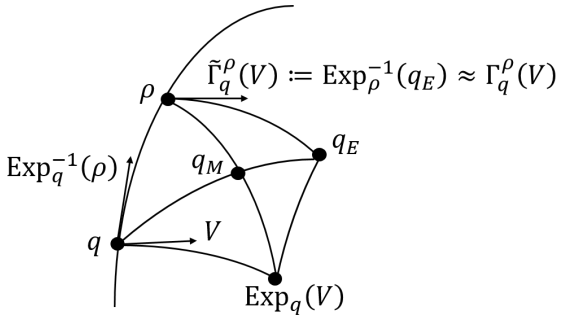


图 5.3　希尔德之梯方法（Schild's ladder method）的图示（改编自 [186]）。

1. 通过指数映射找到 $\mathrm{Exp}_q(V)$；
2. 找到从 $\mathrm{Exp}_q(V)$ 到 ρ 的测地线的中点：$q_M := \mathrm{Exp}_{\mathrm{Exp}_q(V)}\left(\frac{1}{2}\mathrm{Exp}_{\mathrm{Exp}_q(V)}^{-1}(\rho)\right)$；
3. 将从 q 到 q_M 的测地线外推至两倍长度从而找到其终点：$q_E := \mathrm{Exp}_q(2\,\mathrm{Exp}_q^{-1}(q_M))$；
4. 最后，对平行移动后的切向量的近似可取为：$\tilde{\Gamma}_q^\rho(V) := \mathrm{Exp}_\rho^{-1}(q_E) \approx \Gamma_q^\rho(V)$。

综合上述过程，希尔德之梯方法所给出的近似结果可以表达为：

$$\tilde{\Gamma}_q^\rho(V) = \mathrm{Exp}_\rho^{-1}\left(\mathrm{Exp}_q\left(2\,\mathrm{Exp}_q^{-1}\left(\mathrm{Exp}_{\mathrm{Exp}_q(V)}\left(\frac{1}{2}\mathrm{Exp}_{\mathrm{Exp}_q(V)}^{-1}(\rho)\right)\right)\right)\right).$$

此方法所给出的近似 $\tilde{\Gamma}_q^\rho$ 是对 Γ_q^ρ 的一阶近似[186]：$\left\|\Gamma_q^\rho - \tilde{\Gamma}_q^\rho\right\|_{T_\rho\mathscr{P}_2(\mathscr{M})} = o(d_W(q,\rho))$。另外，由上述步骤及表达式可以看出，希尔德之梯方法只需要知道指数映射及其逆映射即可。

假设分布 q 和 ρ 在沃瑟斯坦距离的意义下十分接近，使得希尔德之梯方法能够给出一个很好的近似。下面考虑对于很小的 $\varepsilon > 0$，通过使用 q 和 ρ 的成对近邻样本 $\{x^{(i)}\}_{i=1}^N$ 和 $\{y^{(i)}\}_{i=1}^N$ 将 εV 进行平行移动。而由平行移动的线性性 $\Gamma_q^\rho(\varepsilon V) = \varepsilon\Gamma_q^\rho(V)$，最终可以得到 V 的平行移动 $\Gamma_q^\rho(V)$。

1. 由上一部分得到的结果，分布 $\mathrm{Exp}_q(\varepsilon V)$ 可由 $(\mathrm{id}+\varepsilon V)_{\#}q$ 来表示，因此 $\{x^{(i)} + \varepsilon V(x^{(i)})\}_{i=1}^N$ 可看作 $\mathrm{Exp}_q(\varepsilon V)$ 的一组样本，并且由于 ε 很小，这组样本与 $\{y^{(i)}\}_i$ 仍然是成对近邻的。
2. 由成对近邻的条件，从 ρ 到 $\mathrm{Exp}_q(\varepsilon V)$ 的最优传输映射 $\mathcal{T}$ 可以表示为 $\mathcal{T}(y^{(i)}) = x^{(i)} + \varepsilon V(x^{(i)})$。有了最优传输映射后，由 Ambrosio 等人的著作[54] 定理 7.2.2 可知，从 ρ 到 $\mathrm{Exp}_q(\varepsilon V)$ 的测地线可以表示为 $t \mapsto \left((1-t)\,\mathrm{id} + t\mathcal{T}\right)_{\#}\rho$。将 t 取为 $\frac{1}{2}$，可得此测地线的中点 q_M 的一组样本为 $\left\{\frac{1}{2}\left(y^{(i)} + x^{(i)} + \varepsilon V(x^{(i)})\right)\right\}_i$。
3. 类似地，可以找到 q_E 的一组样本为 $\left\{(1-t)x^{(i)} + \frac{1}{2}t\left(y^{(i)} + x^{(i)} + \varepsilon V(x^{(i)})\right)\right\}_i\Big|_{t=2} = \left\{y^{(i)} + \varepsilon V(x^{(i)})\right\}_i$，并且这组样本与 $\{y^{(i)}\}_i$ 仍然是成对近邻的。

4. 用来近似 $\Gamma_q^\rho(\varepsilon V)$ 的 ρ 处的切向量可表示为 $\big(\mathrm{Exp}_\rho^{-1}(q_E)\big)(y^{(i)}) = \varepsilon V(x^{(i)})$。最后，原切向量 V 的平行移动便可表示为 $\big(\Gamma_q^\rho(V)\big)(y^{(i)}) \approx V(x^{(i)})$。 □

这里所给出的使用有限多个粒子估计沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的指数映射和平行移动的方法看上去并未涉及复杂的几何概念。事实上，这是因为沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 的几何性质是由其支撑空间 $\mathscr{M}$ 所决定的。此处所考虑的是欧氏支撑空间，它是平坦的，因而 $\mathscr{P}_2(\mathscr{M})$ 便也会表现得平坦，从而可有上述各结论。而这些结论在带有曲率的黎曼流形支撑空间 $\mathscr{M}$ 上的推广也将会出现这个黎曼流形上的一些非平凡的几何结构。

5.4.2　ParVI 方法的加速框架

有了沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的指数映射（及其逆映射）和平行移动之后，便可以将 RAG 和 RNes 方法应用于沃瑟斯坦空间上。本节先在沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 的背景下介绍并简化这两个黎曼流形上的加速方法，再通过这两个方法得到 ParVI 方法的加速框架。

RAG 和 RNes 两方法为在沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上关于变量 q 来最小化 $\mathrm{KL}_p(q)$，都引入了辅助变量（auxiliary variable）ρ，并且会在 ρ 的位置上计算 KL 散度的梯度。并且二者在第 k 轮迭代时，更新变量 q 的方式是一样的：

$$q_k = \mathrm{Exp}_{\rho_{k-1}}(\varepsilon V_{k-1}),$$

其中 ε 为离散步长，而 $V_{k-1} := -\operatorname{grad}\mathrm{KL}_p(\rho_{k-1})$ 则可由各 ParVI 方法估计。至于更新辅助变量 ρ 的方式，原本的 RAG 方法需要在每一步更新中求解一个非线性方程。本节在此先通过使用一个近似[①]来将这个过程简化。参考如下推导过程，化简之后的 RAG 方法更新辅助变量 ρ 的方式为：

$$\rho_k = \mathrm{Exp}_{q_k}\left[-\Gamma_{\rho_{k-1}}^{q_k}\left(\frac{k-1}{k}\mathrm{Exp}_{\rho_{k-1}}^{-1}(q_{k-1}) - \frac{k+\alpha-2}{k}\varepsilon V_{k-1}\right)\right],$$

其中 $\alpha > 3$ 是 RAG 方法中需指定的加速因子。

推导　RAG 方法所提出的更新辅助变量 ρ 的方式是需要从下面的非线性方程中解得 ρ_k（参见 RAG 的原论文[123] 算法 2 及式 (5)）：

$$\Gamma_{\rho_k}^{\rho_{k-1}}\left(\frac{k}{\alpha-1}\mathrm{Exp}_{\rho_k}^{-1}(q_k) + \frac{mV_k}{\|V_k\|_{\rho_k}}\right)$$

① 此近似其实也在 RAG 的原始工作[123] 所展示的实际应用中使用，但并未给出一般的简化结果。

$$= \frac{k-1}{\alpha-1}\operatorname{Exp}_{\rho_{k-1}}^{-1}(q_{k-1}) + \frac{mV_{k-1}}{\|V_{k-1}\|_{\rho_{k-1}}} - \frac{k+\alpha-2}{\alpha-1}\varepsilon V_{k-1}.$$

为等式两边同时作用平行移动的逆映射 $\left(\Gamma_{\rho_k}^{\rho_{k-1}}\right)^{-1}$ 并注意到 $\left(\Gamma_{\rho_k}^{\rho_{k-1}}\right)^{-1} = \Gamma_{\rho_{k-1}}^{\rho_k}$，上面的方程可以化为：

$$\begin{aligned}
&\frac{k}{\alpha-1}\operatorname{Exp}_{\rho_k}^{-1}(q_k) + \frac{mV_k}{\|V_k\|_{\rho_k}} \\
&= \Gamma_{\rho_{k-1}}^{\rho_k}\left(\frac{k-1}{\alpha-1}\operatorname{Exp}_{\rho_{k-1}}^{-1}(q_{k-1}) - \frac{k+\alpha-2}{\alpha-1}\varepsilon V_{k-1}\right) + \frac{m\Gamma_{\rho_{k-1}}^{\rho_k}(V_{k-1})}{\|V_{k-1}\|_{\rho_{k-1}}}.
\end{aligned}$$

接下来，考虑将等式左边的 V_k 近似为 $\Gamma_{\rho_{k-1}}^{\rho_k}(V_{k-1})$。再利用性质 $\|V_{k-1}\|_{\rho_{k-1}} = \left\|\Gamma_{\rho_{k-1}}^{\rho_k}(V_{k-1})\right\|_{\rho_k}$，上面的方程可以化为：

$$\frac{k}{\alpha-1}\operatorname{Exp}_{\rho_k}^{-1}(q_k) = \Gamma_{\rho_{k-1}}^{\rho_k}\left(\frac{k-1}{\alpha-1}\operatorname{Exp}_{\rho_{k-1}}^{-1}(q_{k-1}) - \frac{k+\alpha-2}{\alpha-1}\varepsilon V_{k-1}\right).$$

利用性质 $\operatorname{Exp}_{\rho_k}^{-1}(q_k) = -\Gamma_{q_k}^{\rho_k}(\operatorname{Exp}_{q_k}^{-1}(\rho_k))$ 并在等式两边同时作用 $\left(\Gamma_{q_k}^{\rho_k}\right)^{-1} = \Gamma_{\rho_k}^{q_k}$，可以得到：

$$\operatorname{Exp}_{q_k}^{-1}(\rho_k) = -\Gamma_{\rho_k}^{q_k}\Gamma_{\rho_{k-1}}^{\rho_k}\left(\frac{k-1}{k}\operatorname{Exp}_{\rho_{k-1}}^{-1}(q_{k-1}) - \frac{k+\alpha-2}{k}\varepsilon V_{k-1}\right).$$

最后，将 $\Gamma_{\rho_k}^{q_k}\Gamma_{\rho_{k-1}}^{\rho_k}$ 近似为 $\Gamma_{\rho_{k-1}}^{q_k}$，得到的更新辅助变量 ρ 的近似方式为：

$$\rho_k = \operatorname{Exp}_{q_k}\left[-\Gamma_{\rho_{k-1}}^{q_k}\left(\frac{k-1}{k}\operatorname{Exp}_{\rho_{k-1}}^{-1}(q_{k-1}) - \frac{k+\alpha-2}{k}\varepsilon V_{k-1}\right)\right].$$

（推导毕）

至于 RNes 方法，本文重新组织了它更新辅助变量 ρ 的方式：

$$\rho_k = \operatorname{Exp}_{q_k}\left\{c_1 \operatorname{Exp}_{q_k}^{-1}\left[\operatorname{Exp}_{\rho_{k-1}}\left((1-c_2)\operatorname{Exp}_{\rho_{k-1}}^{-1}(q_{k-1}) + c_2 \operatorname{Exp}_{\rho_{k-1}}^{-1}(q_k)\right)\right]\right\},$$

其中 $c_1, c_2 \in \mathbb{R}$ 是本文所引入的新的算法参数。在 RNes 方法的原本形式（参见 Zhang 等人的著作[124] 算法 2）中，算法是由缩减参数 $\beta \in (0,1)$ 以及目标函数梯度的李普希兹系数（Lipschitz coefficient）λ 来表示的。这两种表示形式的关系为：

$$c_1 = \frac{\alpha\varsigma}{\varsigma + \alpha\lambda}, c_2 = \frac{1}{\alpha},$$

其中

$$\begin{aligned}
\alpha &= \left(\sqrt{\beta^2 + 4(1+\beta)\lambda\varepsilon} - \beta\right)/2, \\
\varsigma &= \lambda\left(\sqrt{\beta^2 + 4(1+\beta)\lambda\varepsilon} - \beta\right) / \left(\sqrt{\beta^2 + 4(1+\beta)\lambda\varepsilon} + \beta\right),
\end{aligned}$$

算法 4 ParVI 方法的加速框架：沃瑟斯坦加速梯度方法（Wasserstein accelerated gradient, WAG）和沃瑟斯坦-涅斯捷洛夫方法（Wasserstein Nesterov's method, WNes）

WAG 方法：选定加速因子 $\alpha > 3$；

WNes 方法：选定算法系数 c_1 和 c_2；

随机地初始化一组各不相同的粒子 $\{x_0^{(i)}\}_{i=1}^N$；以相同的值初始化辅助粒子 $\{y_0^{(i)}\}_{i=1}^N = \{x_0^{(i)}\}_{i=1}^N$；

对 $k = 1, 2, \cdots, k_{\text{max}}$, **执行操作:**

　　对 $i = 1, \cdots, N$, **执行操作:**

　　使用 SVGD/Blob/GFSD/GFSF 方法估计 $V(y_{k-1}^{(i)})$；

　　$x_k^{(i)} = y_{k-1}^{(i)} + \varepsilon V(y_{k-1}^{(i)})$;

　　$$y_k^{(i)} = x_k^{(i)} + \begin{cases} \text{WAG: } \dfrac{k-1}{k}(y_{k-1}^{(i)} - x_{k-1}^{(i)}) + \dfrac{k+\alpha-2}{k}\varepsilon V(y_{k-1}^{(i)}); \\ \text{WNes: } c_1(c_2-1)(x_k^{(i)} - x_{k-1}^{(i)}); \end{cases}$$

　　结束

结束

输出所得粒子 $\{x_{k_{\text{max}}}^{(i)}\}_{i=1}^N$ 作为对分布 p 的估计。

是 RNes 方法的原工作[124] 中引入的另外两个量。特别地，在下面将要得到的加速框架的算法（算法 4）中，更新规则里所出现的系数 $c_1(c_2 - 1)$ 可以表示为：

$$c_1(c_2 - 1) = 1 + \beta - \frac{2(1+\beta)(2+\beta)\lambda\varepsilon}{\sqrt{\beta^2 + 4(1+\beta)\lambda\varepsilon} - \beta + 2(1+\beta)\lambda\varepsilon}.$$

由于实际中对目标函数梯度的李普希兹系数没有准确的信息，因此调整原参数 β 和 λ 与调整本节所提参数 c_1 和 c_2 是等价的。最后，RNes 方法在欧氏空间的情况下可以化简为标准的涅斯捷洛夫加速方法。

做好了这些准备，现在便可将沃瑟斯坦空间上估计指数映射和平行移动的方法代入，从而得到加速 ParVI 方法的框架。基于 RAG 方法和 RNes 方法的加速框架实现方式分别被称为沃瑟斯坦加速梯度方法（Wasserstein accelerated gradient, WAG）和沃瑟斯坦-涅斯捷洛夫方法（Wasserstein Nesterov's method, WNes）。经过下面的推导，可得到加速框架的这两个实现方式，列于算法 4 中。注意到在推导过程中，两组粒子成对近邻的条件始终可以得到满足，因为 $\{x^{(i)}\}_i$ 和 $\{y^{(i)}\}_i$ 这两组粒子初始化为成对相同的状态，而之后每一组粒子都是基于另一组粒子以成对的方式更新由步长所控制的一个小量，且两组粒子是交替进行更新的。因此命题 5.1 和命题 5.2 是可以合理使用的。

推导　加速框架的这两个实现方式的推导是类似的，因而此处只给出由 RAG 推导出 WAG 算法的过程。记向量场 $\zeta_{k-1} := \frac{k-1}{k}\operatorname{Exp}_{\rho_{k-1}}^{-1}(q_{k-1}) - \frac{k+\alpha-2}{k}\varepsilon V_{k-1}$，这样 RAG 方法对辅助变量 ρ 的更新方式便可写作 $\rho_k = \operatorname{Exp}_{q_k}\left[-\Gamma_{\rho_{k-1}}^{q_k}(\zeta_{k-1})\right]$。首先假设 q_{k-1} 的样本 $\{x_{k-1}^{(i)}\}_{i=1}^N$ 与 ρ_{k-1} 的样本 $\{y_{k-1}^{(i)}\}_{i=1}^N$ 是成对近邻的，因而可以使用命题 5.1 来估计逆指数映射 $\operatorname{Exp}_{\rho_{k-1}}^{-1}(q_{k-1})(y_{k-1}^{(i)}) = x_{k-1}^{(i)} - y_{k-1}^{(i)}$，从而有 $\zeta_{k-1}(y_{k-1}^{(i)}) = \frac{k-1}{k}(x_{k-1}^{(i)} - y_{k-1}^{(i)}) - \frac{k+\alpha-2}{k}\varepsilon V_{k-1}^{(i)}$。而由 q_k 的更新方式 $x_k^{(i)} = y_{k-1}^{(i)} + \varepsilon V_{k-1}^{(i)}$，可以得知 q_k 的样本 $\{x_k^{(i)}\}_{i=1}^N$ 和 ρ_{k-1} 的样本 $\{y_{k-1}^{(i)}\}_{i=1}^N$ 对于足够小的 $\varepsilon > 0$ 也是成对近邻的。接下来使用命题 5.2 来估计平行移动 $\left(\Gamma_{\rho_{k-1}}^{q_k}(\zeta_{k-1})\right)(x_k^{(i)}) \approx \zeta_{k-1}(y_{k-1}^{(i)})$，代入到 ρ_k 的更新规则中，可得：$y_k^{(i)} = x_k^{(i)} - \left(\Gamma_{\rho_{k-1}}^{q_k}(\zeta_{k-1})\right)(x_k^{(i)}) \approx x_k^{(i)} - \zeta_{k-1}(y_{k-1}^{(i)}) = x_k^{(i)} - \frac{k-1}{k}(x_{k-1}^{(i)} - y_{k-1}^{(i)}) + \frac{k+\alpha-2}{k}\varepsilon V_{k-1}^{(i)}$。这样得到的 $\{y_k^{(i)}\}_{i=1}^N$ 即为 ρ_k 的样本。

下面来检查成对近邻条件是否成立。由算法 4 的初始化方式可知在 $k = 0$ 时有 $x_0^{(i)} = y_0^{(i)}$，显然成对近邻。假设 $\{x_{k-1}^{(i)}\}_{i=1}^N$ 和 $\{y_{k-1}^{(i)}\}_{i=1}^N$ 是成对近邻的，那么对于足够小的 $\varepsilon > 0$，$\zeta_{k-1}(y_{k-1}^{(i)})$ 对于任意 i 都是一个各分量都很小的向量。而由上面所得到的更新规则，这意味着 q_k 的样本 $\{x_k^{(i)}\}_{i=1}^N$ 和 ρ_k 的样本 $\{y_k^{(i)}\}_{i=1}^N$ 是成对近邻的，这提供了为下一轮递推的条件。由归纳法原理可知，对于足够小的 $\varepsilon > 0$，成对近邻条件成立，进而保证了算法 4 的正确性。

（推导毕）

所提加速框架 WAG 和 WNes 继承了 RAG 和 RNes 方法相比原本的梯度流模拟收敛更快的理论保证。根据上一节所得到的理论，各 ParVI 方法从模拟沃瑟斯坦梯度流的角度来看是等价的，因而所提加速框架可以应用于所有 ParVI 方法上。在实际中，所提加速框架在每轮迭代中增加的计算复杂度是线性于粒子数目 N 的，这与 ParVI 方法原本的 $\mathcal{O}(N^2)$ 量级的复杂度相比并不是一个很明显的计算负担。另外，需要强调的是，直接将欧氏空间中的标准涅斯捷洛夫加速方法应用于每一个粒子上这种做法是没有理论支持的，因为每个粒子自身并没有在最优化一个 $\mathscr{M}$ 上的目标函数，而是它们共同所表示的分布在最小化沃瑟斯坦空间上的 KL 散度。最后，本节所得到的沃瑟斯坦空间上更加细致的黎曼结构（命题 5.1 和命题 5.2）可使将其他黎曼流形上的优化技术应用于沃瑟斯坦空间上成为可能，例如黎曼流形上的 BFGS 方法[187-189] 以及黎曼流形上的随机梯度方差消减（variance reduction）方法[156]。这些拓展可以进一步提升 ParVI 方法的表现。

5.5　基于热方程的带宽选择方法

在 5.3 节中所做的理论分析指出，ParVI 方法都需要做一个平滑操作，而这个操作可以通过使用核函数来实现。所以选择核函数的带宽就成为了一个必要且很关键的问题。SVGD 方法使用的是中位数方法（median method）[81]，它是基于数值稳定性上的直觉得到的方法。这里考虑一个更加具有原则性的带宽选择方法。本节首先分析平滑操作的目标并给出带宽应满足的原则，进而推导出一个可行的算法。

如注释 5.1 所述，确定性动力学系统 $\mathrm{d}x = V^{\mathrm{GF}}(x)\,\mathrm{d}t$ 与郎之万动力学系统会产生同样的分布演化规律。特别地，确定性子动力学系统 $\mathrm{d}x = -\nabla \log q_t(x)\,\mathrm{d}t$ 与布朗运动 $\mathrm{d}x = \sqrt{2}\,\mathrm{d}B_t(x)$ 会产生同样的分布演化规律，而这个规律可由热方程（heat equation, HE）来描述：$\partial_t q_t(x) = \Delta q_t(x)$。所以一个好的平滑核函数应使所对应的近似动力学系统所产生的分布演化规律与热方程相契合。这便是选择平滑核函数带宽的原则。

下面依据此原则为 GFSD 方法推导出具体的确定带宽的方法。GFSD 方法使用平滑密度这种平滑操作，即将分布 q_t 近似为平滑分布 $\tilde{q}(x) = \tilde{q}(x; \{x^{(i)}\}_i) = \dfrac{1}{N}\sum_{i=1}^{N} K(x, x^{(i)})$。对于很小的 $\varepsilon > 0$，它所对应的近似动力学系统 $\mathrm{d}x = -\nabla \log \tilde{q}(x)\,\mathrm{d}t$ 会把分布 q_t 的样本 $\{x^{(i)}\}_i$ 移动为 $\{x^{(i)} - \varepsilon \nabla \log \tilde{q}(x^{(i)})\}_i$，而这组粒子近似是分布 $q_{t+\varepsilon}$ 的样本。而另一方面，由 HE 所给出的分布演化规律，上面的新分布 $q_{t+\varepsilon}$（作为密度函数）应该可以由函数 $q_t + \varepsilon \partial_t q_t \approx \tilde{q} + \varepsilon \Delta \tilde{q}$ 来近似。由带宽选择的原则，这两个近似应该尽量保持一致，即 $\tilde{q}\big(x; \{x^{(i)} - \varepsilon \nabla \log \tilde{q}(x^{(i)})\}_i\big)$ 应与 $\tilde{q} + \varepsilon \Delta \tilde{q}$ 尽量接近。将这两个表达式对 ε 展开，可以发现二者的零阶项是一样的，而匹配二者的一阶项即为使如下函数尽可能处处接近于零：$\mathcal{G}(x) := \Delta \tilde{q}(x; \{x^{(i)}\}_i) + \sum_j \nabla_{x^{(j)}} \tilde{q}(x; \{x^{(i)}\}_i) \cdot \nabla \log \tilde{q}(x^{(j)}; \{x^{(i)}\}_i)$。为在实际中达到此目标，可以提出如下优化问题：

$$\min_{w>0} \frac{N}{w^{m+2}} \mathbb{E}_{q(x)}[\mathcal{G}(x)^2] \approx \frac{1}{w^{m+2}} \sum_k \mathcal{G}(x^{(k)})^2,$$

其中 w 表示带宽，而系数 $\dfrac{1}{w^{m+2}}$ 的引入是为了使最终的目标函数成为一个无量纲的量（注意到 x^2/w 是无量纲的），从而减小问题维度对目标函数的影响。以平滑分布 $\tilde{q}$ 直接表示的上述优化问题中的目标函数为：

$$\frac{1}{w^{m+2}} \sum_k \mathcal{G}(x^{(k)})^2$$

$$= \frac{1}{w^{m+2}} \sum_k \Big[\Delta\tilde{q}(x^{(k)}; \{x^{(i)}\}_i) + \sum_j \nabla_{x^{(j)}} \tilde{q}(x^{(k)}; \{x^{(i)}\}_i) \cdot \nabla \log \tilde{q}(x^{(j)}; \{x^{(i)}\}_i) \Big]^2.$$

下面来逐步实例化这个目标函数。设 $\tilde{q}(x; \{x^{(j)}\}_j) = (1/Z) \sum_j c(\|x - x^{(j)}\|^2/(2w))$，其中 $c : \mathbb{R} \to \mathbb{R}$ 是一个正定函数。则目标函数可写为：

$$\sum_k \Big(\sum_j \Big[c_j''(x)\|x - x^{(j)}\|^2 + mwc_j'(x) + \frac{\big(\sum_i c_{ij}' x^{(i)}\big) - \big(\sum_i c_{ij}'\big)x^{(j)}}{\big(\sum_i c_{ij}\big)} \cdot (x - x^{(j)}) c_j'(x) \Big] \Big)^2,$$

其中上标撇 “′” 表示导数，$c_j'(x) := c'(\|x - x^{(j)}\|^2/(2w))$，$c_{ij}' := c_j'(x^{(i)})$，而 $c_{ij} = c(\|x^{(i)} - x^{(j)}\|^2/(2w))$。进一步，对于高斯核函数，$c(r) = (2\pi w)^{-\frac{m}{2}} e^{-r}$，则上述目标函数可表示为 $\frac{1}{(2\pi)^m} \sum_k \sigma_k^2(w)$，其中

$$\sigma_k(w) = \big(\sum_j e_{kj}\|d_{kj}\|^2\big) - hD\big(\sum_j e_{kj}\big) - \sum_j \big(\sum_i e_{ij}\big)^{-1} e_{jk} d_{jk} \cdot \big(\sum_i e_{ij} d_{ij}\big),$$

且

$$\begin{aligned}
\sigma_k'(w) = \quad & \frac{1}{2w^2}\big(\sum_j e_{jk}\|d_{jk}\|^4\big) - \frac{m}{w}\big(\sum_j e_{jk}\|d_{jk}\|^2\big) + \big(\frac{m^2}{2} - m\big)\big(\sum_j e_{jk}\big) \\
& - \frac{1}{2w^2} \sum_j \big(\sum_i e_{ij}\big)^{-1} e_{jk} d_{jk} \cdot \big(\sum_i e_{ij}\|d_{ij}\|^2 d_{ij}\big) \\
& - \frac{1}{2w^2} \sum_j \big(\sum_i e_{ij}\big)^{-1} e_{jk}\|d_{jk}\|^2 d_{jk} \cdot \big(\sum_i e_{ij} d_{ij}\big) \\
& + \frac{1}{2w^2} \sum_j \big(\sum_i e_{ij}\big)^{-2} \big(\sum_i e_{ij}\|d_{ij}\|^2\big) e_{jk} d_{jk} \cdot \big(\sum_i e_{ij} d_{ij}\big) \\
& + \frac{m}{2w} \sum_j \big(\sum_i e_{ij}\big)^{-1} e_{jk} d_{jk} \cdot \big(\sum_i e_{ij} d_{ij}\big),
\end{aligned}$$

其中 $d_{ij} := x^{(i)} - x^{(j)}$，$e_{ij} := e^{-\|d_{ij}\|^2/(2w) - (m/2)\log w}$。

有了这些信息，便可对此目标函数进行优化了。虽然计算 $\sigma_k(w)$ 需要一些计算代价，但这个优化问题是关于一个标量的，因此可以使用一些线搜索（line search）的方法来提高优化效率。实际中的操作会在每次估计向量场 V 之前通过平方插值这个线搜索方法进行一步更新带宽 w 的操作。这个操作只需要计算一次 $\sigma_k(w)$ 的导数以及两次 $\sigma_k(w)$ 的值。

此带宽选择方法被称为热方程（heat equation）方法，即 HE 方法。这个推导是基于 GFSD 方法所采用的平滑密度方法所得到的，因此同样采用了平滑密度方

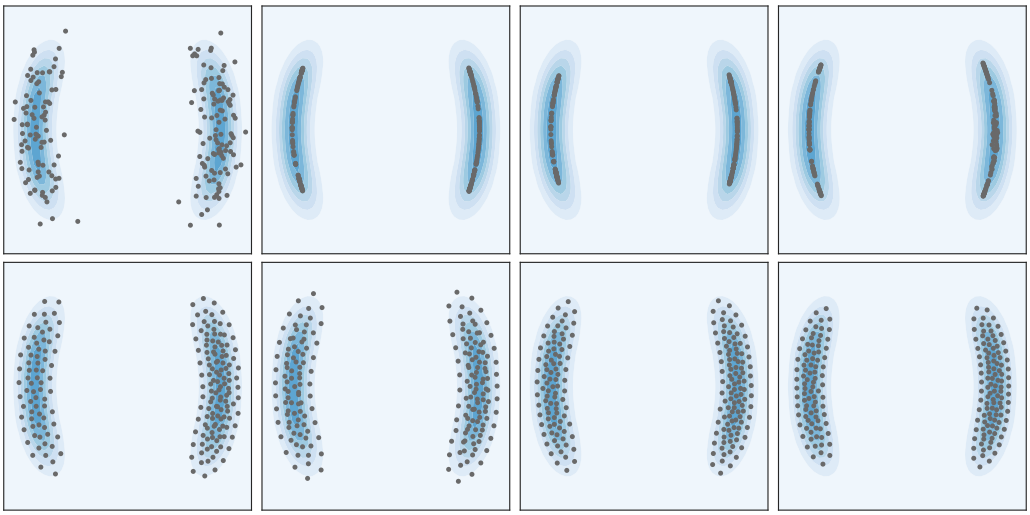


图 5.4　所提 HE 方法（第二行）与中位数方法（第一行）的效果对比。各列所对应的 ParVI 方法分别为 SVGD，Blob，GFSD 和 GFSF。

法的 Blob 方法也可使用此方法。而由 5.3 节中所阐明的平滑密度与平滑函数两操作之间的等价性，此 HE 方法也可用于采用了平滑函数的 ParVI 方法，例如 SVGD 和 GFSF。

5.6　实验

为与 WAG 和 WNes 这两个加速方法的实现方式清晰对应，本节称 ParVI 方法原本所使用的模拟梯度流的方法为沃瑟斯坦梯度下降方法（Wasserstein Gradient Descent, WGD）。此外，尽管 PO 方法并不是以加速的目的而开发的，但这里将它视为一种经验上的加速方法参与比较。以下各实验的代码可从网站“https://github.com/chang-ml-thu/AWGF”下载。

5.6.1　简单模拟实验

首先展示对于选择核函数带宽的任务，使用 HE 方法胜于中位数方法的表现。图 5.4 展示了四种 ParVI 方法（使用原本的 WGD 方法）分别使用 HE 方法和中位数方法在 400 轮更新后所得到的 200 个粒子的分布。每一种方法都以相同的 200 个粒子进行初始化，而这组用来初始化的 200 个粒子采自标准高斯分布 $\mathcal{N}(0,1)$。图中的灰色点即表示最终得到的粒子，而蓝色背景则表示目标分布。这个双峰分布是受 Rezende 等人的工作[72] 中所使用的一个实验场景的启发而设计的。对于二维变量 $Z=(Z_1,Z_2)\in\mathbb{R}^2$，其对数密度函数 $p(Z)$ 为：

$$\log p(Z) = -2(\|Z\|_2^2-3)^2+\log(e^{-2(Z_1-3)^2}+e^{-2(Z_1+3)^2})+\text{const.}$$

表 5.1　贝叶斯逻辑回归模型在 Covertype 数据集上的后验推理实验中各 ParVI 方法所采用的参数

	WGD	PO	WAG	WNes
SVGD	3e-2	(1.0, 0.7, 1e-7), 3e-6	3.9, (0.9, 1e-6)	(300, 0.2), (0.8, 3e-4)
Blob	1e-6	(1.0, 0.7, 1e-7), 3e-7	3.9, (0.9, 1e-6)	(1000, 0.2), (0.9, 1e-5)
GFSD	1e-6	(1.0, 0.7, 1e-7), 3e-7	3.9, (0.9, 1e-6)	(1000, 0.2), (0.9, 1e-5)
GFSF	1e-6	(1.0, 0.7, 1e-7), 3e-7	3.9, (0.9, 1e-6)	(1000, 0.2), (0.9, 1e-5)

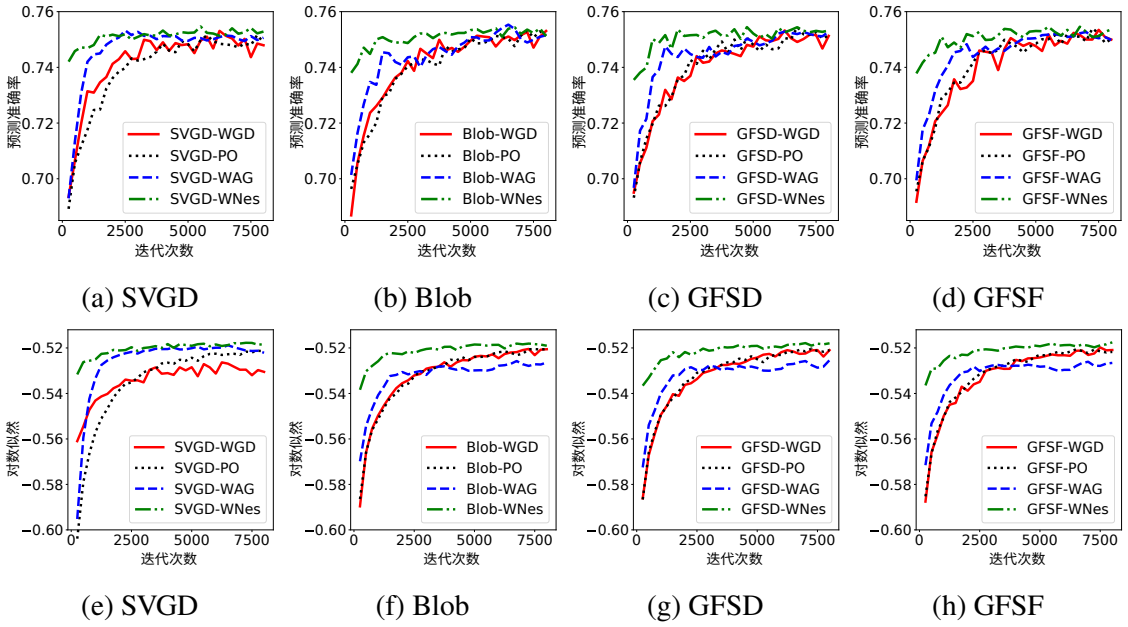
图中所示的区域为 $[-4,4]\times[-4,4]$。SVGD 方法使用固定步长 0.3（而没有采用其原版本中带有动量的 AdaGrad 方法，以保证公平比较），而其他方法都使用固定步长 0.01（这是由于 SVGD 方法中对梯度做了一个以核函数为权重的平均，因而它所计算的向量场与其他方法有不同的量级）。GFSF 方法在对 $\hat{K}$ 求逆之前加入了 $0.01I$ 这个对角阵以保证数值稳定。

由图 5.4 可发现，中位数方法会使粒子最终高度集中在目标分布的峰值位置上，因而没能充分捕捉到目标分布的特性，例如方差等。这是因为它所基于的数值稳定性考量无法保证核函数平滑操作的效果。而 HE 方法则取得了很好的近似结果。粒子不仅有合理的分散程度，而且排列更加整齐，特别是在目标分布的等值线上，粒子几乎是均匀地排列在上面。同时也可注意到，即使使用中位数方法，SVGD 方法也能产生具有一定分散度的粒子。这可能是由于它在更新每个粒子时也考虑了其他粒子处的梯度。

5.6.2　贝叶斯逻辑回归模型实验

此部分实验在贝叶斯逻辑回归模型（Bayesian logistic regression, BLR）的后验推理任务上考察所提加速框架 WAG 和 WNes（参见算法 4）的效果。这里采用与 SVGD 原工作[81] 相同的实验设定（这一设定也由之后的 Blob 方法的原工作[56] 所使用）。具体地，实验中使用 Covertype 数据集，它有 581,012 个样本，其数据维度为 54。每次运行都以 80% : 20% 的比例随机地将数据集分为训练集和测试集。模型结构与 SVGD 原工作[81] 中所使用的模型一样，并取模型权重变量的高斯先验的精度变量的伽马分布（Gamma distribution）先验的参数为 $a_0=1.0$，$b_0=100$（其中 b_0 是尺度参数（scale parameter），不是比率参数（rate parameter））。所有的 ParVI 方法都使用 100 个粒子，并由模型先验随机初始化。它们也都使用随机梯度，对应的随机子数据集大小选定为 50。

各方法在实验中所取的具体参数情况可参见表 5.1。WGD 这一列给出的是步长。PO 这一列的格式为“（衰减指数，记忆率，注入噪声的方差），步长”。这两



(a) SVGD　(b) Blob　(c) GFSD　(d) GFSF

(e) SVGD　(f) Blob　(g) GFSD　(h) GFSF

图 5.5　贝叶斯逻辑回归模型在 Covertype 数据集上的后验推理实验中 WAG 及 WNes 方法的加速效果。(a-d): 以测试准确率衡量；(e-h): 以对数似然衡量。

个方法都使用固定的步长，而 WAG 和 WNes 方法则使用衰减的步长。WAG 这一列的格式为“(加速因子 α，(步长衰减指数，步长尺度))”。WNes 这一列的格式为“(WNes 参数 λ, WNes 参数 β)，(步长衰减指数，步长尺度))”。作为一个特例的是，SVGD-WGD 方法采用了带有动量的 AdaGrad 方法以复现其原工作[81] 中的结果，并选取此方法的记忆率参数为 0.9，步长尺度为 0.03。对 GFSF 方法，在对矩阵 $\hat{K}$ 求逆之前为之加入了小对角阵 $(1 \times 10^{-5})I$。

推理效果分别以在独立的测试集上的分类准确率以及对数似然来衡量。图 5.5 展示了各方法随迭代次数（因为所有方法使用相同的随机子数据集大小，因而这个比较是公平的）收敛的情况，其中每条曲线都是平均了 10 次独立运行结果之后的情况。对于所有的四种 ParVI 方法，WAG 和 WNes 方法的收敛速度都显著地比 WGD 和 PO 方法高。另外，WNes 方法可比 WAG 得到更好的结果，特别是在推理早期，并且也对超参数更加稳定。PO 方法在此任务上的表现与 WGD 方法接近，这与 PO 方法原工作[55] 中的结果一致。此外还可以注意到，四种 ParVI 方法具有相似的表现，这也是一个自然的结果，因为 ParVI 方法都是通过平滑操作对同一个梯度流进行模拟的方法。

5.6.3　贝叶斯神经网络实验

此部分实验采用与 SVGD 方法的原工作[81] 相同的实验设定。具体地，实验中使用的贝叶斯神经网络（Bayesian neural networks）包含一个具有 50 个隐节点

表 5.2　贝叶斯神经网络在 Kin8nm 数据集上的后验推理实验中各 ParVI 方法采用的参数

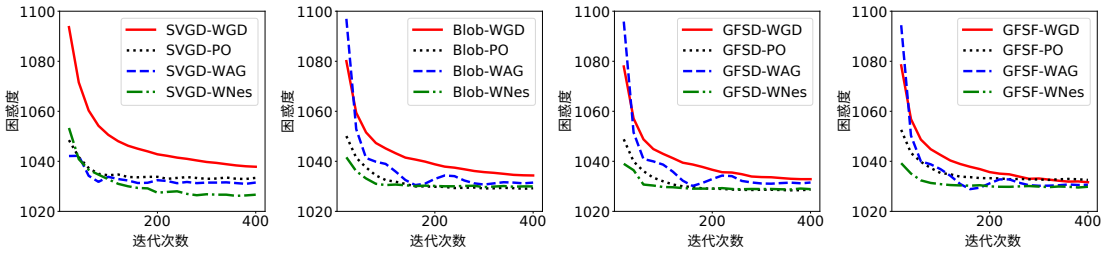
	WGD	PO	WAG	WNes
SVGD	1e-3	(1.0, 0.6, 1e-7), 1e-4	3.6, 1e-6	(1000, 0.2), 1e-4
Blob	(0.5, 3e-5)	(1.0, 0.8, 1e-7), (0.5, 3e-5)	3.5, (0.5, 1e-5)	(3000, 0.2), (0.6, 1e-4)
GFSD	(0.5, 3e-5)	(1.0, 0.8, 1e-7), (0.5, 3e-5)	3.5, (0.5, 1e-5)	(3000, 0.2), (0.6, 1e-4)
GFSF	(0.5, 3e-5)	(1.0, 0.8, 1e-7), (0.5, 3e-5)	3.5, (0.5, 1e-5)	(3000, 0.2), (0.6, 1e-4)

表 5.3　贝叶斯神经网络在 Kin8nm 数据集上的后验推理实验中各 ParVI 方法及其各加速版本的表现。

方法	平均测试均方根误差 ($\times 10^{-2}$)			
	SVGD	Blob	GFSD	GFSF
WGD	8.4±0.2	8.2±0.2	8.0±0.3	8.3±0.2
PO	7.8±0.2	8.1±0.2	8.1±0.2	8.0±0.2
WAG	7.0±0.2	**7.0±0.2**	7.1±0.1	7.0±0.1
WNes	**6.9±0.1**	7.0±0.2	**6.9±0.1**	**6.8±0.1**
方法	平均测试对数似然			
	SVGD	Blob	GFSD	GFSF
WGD	1.042±0.016	1.079±0.021	1.087±0.029	1.044±0.016
PO	1.114±0.022	1.070±0.020	1.067±0.017	1.073±0.016
WAG	1.167±0.015	**1.169±0.015**	1.167±0.017	1.190±0.014
WNes	**1.171±0.014**	1.168±0.014	**1.173±0.016**	**1.193±0.014**

(hidden node) 的隐含层 (hidden layer), 并使用 sigmoid 函数作为激活函数 (activation function)。模型权重变量的高斯先验的精度变量的伽马分布 (Gamma distribution) 先验的参数为 $a_0 = 1.0$, $b_0 = 0.1$。实验中使用 Kin8nm 数据集。它是 UCI 数据集[190] 的一个子集。每次运行都以 90% : 10% 的比例随机地将数据集分为训练集和测试集。各 ParVI 方法都使用 20 个粒子，并使用随机梯度进行估计，对应的随机子数据集大小均选作 100。它们在实验中的详细参数由表 5.2 给出，其中各列的格式与上一部分贝叶斯逻辑回归中 (5.6.2 节) 介绍表 5.1 各列的格式基本一致，例外的是 SVGD 方法为复现其原工作[81] 中的结果而采用了记忆率参数为 0.9 的带有动量的 AdaGrad 方法，而步长尺度由表格给出。另外，本实验中 WGD 方法和 PO 方法也使用衰减的步长，因此表格中也提供了对应的衰减指数。GFSF 方法在对矩阵 $\hat{K}$ 求逆之前为之加入了一个小对角阵 $0.01I$。

表 5.3 展示了各方法在 8,000 轮迭代后所得的结果 (取各方法重复 20 次独立



(a) SVGD　(b) Blob　(c) GFSD　(d) GFSF

图 5.6　隐式狄利克雷分配模型在 ICML 数据集上的后验推理实验中 WAG 及 WNes 方法的加速效果。

运行的均值和标准差)。从表中可以发现所提加速框架的 WAG 和 WNes 这两个实现可以令各 ParVI 方法在固定迭代轮数时都取得更好的结果。PO 方法也提高了这些 ParVI 方法的效果，但不如 WAG 和 WNes 方法那样明显。

5.6.4　隐式狄利克雷分配模型实验

此部分实验在隐式狄利克雷分配模型（latent Dirichlet allocation, LDA）[19] 的后验推理这个非监督学习任务中继续考察所提加速框架的效果。此任务是在给定文档数据后，估计 LDA 模型的话题（topic）这个全局隐变量的后验分布。实验中采用与 Ding 等人的工作[119] 中相同的实验设定。具体地，实验采用拓展的自然式参数化方法（expanded-natural parameterization）[18] 以及坍缩吉布斯采样（collapsed Gibbs sampling）来估计 LDA 模型的话题隐变量的后验分布对数梯度。所使用的数据集是 ICML 数据集①，它包含 765 篇文档，以及 1,918 个单词（不重复计算）。每次运行都以 80% : 20% 的比例随机地将数据集分为训练集和测试集，在训练集每篇 90% 的单词上估计 LDA 模型话题隐变量的后验分布，再在测试集上使用训练得到的后验分布的估计先在每篇 90% 的单词上训练此文档的话题配比隐变量（topic proportion），再使用剩下 10% 的单词来评估训练效果。训练效果是以困惑度（perplexity）来衡量的。困惑度越小，表明训练所得模型越契合测试数据，即可视为推理方法得到了对后验分布更好的估计。

实验中，LDA 模型将其话题隐变量的狄利克雷先验的参数选定为 0.1，而其话题配比隐变量的高斯先验的均值和标准差则分别取为 0.1 和 1.0。话题数选定为 30。坍缩吉布斯采样在每次估计梯度时运行 50 次，并使用所得的样本进行梯度的估计。对于各 ParVI 方法，它们都使用随机梯度进行计算，对应的随机子数据集大小选定为 100。它们都使用 20 个粒子。它们在实验中所选取的其他详细参数由表 5.4 给出，其中各列的格式与贝叶斯逻辑回归中（5.6.2 节）介绍表 5.1 各列的格

① https://cse.buffalo.edu/~changyou/code/SGNHT.zip

表 5.4　隐式狄利克雷分配模型在 ICML 数据集上的后验推理实验中各 ParVI 方法所采用的参数

	WGD	PO	WAG	WNes
SVGD	3.0	(0.7, 0.7, 1e-4), 10.0	2.5, 3.0	(3.0, 0.2), 10.0
Blob	0.3	(0.7, 0.7, 1e-4), 0.30	2.1, 3e-2	(0.3, 0.2), 0.30
GFSD	0.3	(0.7, 0.7, 1e-4), 0.30	2.1, 3e-2	(0.3, 0.2), 0.30
GFSF	0.3	(0.7, 0.7, 1e-4), 0.30	2.1, 3e-2	(0.3, 0.2), 0.30

式基本一致，例外的是所有方法都使用衰减的步长，其中衰减指数为 0.55，衰减初始步数为 1,000，而步长尺度则在表中给出。注意 SVGD 方法在此实验中没有采用它原本采用的带有动量的 AdaGrad 方法。GFSF 方法在对矩阵 $\hat{K}$ 求逆之前为之加入了一个小对角阵 $(1 \times 10^{-5})I$。

所提 WAG 和 WNes 方法的加速效果如图 5.6 所示，其中每条曲线是 10 次独立运行结果的平均。可以发现，对于这四种 ParVI 方法，WAG 和 WNes 都明显地提高了它们的收敛速度。此实验中 PO 方法也可以取得一个可比的加速效果。实验中 WAG 方法表现出了对其加速因子 α 这个参数的敏感性，而且在图 5.6 中也表现出了一些小波动。而 WNes 方法则稳定很多。

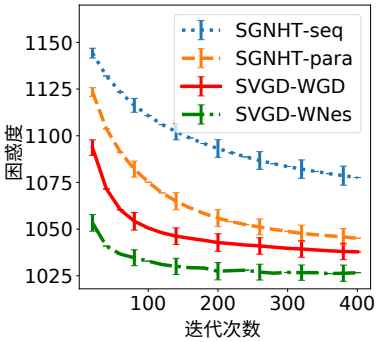


图 5.7　隐式狄利克雷分配模型在 ICML 数据集上的后验推理实验中作为 ParVI 方法和 MCMC 方法代表的 SVGD 方法与 SGNHT 方法的效果对比。

为进一步展示加速 ParVI 方法在解决实际问题中的优势，本部分实验以 SVGD-WNes 为例将它们与一个先进的 MCMC 方法进行对比，即随机梯度诺泽-胡佛恒温器方法（stochastic gradient Nosé-Hoover thermostats, SGNHT）[119]。实验中选定 SGNHT 的固定步长为 0.03，质量参数为 1.0，而扩散参数为 22.4。对于 SGNHT 方法，实验中同时实现了模拟一条链的序列式（sequential, 简记为 “-seq”）采样和模拟多条链并采集所有链最后位置作为样本的平行式（parallel，简记为 “-para”）采样。由于 ParVI 方法使用了 20 个粒子，因此对于序列式采样 SGNHT 采集此链上

最后 20 个样本，而对于平行式采样 SGNHT 选取模拟链数为 20。实验结果展示于图 5.7 中，其中每条曲线是 10 次独立运行结果的平均。可以发现，加速的 ParVI 方法在实验中具有比 MCMC 方法更快的收敛速度。另外，由于各方法都使用了相同数目的粒子（样本），因此这个结果也展示了 ParVI 方法的粒子高效性。

5.7　本章小结与讨论

通过深入探索 ParVI 方法作为沃瑟斯坦梯度流的这个解释，本章为理解 ParVI 方法建立了一个有限多个粒子近似的统一理论，并为提高 ParVI 方法的表现提出了一个加速框架和一个有原则性的带宽选择方法。所提理论发现各 ParVI 方法使用有限多个粒子的近似都是一个平滑操作，并可归于平滑密度和平滑函数这两类形式中。这两类形式的等价性给出了 ParVI 方法之间的一个联系，而平滑操作的必需性则揭示了 ParVI 方法所做的假设。这个理论也启发了两个新的 ParVI 方法的设计与开发。所提加速框架是通过对沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 黎曼结构的深入挖掘而得到的，而所提带宽选择方法是基于对平滑操作的目的的分析而推出的。实验结果展示了所提带宽选择方法所产生的明显更具代表性的粒子从而进一步加强了 ParVI 方法的粒子高效性，以及使用所提加速框架在无显著额外计算开销的情况下为各 ParVI 方法带来的更快的收敛速度。

本章所提的 ParVI 方法的平滑性假设理论可进一步启发新 ParVI 方法的开发，例如利用核函数的谱分解[191] 实现平滑函数的 ParVI 方法。所提 ParVI 方法的加速框架，包括在开发过程中所推得的沃瑟斯坦空间上的逆指数映射和平行移动等实现方法，可启发和实现对 ParVI 方法的更多改进，例如拟牛顿法[187-189] 和方差消减（variance reduction）方法[156]。其他后续工作包括对 ParVI 方法使用有限多个粒子进行近似的非渐进分析，以及对所提带宽选择方法进一步进行简化等。

第 6 章　作为沃瑟斯坦空间上的流的 MCMC 动力学系统

人们已熟知，MCMC 领域中所使用的郎之万动力学系统（Langevin dynamics, LD）是沃瑟斯坦空间（Wasserstein space）上 KL 散度的梯度流。这个观点不仅极大地帮助了对 LD 收敛性质的分析，还启发了最近发展起来的基于粒子的变分推理方法（particle-based variational inference methods, ParVI）。然而目前除去 LD 之外，却没有更多的 MCMC 动力学系统可以从沃瑟斯坦空间上的流的角度来理解。本章工作通过研究与定义一些新的概念，提出一个理论框架，将一般的 MCMC 动力学系统理解为一个纤维黎曼-泊松流形（fiber-Riemannian Poisson manifold, fRP 流形）的沃瑟斯坦空间上的纤维梯度哈密顿流（fiber-gradient Hamiltonian flow, fGH 流）。这个 fGH 流具有一个“守恒项 + 收敛项”的结构，因而可以为一般的 MCMC 动力学系统的行为给出一个直观的理解。本章将现有的 MCMC 方法在所提理论框架下进行了分析。这个理论框架也使一般的 MCMC 动力学系统能够以 ParVI 方法的方式进行模拟，这一方面为 ParVI 领域引入除 LD 之外更多更高效的 MCMC 动力学系统，而另一方面又为 MCMC 领域引入了 ParVI 方法的优点，例如粒子高效性（particle efficiency）。本章为一个特定的 MCMC 动力学系统开发了两个 ParVI 方法，并在实验中展示了这两个方法的优势。

6.1　研究动机

基于动力学系统的马尔可夫链蒙特卡罗方法（dynamics-based Markov chain Monte Carlo methods, MCMC）因其具有的渐进准确保证、可用范围广、采样效率高以及能够高效处理大规模数据的可扩展性等优势，在贝叶斯推理领域中已得到高度关注[60,97,115,192-193]。它们通过模拟一个连续时间*动力学系统*来采样。更准确地说，这个动力学系统指的是一个可保持目标分布不变的扩散过程（diffusion process）。然而，由于模拟过程中样本之间仍然会有一定的正的自相关性（auto-correlation），因而它们还是会表现出较慢的实际收敛速度和较小的有效样本数量（effective sample size）。另一类称为基于粒子的变分推理方法（particle-based variational inference methods, ParVI）则希望通过一个确定性的方式来更新样本，或称粒子，而这个确定性的更新方式则由最小化与目标分布之间的 KL 散度这一原则所确定。由于这类方法通过在模拟时为一组有限多个粒子加入了相互作用的机制，这类方法可充分利用这组粒子的近似能力，因而可以具有更好的粒子高效性（particle efficiency）。而基于优化的原则也可使它们收敛得更快。斯坦因变分梯度下降方法（Stein vari-

ational gradient descent, SVGD）[81] 是其中最有名的代表。这个领域目前在理论方面[56,83,85] 和应用方面[17,25,169-170] 都十分活跃。

MCMC 方法和 ParVI 方法这两者之间关系的研究始于它们在沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$[53-54] 上的解释，其中流形 $\mathscr{M}$ 是其支撑空间（support space）。其定义及性质可参见 5.2.1 节。这是一个非常广泛但同时也有必要结构的空间。由它的结构，其上 KL 散度的梯度流（gradient flow）便可以被定义，而人们已熟知，郎之万动力学系统 (LD)[58,194] 这个特定的 MCMC 动力学系统正是在模拟这个梯度流[61]。而近来的一些分析，包括 Chen 等人的工作[56] 以及上一章中的工作，则揭示了现有的 ParVI 方法也都是在模拟此梯度流，因而它们与 LD 模拟的是一个过程。然而，除了 LD 之外，MCMC 领域中还有很多种类的动力学系统，并且这些动力学系统可以比 LD 收敛得更快或可产生更有效的样本[97,115,119]，但目前却没有 ParVI 方法在模拟这些动力学系统。这些一般的 MCMC 动力学系统也还没有得到作为沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的一个过程这样的解释，这也阻碍了使用 ParVI 形式的模拟过程对它们的模拟。另一方面，当把 LD 看作 $\mathscr{P}_2(\mathscr{M})$ 上 KL 散度的梯度流时，其收敛行为就变得十分清晰[106,195-197]，因为这个梯度流就是使任一分布在 KL 散度的意义下以最快的方式趋向于目标分布的过程。然而，目前除了 LD 之外的 MCMC 动力学系统尚未这种视角的理解。事实上，一个一般的 MCMC 动力学系统只保证它会保持目标分布不变[120]，而并不一定保证会使任一分布都以最快的方式趋向目标分布。所以梯度流的形式很难涵盖一般的 MCMC 动力学系统。

本章提出一个理论框架，给出一般 MCMC 动力学系统在沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的解释。所提框架基于对梯度流这个概念做两方面的推广从而可以涵盖一般的 MCMC 动力学系统：**(a)** 本章引入了一个新的概念，称为*纤维黎曼流形*（fiber-Riemannian manifold）$\mathscr{M}$，它可以只在它的纤维空间（fiber，大致是它的一个个子流形，或者说是它的一个切片（slice））中定义黎曼结构（Riemannian structure），并可进一步发展出来其沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的*纤维梯度流*（fiber-gradient flow）这个新的概念；**(b)** 本章也为流形 $\mathscr{M}$ 引入一个泊松结构，从而可以定义其沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的哈密顿流。将这两方面的推广综合起来，本章可以定义一个纤维黎曼-泊松（fiber-Riemannian Poisson, fRP）流形 $\mathscr{M}$ 及其沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的纤维梯度哈密顿（fiber-gradient Hamiltonian, fGH）流。本章进而发现，任一常规的（regular）MCMC 动力学系统是一个 fRP 流形 $\mathscr{M}$ 的沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的 fGH 流，并且 MCMC 动力学系统与 fRP 流形 $\mathscr{M}$ 的结构有一个对应关系。

这个统一的理论框架为一般的 MCMC 动力学系统的行为提供了一个清晰的视角。在上面所定义的 fGH 流中，哈密顿流这个部分会保持与目标分布的 KL 散

度不变，而纤维梯度流这个部分则会在每一个纤维空间中驱使对应的条件分布朝向对应目标分布的条件分布而演化。若当前分布就是目标分布，即 KL 散度达到最小值时，纤维梯度等于零，而由哈密顿流的性质可知，此时的 fGH 流会保持目标分布不变。对于一般的分布，若此 MCMC 方法对应的纤维梯度流存在，那么它可以驱使每个纤维空间内的收敛，将每个纤维空间中的过程稳定下来，从而使得每个纤维空间中的动力学系统可以容忍使用随机梯度所带来的扰动，进而使对应的 MCMC 方法具有可扩展性。这是对已有讨论[115-116] 向一般 MCMC 动力学系统的推广，而这些已有讨论关注的是哈密顿蒙特卡罗（Hamiltonian Monte Carlo, HMC）[96-98] 这个特定的 MCMC 方法。另一方面，若此 MCMC 方法对应的哈密顿流存在，那么它可以起到让样本（粒子）探索更加广阔区域的作用[97,115]。在所提理论框架中，不同的 MCMC 方法对应着不同的纤维结构，因此对应着不同的流的部分。它们可以依此被划分为三类，其中每一类都具有特定的行为。本章对现有的 17 种 MCMC 方法按照这三类进行了统一的分析和对比。

所提理论框架同时也为 MCMC 领域和 ParVI 领域架起了桥梁。一方面，MCMC 领域中丰富的动力学系统在 ParVI 领域中被解锁，而其中很多动力学系统都具有比 LD 更加优秀的表现。另一方面，MCMC 动力学系统现在可以通过 ParVI 形式进行模拟，从而为 MCMC 领域引入 ParVI 方法的优势，例如粒子高效性。作为一个例子，本章为随机梯度哈密顿蒙特卡罗方法（stochastic gradient Hamiltonian Monte Carlo, SGHMC）[115] 的动力学系统开发了两个 ParVI 形式的模拟方法。通过考察这两个新 ParVI 方法的实际表现，本章展示了在 ParVI 领域中使用 SGHMC 动力学系统的好处，以及在 MCMC 领域中使用 ParVI 形式模拟的优势。

相关工作　Ma 等人[120] 给出了一般 MCMC 动力学系统的完备表示形式。它们的形式可以保证目标分布不变这个原则，但 MCMC 动力学系统在非稳态时（即未处在目标分布时）的表现则未能给出。最近的一些使用福克-普朗克方程（Fokker-Planck equation）来尝试对更加广泛类型的动力学系统进行分析的工作[198-199] 仍然没有脱离梯度流的形式，因此这些推广仍然不足以包含一般的 MCMC 动力学系统。

在将 MCMC 方法与 ParVI 方法建立联系方面，Chen 等人[56] 探索了 LD 和沃瑟斯坦梯度流之间的对应关系，并为这个共同的动力学系统的模拟开发了一些新的实现方法。然而，它们的考量仍然局限在 LD 这个特定的动力学系统上，而没有触及更一般的 MCMC 动力学系统。Gallego 等人[200] 将 SVGD 的算法表示成了一个特定的 MCMC 动力学系统，但并没有考虑现有的 MCMC 动力学系统在 ParVI 视角下的联系。最近，Taghvaei 等人[57] 推导出了一个加速版本的 ParVI 方法并且最终算法与本章所推导的 SGHMC 的其中一种 ParVI 形式很类似（仍然有区别）。

需要说明的是，他们的推导并没有利用 MCMC 动力学系统的概念，并且最终得到的算法只是形式上与 SGHMC 类似，而本章所推导出来的 SGHMC 的两种 ParVI 形式是由所提理论框架所保证的，并且通过这个框架可以清楚地看出所得方法与 SGHMC 的直接联系。另外，通过所提理论框架，这个做法可以用于更多的 MCMC 动力学系统上。

6.2　背景知识

为给 MCMC 动力学系统建立沃瑟斯坦空间上的流的观念，这里首先关注两种特殊的流，即梯度流与哈密顿流。其中流形上的梯度流概念已在 2.1.1.3 节和 2.1.2.2 节介绍，因此不加赘述。沃瑟斯坦空间及其上的梯度流已分别在上一章 5.2.1 节和 5.2.2 节介绍，但本章中需要考虑更多概念，因此将对沃瑟斯坦空间及其上梯度流进行简单回顾，并介绍更多所需概念。哈密顿流则是一个新的概念。本节将介绍一般情况下的哈密顿流以及沃瑟斯坦空间上的哈密顿流。

参考 2.2 节，可知一般 MCMC 动力学系统可在欧氏空间 $\mathbb{R}^m$ 上表达，因此本章只考虑与 $\mathbb{R}^m$ 全局微分同胚（globally diffeomorphic）的流形 $\mathscr{M}$ 就足够了，而这个欧氏空间 $\mathbb{R}^m$ 就是流形 $\mathscr{M}$ 的一个全局坐标系。本章假设所考虑的分布都是绝对连续（absolutely continuous）的（关于欧氏空间上的勒贝格测度或黎曼流形上的黎曼体积形式），从而可以使用其密度函数（关于各自空间上的默认测度）来表示。

6.2.1　沃瑟斯坦空间及其上的梯度流

对于黎曼流形 $(\mathscr{M}, g)$ 的沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$，它也有一个黎曼结构[52-54]。沃瑟斯坦空间在分布 $q \in \mathscr{P}_2(\mathscr{M})$ 处的切空间可表示为（可参见 Villani 的著作[53] 定理 13.8 或 Ambrosio 等人的著作[54] 定理 8.3.1 及定义 8.4.1）：

$$T_q\mathscr{P}_2(\mathscr{M}) = \overline{\{\operatorname{grad} f \mid f \in C_c^\infty(\mathscr{M})\}}^{\mathscr{L}_q^2(\mathscr{M})},$$

其中 $C_c^\infty(\mathscr{M})$ 表示流形 $\mathscr{M}$ 上具有紧致支撑集的光滑函数的集合，$\mathscr{L}_q^2(\mathscr{M})$ 是如下希尔伯特空间 $\{V \in \mathscr{T}(\mathscr{M}) \mid \mathbb{E}_q[g(V,V)] < \infty\}$ 带有内积 $\langle V, U\rangle_{\mathscr{L}_q^2} := \mathbb{E}_{q(x)}[g_x(V(x), U(x))]$，而上划线表示取闭包（closure）操作。切空间 $T_q\mathscr{P}_2$ 可从 $\mathscr{L}_q^2(\mathscr{M})$ 继承一个内积，这个内积便定义了 $\mathscr{P}_2(\mathscr{M})$ 的一个黎曼结构，且这个黎曼结构是与沃瑟斯坦距离相容的，即它所引出的黎曼距离刚好就是沃瑟斯坦距离[175]。有了这个结构，沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的 KL 散度 $\mathrm{KL}_p(q) := \int_{\mathscr{M}} \log(q/p)\,\mathrm{d}q, \forall q \in \mathscr{P}_2(\mathscr{M})$ 的梯度便可以显式给出（可参见 Villani 的著作[53] 公式 15.2 以及定

理 23.18）：

$$\operatorname{grad}\mathrm{KL}_p(q) = \operatorname{grad}\log(q/p) \quad \in T_q\mathscr{P}_2(\mathscr{M}). \tag{6-1}$$

注意到 $T_q\mathscr{P}_2(\mathscr{M})$ 是希尔伯特空间 $\mathscr{L}^2_q(\mathscr{M})$ 的线性子空间，因此可以唯一地定义一个正交投影（orthogonal projection）$\pi_q : \mathscr{L}^2_q(\mathscr{M}) \to T_q\mathscr{P}_2$。对于任意 $V \in \mathscr{L}^2_q(\mathscr{M})$，投影后的向量场 $\pi_q(V)$ 都是在切空间 $T_q\mathscr{P}_2$ 中唯一满足 $\operatorname{div}(qV) = \operatorname{div}(q\pi_q(V))$ 成立的向量场（可参见 Ambrosio 等人的著作[54] 引理 8.4.2），其中 div 表示流形 $\mathscr{M}$ 上向量场的散度。若记 q 为对应分布在坐标空间 $\mathbb{R}^m$ 上关于勒贝格测度的密度函数，则向量场 V 的散度可在坐标空间中表示为：$\operatorname{div}(qV) = \partial_i(qV^i)$。这个投影也具有一个物理上的直观解释。令 $V \in \mathscr{L}^2_q(\mathscr{M})$ 是流形 $\mathscr{M}$ 上的一个向量场，并令它所对应的流作用在分布 q 的随机变量 x 上。变换之后的随机变量 $F_t(x)$ 对应着一个分布 q_t，因而由向量场 V 可以引出一条分布曲线 $(q_t)_t$。而这样的曲线 $(q_t)_t$ 在分布 q 处的切向量正好就是 $\pi_q(V)$。

6.2.2　一般流形及沃瑟斯坦空间上的哈密顿流

哈密顿流（Hamiltonian flow）是对经典力学（classical mechanics）中哈密顿动力学系统（Hamiltonian dynamics）的抽象和推广[142]。它是由流形 $\mathscr{M}$ 上的一个泊松结构（Poisson structure）（可参见 [201]）所定义的。泊松结构可通过泊松括号（Poisson bracket）$\{\cdot,\cdot\} : C^\infty(\mathscr{M}) \times C^\infty(\mathscr{M}) \to C^\infty(\mathscr{M})$ 来表示，即一个满足莱布尼兹法则（Leibniz rule）的 $C^\infty(\mathscr{M})$ 上的李括号（Lie bracket），也可以等价地通过一个二重向量场（bivector field）$\chi : T^*\mathscr{M} \times T^*\mathscr{M} \to C^\infty(\mathscr{M})$ 来表示，两者之间的对应关系为 $\chi(\mathrm{d}f, \mathrm{d}h) = \{f, h\}, \forall f, h \in C^\infty(\mathscr{M})$。将二重向量场在坐标空间中表达，有 $\chi_x(\mathrm{d}f(x), \mathrm{d}h(x)) = \chi^{ij}(x)\partial_i f(x)\partial_j h(x)$，其中矩阵 $(\chi^{ij}(x))$ 要求是反对称的，并且满足雅可比恒等式（Jacobi identity）：

$$\chi^{il}\partial_l\chi^{jk} + \chi^{jl}\partial_l\chi^{ki} + \chi^{kl}\partial_l\chi^{ij} = 0, \forall i, j, k. \tag{6-2}$$

由流形 $\mathscr{M}$ 上的一个泊松结构可以引出其上一光滑函数 f 的哈密顿向量场（Hamiltonian vector field）：$V_f[\cdot] := \{\cdot, f\}$（此处将切向量看作在函数上的算符）。其坐标表达式为：

$$V_f(x) = \chi^{ij}(x)\partial_j f(x)\partial_i \quad \in T_x\mathscr{M}. \tag{6-3}$$

哈密顿向量场 V_f 所引出的流即为哈密顿流 $\{(F_t(x))_t\}$。光滑函数 f 也被称作哈密顿流 V_f 的哈密顿量（Hamiltonian）。哈密顿流最重要的性质是它可保持哈密顿量

f 守恒：$f(F_t(x))$ 关于 t 是一个常量。哈密顿流的概念可能在辛流形（symplectic manifold）的设定下，或者更加具体地，在流形的切丛的设定下更为人所知（可参见 Da Silva 的著作[202] 或 Marsden 等人的著作[142]）。但是这些概念都没有此处所考虑的泊松流形那么广泛（例如，辛流形和切丛都一定是偶数维的），因而不能满足本章解释一般 MCMC 动力学系统这个任务的要求。

对于沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$，流形 $\mathscr{M}$ 的泊松结构 $\{\cdot,\cdot\}_{\mathscr{M}}$ 也可以为它引出一个泊松结构。考虑沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的线性函数 $\mathcal{F}_f : q \mapsto \mathbb{E}_q[f]$ 其中 $f \in C_c^\infty(\mathscr{M})$。可为这些线性函数定义一个泊松括号为（可参见 Lott 的著作[183] 第 6 节，或者 Gangbo 等人的著作[203] 7.2 节）：

$$\{\mathcal{F}_f, \mathcal{F}_h\}_{\mathscr{P}_2(\mathscr{M})} := \mathcal{F}_{\{f,h\}_{\mathscr{M}}}. \tag{6-4}$$

这个泊松括号可以通过线性化的概念推广至沃瑟斯坦空间上的任一光滑函数 $\mathcal{F}$。函数 $\mathcal{F}$ 在分布 q 处的线性化定义为一个在 q 处与 $\mathcal{F}$ 具有相同梯度的线性函数 $\mathcal{F}_f$：$\operatorname{grad}\mathcal{F}_f(q) = \operatorname{grad}\mathcal{F}(q)$。进而泊松括号可以拓展为 $\{\mathcal{F},\mathcal{H}\}_{\mathscr{P}_2}(q) := \{\mathcal{F}_f,\mathcal{F}_h\}_{\mathscr{P}_2}(q)$（Gangbo 等人的著作[203] 注释 7.8），其中 $\mathcal{F}_f$ 和 $\mathcal{F}_h$ 分别是光滑函数 $\mathcal{F}$ 和 $\mathcal{H}$ 在 q 处的线性化。由此泊松结构可在沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上定义函数 $\mathcal{F}$ 的哈密顿向量场（Gangbo 等人的著作[203] 7.2 节）为：

$$\mathcal{V}_F(q) = \mathcal{V}_{\mathcal{F}_f}(q) = \pi_q(V_f) \quad \in T_q\mathscr{P}_2(\mathscr{M}). \tag{6-5}$$

关于沃瑟斯坦空间上的泊松结构这个方向上的研究，Ambrosio 等人[204] 研究了 $\mathscr{P}_2(\mathscr{M})$ 上哈密顿流的存在性和模拟方法，其中 $\mathscr{M}$ 是一个辛欧氏空间，并在一些特定条件下验证了哈密顿量守恒的性质。Gangbo 等人[203] 研究了函数空间 $C_c^\infty(\mathscr{M})$ 的代数对偶空间（algebraic dual）$(C_c^\infty(\mathscr{M}))^*$ 上的泊松结构。这个空间是能够包含沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 的。他们发现，由函数空间 $C_c^\infty(\mathscr{M})$ 的李代数结构（Lie structure）所引出的 $(C_c^\infty(\mathscr{M}))^*$ 上的泊松结构与式 (6-4) 是一致的。他们所考虑的情况仍然是一个辛欧氏空间 $\mathscr{M}$，但他们的分析和推导过程以及结论都可以直接用于一个黎曼-泊松流形上。Lott[183] 考虑泊松流形 $\mathscr{M}$ 上所有绝对连续的分布所构成的空间上的泊松结构，并采取与式 (6-4) 相同的形式。他发现，这个泊松结构正是 Gangbo 等人[203] 所考虑的 $(C_c^\infty(\mathscr{M}))^*$ 上的泊松结构在上述分布流形上的限制。

6.3　MCMC 动力学系统作为沃瑟斯坦空间上的流的解释

此部分将展示本章工作的主要发现，即一般 MCMC 动力学系统与沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的特定流的联系。下面将首先深入研究这两个概念并定义一些新的

更广泛的概念作为建立这个联系的技术准备，然后提出统一的理论框架来正式地描述这个联系，并对现有 MCMC 方法在此框架下进行分析。

6.3.1　技术发展和概念定义

本节首先来挖掘 MCMC 动力学系统及沃瑟斯坦空间上的流的相关知识，并引出一些新的概念。

在 MCMC 动力学系统方面　一般 MCMC 动力学系统可由 2.2 节中所介绍的完备表示形式（式 (2-2)）描述为欧氏空间 $\mathbb{R}^m$ 中的一个扩散过程。注意到沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的流都是确定性的过程，而 MCMC 方法则会涉及随机扩散的过程。因此这里将首先将 MCMC 动力学系统重新表示为一个等价的确定性动力学系统，以方便统一框架的建立。此处所说两个动力学系统是等价的（equivalent），如果它们产生相同的分布曲线。

引理 6.1 (等价的确定性 MCMC 动力学系统)：　假设式 (2-2) 所表达的 MCMC 动力学系统具有对称的扩散矩阵 D。则这个 MCMC 动力学系统等价于下面的 $\mathbb{R}^m$ 中的确定性动力学系统：

$$\begin{aligned} \mathrm{d}x &= W_t(x)\,\mathrm{d}t, \\ (W_t)^i &= D^{ij}\partial_j \log(p/q_t) + Q^{ij}\partial_j \log p + \partial_j Q^{ij}, \end{aligned} \tag{6-6}$$

其中 q_t 是 t 时刻 x 的分布的密度函数。

证明 给定动力学系统 (2-2)，它所产生的分布曲线 $(q_t)_t$ 可由福克-普朗克方程（Fokker-Planck equation）（可参见 [177]）给出：

$$\partial_t q_t = -\partial_i(q_t H^i) + \partial_i\partial_j(q_t D^{ij}).$$

此式可进一步化简：

$$\begin{aligned} \partial_t q_t =& -(\partial_i q_t)H^i - q_t(\partial_i H^i) + q_t(\partial_i\partial_j D^{ij}) + (\partial_i\partial_j q_t)D^{ij} + (\partial_i q_t)(\partial_j D^{ij}) + (\partial_j q_t)(\partial_i D^{ij}) \\ =& -(\partial_i q_t)(\partial_j D^{ij} + \partial_j Q^{ij}) - (\partial_i q_t)(D^{ij} + Q^{ij})\frac{\partial_j p}{p} - q_t\partial_i\partial_j(D^{ij} + Q^{ij}) \\ & - q_t(\partial_i D^{ij} + \partial_i Q^{ij})\frac{\partial_j p}{p} - q_t(D^{ij} + Q^{ij})(\frac{\partial_i\partial_j p}{p} - \frac{(\partial_i p)(\partial_j p)}{p^2}) \\ & + q_t(\partial_i\partial_j D^{ij}) + (\partial_i\partial_j q_t)D^{ij} + (\partial_i q_t)(\partial_j D^{ij}) + (\partial_j q_t)(\partial_i D^{ij}) \\ =& \quad (\partial_i q_t - \frac{q_t}{p}\partial_i p)(\partial_j D^{ij} - \partial_j Q^{ij}) - \frac{1}{p}(\partial_i q_t)(\partial_j p)(D^{ij} + Q^{ij}) \end{aligned}$$

$$-\frac{q_t}{p}(\partial_i\partial_j p)D^{ij}+\frac{q_t}{p^2}(\partial_i p)(\partial_j p)D^{ij}+(\partial_i\partial_j q_t)D^{ij},$$

其中最后一个等式利用了扩散矩阵 D 的对称性以及卷曲矩阵 Q 的反对称性：$(\partial_j p)(\partial_i D^{ij})=(\partial_i p)(\partial_j D^{ji})=(\partial_i p)(\partial_j D^{ij})$，类似地，$(\partial_j p)(\partial_i Q^{ij})=-(\partial_i p)(\partial_j Q^{ij})$；$\partial_i\partial_j Q^{ij}=\partial_j\partial_i Q^{ji}=-\partial_i\partial_j Q^{ij}$ 所以 $\partial_i\partial_j Q^{ij}=0$，类似地，$(\partial_i p)(\partial_j p)Q^{ij}=0$，$(\partial_i\partial_j p)Q^{ij}=0$。

引理中所给出的确定性动力学系统 $\mathrm{d}x=W_t(x)\,\mathrm{d}t$（其中 $W_t(x)$ 由式 (6-6) 定义）所给出的分布曲线为：

$$\begin{aligned}
\partial_t q_t =& -\partial_i(q_t(W_t)^i)=-(\partial_i q_t)(W_t)^i-q_t(\partial_i(W_t)^i)\\
=& -(\partial_i q_t)D^{ij}(\frac{\partial_j p}{p}-\frac{\partial_j q_t}{q_t})-(\partial_i q_t)Q^{ij}(\frac{\partial_j p}{p})-(\partial_i q_t)(\partial_j Q^{ij})\\
& -q_t(\partial_i D^{ij})(\frac{\partial_j p}{p}-\frac{\partial_j q_t}{q_t})-q_t D^{ij}\left(\frac{\partial_i\partial_j p}{p}-\frac{(\partial_j p)(\partial_i p)}{p^2}-\frac{\partial_i\partial_j q_t}{q_t}+\frac{(\partial_j q_t)(\partial_i q_t)}{q_t^2}\right)\\
& -q_t(\partial_i Q^{ij})\frac{\partial_j p}{p}-q_t Q^{ij}(\frac{\partial_i\partial_j p}{p}-\frac{(\partial_j p)(\partial_i p)}{p^2})-q_t(\partial_i\partial_j Q^{ij})\\
=& \quad(\partial_i q_t-\frac{q_t}{p}\partial_i p)(\partial_j D^{ij}-\partial_j Q^{ij})-\frac{1}{p}(\partial_i q_t)(\partial_j p)(D^{ij}+Q^{ij})\\
& -\frac{q_t}{p}(\partial_i\partial_j p)D^{ij}+\frac{q_t}{p^2}(\partial_i p)(\partial_j p)D^{ij}+(\partial_i\partial_j q_t)D^{ij},
\end{aligned}$$

其中最后一个等式也使用了上面提到的性质。由此可见，这两个动力学系统会产生同样的分布曲线，因此它们是等价的。 □

对于任意分布 $q\in\mathscr{P}_2(\mathbb{R}^m)$，投影后的向量场 $\pi_q(W)$ 是 q 处的一个切向量，因此 W 可定义一个 $\mathscr{P}_2(\mathbb{R}^m)$ 上的向量场，而它可以引出沃瑟斯坦空间上的一个流。这种方式给出了 MCMC 动力学系统作为沃瑟斯坦空间上的流的第一个视角。但这个形式仍然无法为此动力学系统的行为给出一个明晰的解释。为此，本章将会在定理 6.1 中给出另外一个等价的沃瑟斯坦空间上的流，它的鲜明结构可以给 MCMC 动力学系统一个直观的理解。

此外，引理 6.1 在理解 MCMC 动力学系统的巴伯生成器（Barbour's generator）[205] $\mathcal{B}$ 方面也有其独立的价值。这个生成器的重要性在于，由于它具有性质 $\mathbb{E}_p[\mathcal{B}f]=0$，因此它可被用于通过斯坦因方法（Stein's method）[206] 构造分布度量中，进而可以开发 ParVI 方法。例如，标准的郎之万动力学系统所引出的巴伯生成器刚好就是斯坦因算符（Stein's operator），而这个算符进而可以产生一个称为斯坦因差异量（Stein discrepancy）[207] 的分布度量，而这个度量最终启发了 SVGD 这个 ParVI 方法的产生。具体地，对于一个 MCMC 动力学系统，巴伯生成器将一个函数 $f\in C_c^\infty(\mathbb{R}^m)$ 映射为另一个函数：$(\mathcal{B}f)(x):=\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_{q_t}[f]\big|_{t=0}$，其

中 $(q_t)_t$ 是在此 MCMC 动力学系统下演化的分布曲线，并服从初始条件 $q_0 = \delta_x$（狄拉克度量）。以沃瑟斯坦空间 $\mathscr{P}_2(\mathbb{R}^m)$ 上的线性函数 $\mathcal{F}_f$ 来表示，可以发现 $(\mathcal{B}f)(x) = \dfrac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_f(q_t)\big|_{t=0} = \langle \operatorname{grad}\mathcal{F}_f, \pi_{q_0}(W_0)\rangle_{T_{q_0}\mathscr{P}_2}$ 正是 $\mathcal{F}_f$ 在 q_0 处沿着曲线 $(q_t)_t$ 的方向导数（directional derivative）。参见如下推导，这个发现可得到如下表达式：

$$\mathcal{B}f = \frac{1}{p}\partial_j\left[p\left(D^{ij} + Q^{ij}\right)(\partial_i f)\right]. \tag{6-7}$$

这与已有结果（例如 Gorham 等人的著作[208] 定理 2）是吻合的。

推导　巴伯生成器可被看作方向导数 $(\mathcal{B}f)(x) = \dfrac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_f(q_t)\Big|_{\substack{q_0=\delta_x\\ t=0}}$ on $\mathscr{P}_2(\mathbb{R}^m)$。由梯度的定义，这个表达式可以写为：$(\mathcal{B}f)(x) = \langle \operatorname{grad}\mathcal{F}_f, \pi_{q_0}(W_0)\rangle_{T_{q_0}\mathscr{P}_2} = \langle \operatorname{grad}\mathcal{F}_f, W_0\rangle_{\mathscr{L}^2_{q_0}}$，其中 $\pi_{q_0}(W_0)$ 是分布曲线 $(q_t)_t$ 在时刻 0 处的切向量（由引理 6.1），而最后一个等式成立是由于 π_q 是从 $\mathscr{L}^2_q$ 到 $T_q\mathscr{P}_2$ 的正交投影，以及 $\operatorname{grad}\mathcal{F}_f \in T_{q_0}\mathscr{P}_2$（参见 6.2.1 节）。

在继续推导之前，首先介绍一个分布的弱导数（weak derivative）的概念（可参见 Nicolaescu 的著作[132] 定义 10.2.1）。对于一个有光滑密度函数 q 的绝对连续分布，对于任意 $f \in C_c^\infty(\mathbb{R}^m)$，由分部积分（integration by parts）的规则可以写出：

$$\int_{\mathbb{R}^m} f(x)(\partial_i q(x))\,\mathrm{d}x = \int_{\mathbb{R}^m} \partial_i(f(x)q(x))\,\mathrm{d}x - \int_{\mathbb{R}^m}(\partial_i f(x))q(x)\,\mathrm{d}x.$$

而由高斯定理（Gauss's theorem）（可参见 Abraham 等人的著作[131] 定理 8.2.9），可得结论 $\displaystyle\int_{\mathbb{R}^m}\partial_i(f(x)q(x))\,\mathrm{d}x = \lim_{R\to+\infty}\int_{\mathbb{S}^{m-1}(R)}(f(y)q(y))v_i(y)\,\mathrm{d}y$，其中 $\mathbb{S}^{m-1}(R)$ 是 $\mathbb{R}^m$ 中半径为 R 的 $m-1$ 维超球面，$y \in \mathbb{S}^{m-1}(R)$，而 v_i 则是超球面 $\mathbb{S}^{m-1}(R)$ 上的单位法向量 v（指向球外部）的第 i 个分量。由于 f 具有紧致的支撑集，且 $\displaystyle\lim_{\|x\|\to+\infty} q(x) = 0$，因此选定一个足够大的半径 R 之后，可以得到 $f(y)q(y) = 0$。这使得上面的积分等于零，因此有：

$$\int_{\mathbb{R}^m} f(x)(\partial_i q(x))\,\mathrm{d}x = -\int_{\mathbb{R}^m}(\partial_i f(x))q(x)\,\mathrm{d}x, \forall f \in C_c^\infty(\mathbb{R}^m).$$

这一性质可以直接用来定义一种微分 $\partial_i q$，从而对于非绝对连续的分布也可以定义其微分，例如对狄拉克测度 δ_{x_0}：

$$\int_{\mathbb{R}^m} f(x)(\partial_i \delta_{x_0}(x))\,\mathrm{d}x := \int_{\mathbb{R}^m}(\partial_i f(x))\delta_{x_0}(x)\,\mathrm{d}x = \partial_i f(x_0).$$

这种导数的定义方式即为弱导数。

现在继续进行推导。利用式 (6-6) 中的形式并注意到 $q_0 = \delta_{x_0}$，可有结论：

$$
\begin{aligned}
&(\mathcal{B}f)(x_0) = \langle \operatorname{grad} \mathcal{F}_f, W_0 \rangle_{\mathscr{L}^2_{q_0}} \\
&= \mathbb{E}_{q_0(x)}[\langle \operatorname{grad} f(x), W_0(x) \rangle_{\mathbb{R}^m}] = \mathbb{E}_{q_0}[(\partial_i f) W_0^i] \\
&= \mathbb{E}_{q_0}\left[D^{ij}(\partial_i f)\big(\partial_j \log(p/q_0)\big) + Q^{ij}(\partial_i f)(\partial_j \log p) + (\partial_j Q^{ij})(\partial_i f)\right] \\
&= \quad \left[D^{ij}(\partial_i f)(\partial_j \log p)\right](x_0) - \int_{\mathbb{R}^m} \big(D^{ij}(\partial_i f)\big)(x)(\partial_j q_0)(x)\,\mathrm{d}x \\
&\quad + \left[Q^{ij}(\partial_i f)(\partial_j \log p) + (\partial_j Q^{ij})(\partial_i f)\right](x_0) \\
&= \left[D^{ij}(\partial_i f)(\partial_j \log p) + \frac{1}{p}\partial_j(pQ^{ij})(\partial_i f)\right](x_0) + \int_{\mathbb{R}^m} \partial_j\big(D^{ij}(\partial_i f)\big)(x) q_0(x)\,\mathrm{d}x \\
&= \left[D^{ij}(\partial_i f)(\partial_j \log p) + \frac{1}{p}\partial_j(pQ^{ij})(\partial_i f)\right](x_0) + \left[\partial_j\big(D^{ij}(\partial_i f)\big)\right](x_0) \\
&= \left[D^{ij}(\partial_i f)(\partial_j \log p) + \frac{1}{p}\partial_j(pQ^{ij})(\partial_i f) + (\partial_j D^{ij})(\partial_i f) + D^{ij}(\partial_i \partial_j f)\right](x_0) \\
&= \left[\frac{1}{p}\partial_j\big(p(D^{ij} + Q^{ij})\big)(\partial_i f) + D^{ij}(\partial_i \partial_j f)\right](x_0) \\
&= \left[\frac{1}{p}\partial_j\big(p(D^{ij} + Q^{ij})\big)(\partial_i f) + (D^{ij} + Q^{ij})(\partial_i \partial_j f)\right](x_0) \\
&= \left[\frac{1}{p}\partial_j\left[p\left(D^{ij} + Q^{ij}\right)(\partial_i f)\right]\right](x_0),
\end{aligned}
$$

其中倒数第二个等式成立是由于 $Q^{ij}(\partial_i \partial_j f) = 0$（因为 Q 是反对称的）。这样便可得到式 (6-7)。

（推导毕）

在沃瑟斯坦空间上的流的方面　在从沃瑟斯坦空间上找到一个具有直观结构的流来解释一般 MCMC 动力学系统之前，本节先深入挖掘一下沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的流，其中 $\mathscr{M}$ 同时具有一个黎曼结构和一个泊松结构[①]。KL 散度 KL_p 的梯度已经由式 (6-1) 给出，但由于其非线性性，它的哈密顿向量场却不易直接得到。本节首先为它的哈密顿向量场推导出一个显式的表达。

引理 6.2 (KL 散度在沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的哈密顿向量场)：令 χ 是流形 $\mathscr{M}$ 的泊松结构的二重向量场形式，并考虑由此泊松结构所引出的其沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的泊松结构（参见 6.2.2 节）。则 KL 散度在沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上

① 本章不考虑黎曼结构与泊松结构之间的相容性（compatibility），因而本章所考虑的流形不同于 Kähler 流形，且比 Kähler 流形更加广泛。

的 KL_p 的哈密顿向量场为：

$$\mathcal{V}_{\mathrm{KL}_p}(q) = \pi_q(V_{\log(q/p)}) = \pi_q(\chi^{ij}\partial_j \log(q/p)\partial_i).$$

证明 注意到 KL 散度 $\mathrm{KL}_p(q) = \int_{\mathscr{M}} \log(q/p)\,\mathrm{d}q$ 是沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的非线性函数，因此首先需要对它进行线性化。考虑沃瑟斯坦空间上一定点 $q_0 \in \mathscr{P}_2(\mathscr{M})$。式 (6-1) 给出了 KL 散度在 q_0 处的梯度：$\operatorname{grad}\mathrm{KL}_p(q_0) = \operatorname{grad}\log(q_0/p)$。考虑沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的如下线性函数：

$$F : q \mapsto \int_{\mathscr{M}} \log(q_0/p)\,\mathrm{d}q. \tag{6-8}$$

由已有知识（例如 Villani 的著作[53] 例 15.10，Ambrosio 等人的著作[54] 引理 10.4.1，或 Santambrogio 的著作[209] 式 4.10）可知，它在 q_0 处的梯度为：

$$\big(\operatorname{grad}\mathcal{F}\big)(q_0) = \operatorname{grad}\left(\frac{\delta\mathcal{F}}{\delta q}\bigg|_{q=q_0}\right),$$

其中 $\dfrac{\delta\mathcal{F}}{\delta q}$ 表示 $\mathcal{F}$ 的一阶函数变分（first-order functional variation）。对于式 (6-8) 中所给出的 $\mathcal{F}$，这个变分在 $q = q_0$ 处等于 $\log(q_0/p)$。现在可以知道 $\operatorname{grad}\mathcal{F}(q_0) = \operatorname{grad}\log(q_0/p) = \operatorname{grad}\mathrm{KL}_p(q_0)$，因此 $\mathcal{F}(q)$ 是 $\mathrm{KL}_p(q)$ 在 $q = q_0$ 处的线性化，且式 (6-5) 中对应的 $f \in C_c^\infty(\mathscr{M})$ 为 $\log(q_0/p)$，因此

$$\mathcal{V}_{\mathrm{KL}_p}(q_0) = \pi_{q_0}(V_{\log(q_0/p)}).$$

参考式 (6-3)，可有结论 $V_{\log(q_0/p)} = \chi^{ij}\partial_j \log(q_0/p)\partial_i$。最后，由 q_0 的一般性，便可写出引理中的结果。 □

注意其中的投影算符 π_q 只是为了给出一个沃瑟斯坦空间上合法的向量场，它并不会对动力学系统带来差别，因为 V 和 $\pi_q(V)$ 会在 q 处产生相同的分布曲线，即它们是局部等价的。

接下来，为了能使所提理论框架可以涵盖一般的 MCMC 动力学系统，本章引入一个新的概念，称作纤维黎曼流形（fiber-Riemannian manifold），并在其上定义相关的概念。这个概念是黎曼流形的一个推广，它解除了黎曼流形对其黎曼结构非退化（non-degenerate）的要求。

定义 6.1 (纤维黎曼流形)：称一个流形 $\mathscr{M}$ 是一个纤维黎曼流形（fiber-Riemannian manifold），如果它是一个纤维丛（fiber bundle）且其每一个纤维空间（fiber）上都定义有一个黎曼结构。

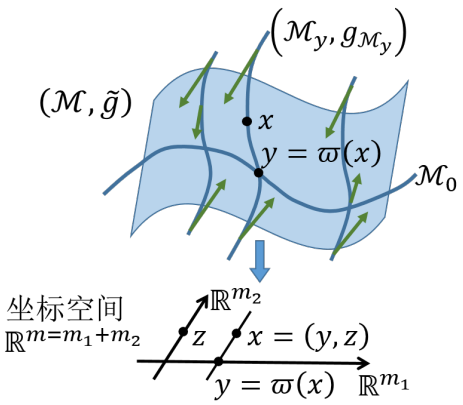


图 6.1　纤维黎曼流形 $(\mathscr{M}, \tilde{g})$ $(m_1 = m_2 = 1)$ 及其上的纤维梯度（绿色箭头）的图示

相关概念的图示由图 6.1 给出。大致来讲，m 维流形 $\mathscr{M}$（其中 $m = m_1 + m_2$）是一个纤维丛意味着存在一个 m_1 维流形 $\mathscr{M}_0$（基空间，base space）和一个 m_2 维流形 $\mathscr{F}$（公共纤维空间，common fiber）使得 $\mathscr{M}$ 局部等价于乘积流形 $\mathscr{M}_0 \times \mathscr{F}$（可参见 Nicolaescu 的著作[132] 定义 2.1.21）。记向基空间的投影 $\mathscr{M} \to \mathscr{M}_0$ 为 ϖ。此投影是满射，因而可以定义通过点 $x \in \mathscr{M}$ 的纤维空间为 $\mathscr{M}$ 的一个子流形 $\mathscr{M}_{\varpi(x)} := \varpi^{-1}(\varpi(x))$。由定义，对于任意 x，所对应的纤维空间 $\mathscr{M}_{\varpi(x)}$ 都微分同胚于 $\mathscr{F}$。流形 $\mathscr{M}$ 的坐标系也可由此结构分解：$x = (y, z)$ 其中 $y \in \mathbb{R}^{m_1}$ 是基空间 $\mathscr{M}_0$ 的坐标系，而 $z \in \mathbb{R}^{m_2}$ 是纤维空间 $\mathscr{M}_{\varpi(x)} = \mathscr{M}_y$ 的坐标系。在同一纤维空间 $\mathscr{M}_y$ 中的点的坐标具有相同的 y 的部分。另外，本章也允许 m_1 或 m_2 其中之一等于零的情况发生。

一个纤维黎曼流形 $\mathscr{M}$ 会为其每个纤维空间 $\mathscr{M}_y$ 定义一个黎曼结构 $g_{\mathscr{M}_y}$。在纤维空间的坐标系中，这个黎曼结构可表示为矩阵 $\big((g_{\mathscr{M}_y})_{ij}(z)\big)_{m_2 \times m_2}$。考虑 $\mathscr{M}$ 上的光滑函数 f。对于任一纤维空间 $\mathscr{M}_y$，它可看作其上的函数，并可写作 $f(x) = f(y, z)$，其中 y 是定值，而 z 是 $\mathscr{M}_y$ 上的点（或坐标）。利用 $\mathscr{M}_y$ 的黎曼结构可在其上定义 $f(y, z)$ 关于 z 的梯度，并可在其坐标系中表示为 $(g_{\mathscr{M}_y})^{ij}(z)\partial_{z^j} f(y, z)$。考虑 $\mathscr{M}$ 上所有的纤维空间，将函数 f 在各纤维空间上的梯度取并集，可以得到整个流形 $\mathscr{M}$ 上的一个向量场，坐标表示为 $\big(0_m, (g_{\mathscr{M}_{\varpi(x)}})^{ij}(z)\partial_{z^j} f(\varpi(x), z)\big)$。此向量场被称为光滑函数 f 在纤维黎曼流形 $\mathscr{M}$ 上的纤维梯度（fiber-gradient）$\mathrm{grad}_{\mathrm{fib}}\, f$。为将纤维梯度表示为与梯度类似的形式，定义纤维黎曼结构（fiber-Riemannian structure）$\tilde{g}$ 为：

$$\big(\tilde{g}^{ij}(x)\big)_{m \times m} = \begin{pmatrix} 0_{m_1 \times m_1} & 0_{m_1 \times m_2} \\ 0_{m_2 \times m_1} & \big((g_{\mathscr{M}_{\varpi(x)}})^{ij}(z)\big)_{m_2 \times m_2} \end{pmatrix}. \tag{6-9}$$

利用此定义，纤维梯度的坐标表示可以写为：

$$\operatorname{grad}_{\text{fib}} f(x) = \tilde{g}^{ij}(x)\partial_j f(x)\partial_i.$$

注意其中 $(\tilde{g}^{ij})$ 通常（$m_1 \geqslant 1$）是一个奇异矩阵，而黎曼结构 g 的坐标表示 (g_{ij}) 或 (g^{ij}) 一定是非奇异的（参见 2.1.2.1 节），因而无法通过它来定义一个黎曼结构 g，使得 $(\mathscr{M}, \tilde{g})$ 不是一个黎曼流形。另外，纤维梯度在 x 处的值 $\operatorname{grad}_{\text{fib}} f(x)$ 是纤维空间 $\mathscr{M}_{\varpi(x)}$ 上的切向量，因而它与纤维空间相切，而其产生的流也只会将点在各自的纤维空间内移动。

由于纤维黎曼流形 $\mathscr{M}$ 的沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 难以分解出纤维黎曼结构，下面转而考虑所有条件分布所构成的空间：$\widetilde{\mathscr{P}}_2(\mathscr{M}) := \{q(z|y) \in \mathscr{P}_2(\mathscr{M}_y) \mid y \in \mathscr{M}_0\}$。由此定义可知，此分布空间可以局部分解为乘积空间 $\mathscr{M}_0 \times \mathscr{P}_2(\mathscr{M}_{\varpi(x)})$。由于纤维空间 $\mathscr{M}_{\varpi(x)}$ 具有黎曼结构，因此其沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M}_{\varpi(x)})$ 也有一个黎曼结构（参见 6.2.1 节），而此沃瑟斯坦空间正是 $\widetilde{\mathscr{P}}_2(\mathscr{M})$ 的纤维空间，因此 $\widetilde{\mathscr{P}}_2(\mathscr{M})$ 是一个纤维黎曼流形。在纤维空间 $\mathscr{P}_2(\mathscr{M}_y)$ 上，参见式 (6-1)，可以写出条件分布的 KL 散度 $\mathrm{KL}_{p(\cdot|y)}$ 的梯度在 $q(\cdot|y) \in \mathscr{P}_2(\mathscr{M}_{\varpi(x)})$ 处的表达式：作为 $\mathscr{M}$ 上的向量场 $\big(0_m, (g_{\mathscr{M}_y})^{ij}\partial_{z^j} \log \dfrac{q(\cdot|y)}{p(\cdot|y)}\big)$。因此分布空间 $\widetilde{\mathscr{P}}_2(\mathscr{M})$ 上的 KL 散度 $\mathrm{KL}_{p(x)}$ 的纤维梯度在 q 处的表达式为：

$$\big(\operatorname{grad}_{\text{fib}} \mathrm{KL}_p\big)(q)(x) = \tilde{g}^{ij}(x)\partial_j \log \frac{q(z|y)}{p(z|y)} = \tilde{g}^{ij}(x)\partial_j \log\big(q(x)/p(x)\big),$$

其中最后一个等式成立是因为在与 $(\tilde{g}^{ij})$ 相乘后，只有关于变量 z 的偏导数才能起到作用，而 $\partial_{z^j} \log q(y,z) = \partial_{z^j} \log q(z|y)$。在经 π 投影后，$\operatorname{grad}_{\text{fib}} \mathrm{KL}_p$ 便可成为整个纤维黎曼流形 $\mathscr{M}$ 的沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的一个向量场。注意到沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 可局部等价地表示为 $\mathscr{P}_2(\mathscr{M}_0) \times \widetilde{\mathscr{P}}_2(\mathscr{M})$，而 $\widetilde{\mathscr{P}}_2(\mathscr{M})$ 不是一个黎曼流形，因而难以直接在 $\mathscr{P}_2(\mathscr{M})$ 上得到 KL 散度的纤维梯度。

6.3.2　统一的理论框架

本节首先为 MCMC 动力学系统引入一个常规性假设（regularity assumption）以便更准确地描述理论框架。这个假设几乎所有现有 MCMC 方法都满足，并会在本节最后进一步讨论放松此假设。

假设 6.1 (常规 MCMC 动力学系统)：一个 MCMC 动力学系统被称为是常规的（regular），如果它在式 (2-2) 的表示形式中的矩阵 (D, Q) 还满足：**(a)** 扩散矩阵 $D = C$ 或 $D = 0$ 或 $D = \begin{pmatrix} 0 & 0 \\ 0 & C \end{pmatrix}$，其中 $C(x)$ 处处对称正定；**(b)** 卷曲矩阵 $Q(x)$

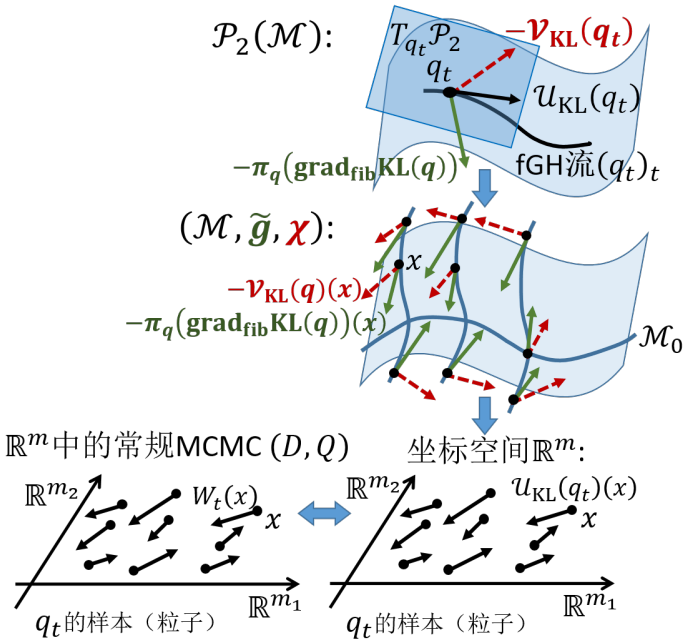


图 6.2　所提统一理论框架（定理 6.1）的图示：一个常规 MCMC 动力学系统等价于一个 fRP 流形 $\mathscr{M}$ 的沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的 fGH 流 $\mathcal{U}_{\mathrm{KL}_p}$。在分布 q_t 处的纤维梯度和哈密顿向量场作为流形 $\mathscr{M}$ 上的向量场分别以绿色实箭头和红色虚箭头表示。

处处满足式 (6-2)。

下面的定理正式叙述了所提的统一的理论框架。图 6.2 给出了此定理的图示。

定理 6.1 (统一理论框架：常规 MCMC 动力学系统与 $\mathscr{P}_2(\mathscr{M})$ 上 fGH 流的等价性)：称 $(\mathscr{M},\tilde{g},\chi)$ 是一个纤维黎曼-泊松（fiber-Riemannian Poisson, fRP）流形，并定义其沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的纤维梯度哈密顿（fiber-gradient Hamiltonian, fGH）流为：

$$\begin{aligned}\mathcal{U}_{\mathrm{KL}_p} &:= -\pi(\mathrm{grad}_{\mathrm{fib}}\,\mathrm{KL}_p) - \mathcal{V}_{\mathrm{KL}_p},\\ \mathcal{U}_{\mathrm{KL}_p}(q) &= \pi_q\big((\tilde{g}^{ij}+\chi^{ij})\partial_j \log(p/q)\partial_i\big).\end{aligned} \tag{6-10}$$

则有如下结论：**(a)** 任一 $\mathbb{R}^m$ 上的以分布 p 为平稳分布的常规 MCMC 动力学系统等价于一个特定 fRP 流形 $\mathscr{M}$ 的沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的 fGH 流 $\mathcal{U}_{\mathrm{KL}_p}$；**(b)** 反过来，对于任一 fRP 流形 $\mathscr{M}$，其沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的 fGH 流 $\mathcal{U}_{\mathrm{KL}_p}$ 等价于 $\mathscr{M}$ 的坐标空间中的一个以 p 为平稳分布的 MCMC 动力学系统；**(c)** 进一步，在上述两种情况下，fRP 流形 $\mathscr{M}$ 的纤维黎曼结构 $\tilde{g}$ 和泊松结构 χ 的坐标表示分别等于常规 MCMC 动力学系统的扩散矩阵 D 和卷曲矩阵 Q。

证明 对于一固定的分布 $q \in \mathscr{P}_2(\mathscr{M})$，流形 $\mathscr{M}$ 上的两个（确定性）动力学系统将会产生同样的分布曲线，如果它们在 ϖ 的投影下给出切空间 $T_q\mathscr{P}_2(\mathscr{M})$ 中相同

的切向量。所以证明 $\pi_q(W) = \mathcal{U}_{\mathrm{KL}_p}(q)$ 便足以说明定理中的两个动力学系统等价，其中 W 是引理 6.1 所给出的等价确定性 MCMC 动力学系统。这进一步等价于证明 $\pi_q(W - \mathcal{U}_{\mathrm{KL}_p}(q)) = 0_{\mathscr{L}^2_q}$，再使用投影 ϖ 的定义可得：$\mathrm{div}\left(q(W - \mathcal{U}_{\mathrm{KL}_p}(q))\right) = \mathrm{div}(q0_{\mathscr{L}^2_q}) = 0$（参见 6.2.1 节）。

首先考虑情况 (b)：给定一个 fRP 流形 $(\mathscr{M}, \tilde{g}, \chi)$，可定义一个 MCMC 动力学系统其扩散矩阵为 $(\tilde{g}^{ij})$ 而卷曲矩阵为 χ^{ij}，即分别等于纤维黎曼结构和泊松结构的坐标表示。这是一个常规 MCMC 动力学系统，因为假设 6.1 可以由 $(\tilde{g}^{ij})$ 的性质（参见式 (6-9)）及 (χ^{ij}) 的性质（参见 6.2.2 节）所满足。由引理 6.1，其等价的确定性动力学系统可由如下向量场给出：

$$W^i = \tilde{g}^{ij}\partial_j \log(p/q) + \chi^{ij}\partial_j \log p + \partial_j \chi^{ij}.$$

因此，

$$\begin{aligned}
&\mathrm{div}\left(q(W - \mathcal{U}_{\mathrm{KL}_p}(q))\right) \\
=\ &\mathrm{div}\left(q\big(\tilde{g}^{ij}\partial_j \log(p/q) + \chi^{ij}\partial_j \log p + \partial_j \chi^{ij} - (\tilde{g}^{ij} + \chi^{ij})\partial_j \log(p/q)\big)\,\partial_i\right) \\
=\ &\mathrm{div}\left(q(\partial_j \chi^{ij} + \chi^{ij}\partial_j \log q)\partial_i\right) = \mathrm{div}\left((q\partial_j \chi^{ij} + \chi^{ij}\partial_j q)\partial_i\right) \\
=\ &\mathrm{div}\left(\partial_j(q\chi^{ij})\partial_i\right) = \partial_i\partial_j(q\chi^{ij}) \\
=\ &0,
\end{aligned}$$

其中最后一个等式成立是由于矩阵 (χ^{ij}) 的反对称性。这表明，上面所构造的 MCMC 动力学系统与 fRP 流形 $\mathscr{M}$ 上的纤维梯度哈密顿流（fGH 流）$\mathcal{U}_{\mathrm{KL}_p}$ 等价。

对于情况 (a)，给定任一常规 MCMC 动力学系统，其矩阵表示 (D, Q) 满足假设 6.1，因而可以定义一个 fRP 流形 $(\mathscr{M}, \tilde{g}, \chi)$，其结构可通过在坐标空间中的表示来确定：$\tilde{g}^{ij} := D^{ij}$, $\chi^{ij} := Q^{ij}$。假设 6.1 可保证这样的 $\tilde{g}$ 是一个正确的纤维黎曼结构，而 χ 是一个正确的泊松结构。在这个构造出的流形上，可以重复上述过程，得到它上面的 fGH 流 $\mathcal{U}_{\mathrm{KL}_p}$，以及与之等价的常规 MCMC 动力学系统。这个常规 MCMC 动力学系统的等价确定性动力学系统可表达为如下向量场：

$$W^i = D^{ij}\partial_j \log(p/q) + Q^{ij}\partial_j \log p + \partial_j Q^{ij}.$$

它与原来所给的常规 MCMC 动力学系统的等价确定性动力学系统是一样的。这表明，原来所给的常规 MCMC 动力学系统与所构造的 fRP 流形上的 fGH 流 $\mathcal{U}_{\mathrm{KL}_p}$ 是等价的。

最后，由上述两种情况下的构造和分析可给出陈述 (c)。 □

定理中所给出的形式将常规 MCMC 动力学系统与沃瑟斯坦空间上的流统一了起来，并为一般 MCMC 动力学系统的行为给出了一个直接的解释。MCMC 方法最基本的要求，即目标分布 p 是平稳分布，在所提框架中变得十分显然，因为 $\mathcal{U}_{\mathrm{KL}_p}(p) = 0$。哈密顿流 $-\mathcal{V}_{\mathrm{KL}_p}$ 保持 KL_p（与分布 p 的差异）守恒，并可同时鼓励粒子在样本空间中探索更广阔的区域，从而可以加速收敛并降低样本自相关性[139]。纤维梯度流 $-\operatorname{grad}_{\mathrm{fib}} \mathrm{KL}_p$ 则在每一个纤维空间 $\mathscr{M}_{\varpi(x)}$ 中最小化 $\mathrm{KL}_{p(\cdot|y)}$（其中 $y = \varpi(x)$），驱使 $q_t(\cdot|y)$ 靠近 $p(\cdot|y)$ 从而促使动力学系统收敛。下面将讨论这个一般的行为在各具体情况下的分析。

6.3.3　统一框架下现有 MCMC 方法的分析

本节将在所提统一框架下为现有的 MCMC 方法做一个具体的分析。由扩散矩阵 D 的性质，这些 MCMC 方法可以分为三类。每一类都对应着 fRP 流形的一个特定纤维结构，因而对应的动力学系统会具有特定的行为。

类型 1:　D 是非奇异的（对应式 (6-9) 中 $m_1 = 0$）
这种情况下，纤维结构所对应的基空间 $\mathscr{M}_0$ 退化，而 $\mathscr{M}$ 自己就是其唯一的纤维空间，因此 $\mathscr{M}$ 此时就是一个黎曼流形，其黎曼结构的坐标表示为 $(g_{ij}) = D^{-1}$。条件分布流形 $\widetilde{\mathscr{P}_2}(\mathscr{M})$ 在此情况下就是 $\mathscr{M}$ 的沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$，因此纤维梯度流就是 $\mathscr{P}_2(\mathscr{M})$ 上的梯度流。此时 fGH 流变为梯度哈密顿流：

$$\mathcal{U}_{\mathrm{KL}_p} = -\pi(\operatorname{grad} \mathrm{KL}_p) - \mathcal{V}_{\mathrm{KL}_p}.$$

这意味着动力学系统的收敛性：哈密顿流 $-\mathcal{V}_{\mathrm{KL}_p}$ 保持 KL_p 守恒，而梯度流 $-\operatorname{grad} \mathrm{KL}_p$ 则在 $\mathscr{P}_2(\mathscr{M})$ 上最快地最小化 KL_p，因此它们一起还是会单调地最小化 KL_p，将分布曲线引向唯一的最小点 p。由于这个促使收敛的机制，这类动力学系统的模拟过程是可以允许随机梯度带来的适度扰动的。这个平稳分布的唯一性也与 Ma 等人[120] 所给的结论一致。

郎之万动力学系统（Langevin dynamics, LD）[58] 属于这一类 MCMC 动力学系统。它在实际中可使用基于整个数据集的准确梯度进行模拟[59]，也可使用基于随机子数据集的随机梯度（stochastic gradient, SG）进行模拟[60]。它的卷曲矩阵 $Q = 0$ 在流形 $\mathscr{M}$ 上定义了一个平凡的泊松结构，使得它的哈密顿流为零，因此它的 fGH 流中便只有梯度流。这使得人们能够对它的渐进与非渐进的收敛性质进行充分的分析（例如 [106-107,195-197,210]）。它的黎曼流形版本[127] 将 D 选择为费舍尔信息矩阵的逆，这样 $\mathscr{M}$ 就成为了信息几何[46] 中所考虑的分布流形。Patterson 等人[18] 进一步使用随机梯度对它进行模拟。

类型 2:　$D = 0$（对应式 (6-9) 中 $m_2 = 0$）

这种情况下，纤维结构所对应的基空间 $\mathscr{M}_0 = \mathscr{M}$ 而所有的纤维空间都是退化的。其 fGH 流 $\mathcal{U}_{\mathrm{KL}_p}$ 只包含哈密顿流 $-\mathcal{V}_{\mathrm{KL}_p}$。它可保持 KL_p 守恒并鼓励样本空间中的大范围探索。注意到在此情况下，KL 散度 KL_p 的减小并没有得到保证，因此在这类 MCMC 动力学系统的模拟过程中需要很小心。特别地，这类 MCMC 动力学系统不可使用多条平行的链进行模拟，除非这些链一开始就是以目标分布 p 进行的初始化。因此，这类 MCMC 动力学系统不适合使用 ParVI 形式进行模拟。动力学系统中稳定性因素的缺失也解释了它们直接使用随机梯度进行模拟的问题，因为随机梯度带来的扰动无法被控制。这个是对 HMC 的已有讨论[115-116] 向这类 MCMC 动力学系统的推广。

著名的哈密顿动力学系统（可参见 Marsden 等人的著作[142] 第 2 章）是此类中的代表，而 HMC 是基于对它进行模拟的采样方法。为从 ℓ 维样本空间 $\mathscr{Z}$ 上的目标分布 $p(Z)$ 中采样，目标变量（通常即为贝叶斯模型隐变量）Z 被增广为 $x = (Z, r)$，其中向量 $r \in \mathbb{R}^\ell$ 被称作动量（momentum）。在所提框架下，这个做法相当于是取 fRP 流形 $\mathscr{M}$ 为流形 $\mathscr{Z}$ 的余切丛 $T^*\mathscr{Z}$，其标准泊松结构（可参见 Da Silva 的著作[202] 第 2 章）对应着 $Q = (\chi^{ij}) = \begin{pmatrix} 0 & -I_\ell \\ I_\ell & 0 \end{pmatrix}$。另外，为定义增广之后的目标分布 $p(x) = p(Z)p(r|Z)$，一个特定的条件分布 $p(r|Z)$（正式地说，一个特定的 disintegration[139]）也被指定。HMC 可产生比 LD 更加有效的样本，因为哈密顿流可使样本走得更远从而降低自相关性[139]。如前面所提，HMC 的动力学系统并不能保证收敛。但它可依赖其模拟过程的遍历性（ergodicity）来保证收敛[98,104]。它以一个很细致的方式进行模拟：它使用的跳蛙积分器（leap-frog integrator）是辛的（symplectic）且是二阶的，并且动量 r 也会不断地从 $p(r|Z)$ 中重新采样。

HMC 考虑 $\mathscr{Z}$ 为欧氏空间的情况，并选择 $p(r|Z) = \mathcal{N}(0, \Sigma)$，而 Zhang 等人的工作[102] 则选取 $p(r|Z)$ 为多项伽马分布（monomial Gamma distribuiton）。Girolami 等人[127] 考虑了 $(\mathscr{Z}, g)$ 是黎曼流形的情况，并选取 $p(r|Z)$ 为 $\mathcal{N}\big(0, (g_{ij}(Z))\big)$（可看作余切空间 $T^*_Z\mathscr{Z}$ 中的标准高斯分布）。Byrne 等人[44] 通过考虑流形的嵌入空间，将哈密顿动力学系统在没有全局坐标系的流形上进行模拟，而 Lan 等人[128] 则考虑动力学系统的拉格朗日形式（即将动量（momentum）即余切向量（cotangent vector）替换为速度（velocity）即切向量（tangent vector））以得到更高效的模拟方法。

类型 3:　$D \neq 0$ 且 D 是奇异的（对应式 (6-9) 中 $m_1, m_2 \geqslant 1$）

这种情况下，基空间 $\mathscr{M}_0$ 和任一纤维空间 $\mathscr{M}_{\varpi(x)}$ 都是非退化的。fGH 流中的纤维

梯度流会在每一个纤维空间 $\mathscr{M}_{\varpi(x)}$ 中稳定对应的动力学系统，不过这对于随机梯度 MCMC 方法来说已经足够了，因为随机梯度只会出现在它们的纤维空间中。

SGHMC[115] 是这类方法中的第一个实例。哈密顿动力学系统类似，它也选取 $\mathscr{M}=T^*\mathscr{Z}$ 并使用同样的卷曲矩阵 Q，但它的扩散矩阵 $D_{2\ell\times 2\ell}$ 则取为假设 6.1 中的形式，其中 $C_{\ell\times\ell}$ 是一个常量。其逆矩阵 C^{-1} 为每一个纤维空间 $\mathscr{M}_{\varpi(x)}$ 定义了一个黎曼结构。在所提框架下，这个选择使得 fRP 流形 $\mathscr{M}$ 的纤维结构与纤维丛 $T^*\mathscr{Z}$ 的一致：$\mathscr{M}_0=\mathscr{Z}$，$\mathscr{M}_y=T^*_Z\mathscr{Z}$，且 $x=(y,z)=(Z,r)$。选定一个用于增广目标分布的条件分布 $p(r|Z)$，再应用引理 6.1，可以得到 SGHMC 的等价确定性动力学系统（以矩阵形式表示）：

$$\begin{cases}\dfrac{\mathrm{d}Z}{\mathrm{d}t}=-\nabla_r\log p(r|Z),\\[2ex]\dfrac{\mathrm{d}r}{\mathrm{d}t}=\nabla_Z\log p(Z)+\nabla_Z\log p(r|Z)+C\nabla_r\log\dfrac{p(r|Z)}{q(r|Z)}.\end{cases}\tag{6-11}$$

可以注意到，这个动力学系统是在哈密顿动力学系统中加入了动力学系统 $\dfrac{\mathrm{d}r}{\mathrm{d}t}=C\nabla_r\log\dfrac{p(r|Z)}{q(r|Z)}$，而这个被加入的动力学系统正是 $\mathscr{P}_2(\mathscr{M})$ 上的纤维梯度流 $-(\mathrm{grad}_{\mathrm{fib}}\,\mathrm{KL}_p)(q)$，或者说是纤维空间 $T^*_Z\mathscr{Z}$ 上的梯度流 $-(\mathrm{grad}\,\mathrm{KL}_{p(\cdot|Z)})(q(\cdot|Z))$。它会将 $q(\cdot|Z)$ 推向 $p(\cdot|Z)$。当使用随机梯度时，$Z\in\mathscr{Z}$ 的动力学系统不受影响，而在每个纤维空间 $T^*_Z\mathscr{Z}$ 中，随机梯度会带来一个噪声，从而会对 $q(\cdot|Z)$ 的演化造成扰动。而纤维空间中的纤维梯度则可以为 $q(\cdot|Z)$ 提供一个稳定的更新方向，从而弥补随机噪声带来的扰动，使得动力学系统对随机梯度变得鲁棒。

这类方法中的另一个有名的例子是随机梯度诺泽-胡佛恒温器方法（stochastic gradient Nosé-Hoover thermostats, SGNHT）[119]。它在增广变量 (Z,r) 的基础上，继续增广一个称为恒温器变量（thermostats）$\xi\in\mathbb{R}$ 的标量，从而可以更好地平衡随机梯度的噪声。在所提框架下来看，恒温器变量 ξ 增广了基空间 $\mathscr{M}_0$，而纤维空间则与 SGHMC 相同。

SGHMC 和 SGNHT 两方法都选择高斯条件分布 $p(r|Z)=\mathcal{N}(0,\Sigma^{-1})$，而随机梯度多项伽马恒温器方法（stochastic gradient monomial Gamma thermostats, SGMGT）[118] 将它选择为多项伽马分布，而 Lu 等人[103] 则根据一个相对论形式的能量函数选择 $p(r|Z)$ 以更好地调节动量 r 每一维的尺度。Ma 等人[120] 进一步考虑了 SGHMC 向黎曼流形 $(\mathscr{Z},g)$ 情况的拓展，而本文在第 3 章中提出了 SGHMC 和 SGNHT 在黎曼流形上的版本（其中 SGHMC 的版本与 Ma 等人[120] 的拓展不同）并考虑在流形的嵌入空间中模拟从而可适用于超球面这样的没有全局坐标系的流形。在本章的框架下，这些处理是将 $p(r|Z)$ 选为 $\mathcal{N}\big(0,(g_{ij}(Z))\big)$，并在每个纤维空间 $\mathscr{M}_y=T^*_Z\mathscr{Z}$ 中定义了黎曼结构 $\Big(\sqrt{(g^{ij}(Z))}^{\top}C^{-1}\sqrt{(g^{ij}(Z))}\Big)_{\ell\times\ell}$。

讨论　由于式 (2-2)、式 (6-6) 及式 (6-10) 这三个等价的动力学系统关于 D 和 Q 或者 $(\tilde{g}^{ij})$ 和 (χ^{ij}) 都是线性的，因而 MCMC 动力学系统可以进行组合。由上面的分析可知，SGHMC 动力学系统可以看作余切丛 $T^*\mathscr{Z}$ 上的哈密顿动力学系统与纤维空间（即余切空间）$T^*_Z\mathscr{Z}$ 上的 LD 的组合。另外一个例子是，SGMGT 方法的原工作[118] 中，作者也将属于类型 3 的 SGMGT 动力学系统与属于类型 1 的 LD 进行组合，得到一个属于类型 1 的新方法。由于类型 1 的动力学系统是在整个流形 $\mathscr{M}$ 上最小化 KL_p 而不再仅是在纤维空间上，因此组合后的类型 1 方法会比原本的 SGMGT 方法具有更好的收敛表现。这个判断与他们工作中所展示的实验结果相符。

下面对假设 6.1 进行讨论。文中提到的所有 MCMC 动力学系统都满足假设 6.1(a)，而除了与 SGNHT 相关的方法，其他的 MCMC 动力学系统也都满足假设 6.1(b)。假设 6.1(b) 存在例外，但可以注意到，由定理 6.1 的推导过程可知，假设 6.1(b) 只是为了满足 $\mathscr{M}$（进而 $\mathscr{P}_2(\mathscr{M})$）是一个泊松流形这个要求，而在推导过程中则没有用到。哈密顿向量场的定义及其哈密顿量守恒这个关键性质也都可以不依赖这个假设。所以，也许可以通过定义比泊松流形更加广泛的数学概念来解除假设 6.1(b)，并建立必要的概念和性质从而将所提框架进一步推广。假设 6.1(a) 也有望解除，例如可通过将一般的半正定矩阵使用可逆坐标变换将其变换为假设 6.1(a) 中所要求的形式，即在另一个坐标系中来看，对应的 MCMC 动力学系统是常规的。这些更加深入的探索方向可作为本章工作的后续工作。

6.4　MCMC 方法的 ParVI 形式模拟

所提统一框架（定理 6.1）将一个 MCMC 动力学系统识别为一个 fRP 流形 $\mathscr{M}$ 的沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上的 fGH 流，此流由式 (6-10) 显式表达。引理 6.1 则给出了另一种等价的确定性动力学系统，可与 fGH 流在 $\mathscr{P}_2(\mathscr{M})$ 上产生相同的分布曲线。这些发现可使一个 MCMC 方法通过借鉴 ParVI 方法中使用有限多个粒子的模拟方法来从沃瑟斯坦空间上的流的角度进行模拟，这便是 MCMC 方法的 ParVI 形式模拟。ParVI 方法和 MCMC 方法的这种结合为 ParVI 领域极大地拓展了可使用的动力学系统，也为 MCMC 方法带来了 ParVI 方法的好处，例如粒子高效性。

本节针对 SGHMC 这个特定的动力学系统来开发它所对应的 ParVI 形式的模拟。SGHMC 动力学系统选择条件分布 $p(r|Z) = \mathcal{N}(0, \Sigma)$ 其中协方差矩阵 Σ 是定值，因而动量 r 与变量 Z 在目标分布中是独立的，进而由引理 6.1 所得的 SGHMC

的确定性动力学系统式 (6-11) 可写作：

$$\begin{cases} \dfrac{\mathrm{d}Z}{\mathrm{d}t} = \Sigma^{-1} r, \\ \dfrac{\mathrm{d}r}{\mathrm{d}t} = \nabla_Z \log p(Z) - C\Sigma^{-1} r - C\nabla_r \log q(r). \end{cases} \tag{6-12}$$

而统一框架（定理 6.1）则给出了另一个等价的动力学系统。SGHMC 所对应的 fGH 流（式 (6-10)）为：

$$\begin{cases} \dfrac{\mathrm{d}Z}{\mathrm{d}t} = \Sigma^{-1} r + \nabla_r \log q(r), \\ \dfrac{\mathrm{d}r}{\mathrm{d}t} = \nabla_Z \log p(Z) - C\Sigma^{-1} r - C\nabla_r \log q(r) - \nabla_Z \log q(Z). \end{cases} \tag{6-13}$$

对这些基于流的动力学系统进行模拟的关键问题，是使用有限多个粒子的情况下，密度函数 q 是未知的。本文在上一章中，对 ParVI 领域中的这个问题进行了深入分析，并发现已有 ParVI 方法都是通过一个平滑操作（smoothing）来使用有限多个粒子来估计对数梯度 $\nabla \log q$ 的，并且这个平滑操作可以通过平滑密度或者通过平滑函数的方式实现。本节在此采用属于平滑密度操作的 Blob 方法[56] 来实现 SGHMC 的 ParVI 形式模拟。设 $\{r^{(i)}\}_i$ 为分布 $q(r)$ 的一组样本，而 K_r 是关于变量 r 的一个核函数（kernel）。则 Blob 方法使用如下方法给出对 $\nabla_r \log q(r)$ 的近似（可参见上一章 5.2.3 节）：

$$-\nabla_r \log q(r^{(i)}) \approx -\frac{\sum_k \nabla_{r^{(i)}} K_r^{(i,k)}}{\sum_j K_r^{(i,j)}} - \sum_k \frac{\nabla_{r^{(i)}} K_r^{(i,k)}}{\sum_j K_r^{(j,k)}},$$

其中 $K_r^{(i,j)} := K_r(r^{(i)}, r^{(j)})$。对对数梯度 $-\nabla_Z \log q(Z)$ 的近似可通过类似的方式实现。注意到在 $r^{(i)}$ 处核函数的梯度 $-\nabla_{r^{(i)}} K_r^{(i,k)}$ 会给出一个远离 $r^{(k)}$ 的方向，因此这个估计有效地为粒子之间引入了一个排斥相互作用（repulsive interaction），这与 SVGD 原论文[81] 中的讨论类似。另外，原初的 SGHMC 方法可以通过将 $-C\nabla_r \log q(r)\,\mathrm{d}t$ 替换为 $\mathcal{N}(0, 2C\,\mathrm{d}t)$ 来对动力学系统 (6-12) 进行随机性的模拟，但是动力学系统 (6-13) 则没有对应的随机性模拟方法。直接将 $\nabla_r \log q(r)$ 和 $\nabla_Z \log q(Z)$ 替换为对应变量上的随机噪声在此情况下是不合理的。例如 $\nabla_Z \log q(Z)$ 是用来更新变量 r 的，因而为变量 Z 加入噪声是不起效果的，而转而为变量 r 加入噪声则对应的是 $\nabla_r \log q(r)$ 所产生的更新，而不是 $\nabla_Z \log q(Z)$。最后，将 Blob 方法对梯度 $\nabla \log q$ 的估计代入，可得 pSGHMC-det 方法对粒子的

更新方式为：

$$\begin{cases} Z^{(i)} \leftarrow Z^{(i)} + \varepsilon \Sigma^{-1} r^{(i)}, \\ r^{(i)} \leftarrow r^{(i)} + \varepsilon \nabla_Z \log p(Z^{(i)}) \\ \qquad - \varepsilon C \Big(\Sigma^{-1} r^{(i)} + \dfrac{\sum_k \nabla_{r^{(i)}} K_r^{(i,k)}}{\sum_j K_r^{(i,j)}} + \sum_k \dfrac{\nabla_{r^{(i)}} K_r^{(i,k)}}{\sum_j K_r^{(j,k)}} \Big), \end{cases} \tag{6-14}$$

而 pSGHMC-fGH 方法的更新方式为：

$$\begin{cases} Z^{(i)} \leftarrow Z^{(i)} + \varepsilon \Big(\Sigma^{-1} r^{(i)} + \dfrac{\sum_k \nabla_{r^{(i)}} K_r^{(i,k)}}{\sum_j K_r^{(i,j)}} + \sum_k \dfrac{\nabla_{r^{(i)}} K_r^{(i,k)}}{\sum_j K_r^{(j,k)}} \Big), \\ r^{(i)} \leftarrow r^{(i)} + \varepsilon \nabla_Z \log p(Z^{(i)}) \\ \qquad - \varepsilon \Big(\dfrac{\sum_k \nabla_{Z^{(i)}} K_Z^{(i,k)}}{\sum_j K_Z^{(i,j)}} + \sum_k \dfrac{\nabla_{Z^{(i)}} K_Z^{(i,k)}}{\sum_j K_Z^{(j,k)}} \Big) \\ \qquad - \varepsilon C \Big(\Sigma^{-1} r^{(i)} + \dfrac{\sum_k \nabla_{r^{(i)}} K_r^{(i,k)}}{\sum_j K_r^{(i,j)}} + \sum_k \dfrac{\nabla_{r^{(i)}} K_r^{(i,k)}}{\sum_j K_r^{(j,k)}} \Big), \end{cases} \tag{6-15}$$

其中 ε 是步长，K_Z 是关于变量 Z 的核函数，并且 $K_Z^{(i,j)} := K_Z(Z^{(i)}, Z^{(j)})$。

这两个 ParVI 形式的 SGHMC 动力学系统的模拟方法分别被称为 pSGHMC-det（对应式 (6-12) 或式 (6-14)）和 pSGHMC-fGH（对应式 (6-13) 或式 (6-15)）（“p” 代表粒子（particle），“det” 代表确定性的（deterministic），而 fGH 代表 fGH 流）。与原初的 SGHMC 相比，这两个所提方法都是通过确定性的方式进行模拟，并且显式地出现了粒子之间的排斥相互作用，因而可以更快收敛并且具有粒子高效性。另一方面，SGHMC 动力学系统在向目标分布的收敛过程中可比 LD 更加高效，所以对应的 ParVI 方法也会比已有的基于 LD 的 ParVI 方法（例如 SVGD 和 Blob 方法）更加高效。读者可能会注意到，pSGHMC-det 方法与直接将带有动量的随机梯度下降方法（stochastic gradient descent with momentum, SGDM）[172] 应用于 Blob 所得方法十分类似，但需要强调的是，后者的推导过程没有理论保证，因为 Blob 方法是在沃瑟斯坦空间 $\mathscr{P}_2(\mathscr{M})$ 上最小化 KL 散度的，而不是在 $\mathscr{M}$ 优化某个目标函数的。此外，这两个 ParVI 方法也可以进一步受益于 ParVI 领域中的诸多先进技术，例如上一章所得到的 ParVI 方法的加速框架（5.4 节），HE 带宽选择方法（5.5 节），以及可近似 $-\nabla \log q$ 的更多方法（例如上一章所提的 GFSD 和 GFSF 方法，参见 5.3.3 节）。

图 6.3　各动力学系统的模拟结果。各行分别对应着 Blob，SGHMC，pSGHMC-det，和 pSGHMC-fGH 方法，每行相邻两图相差 300 轮迭代，而每行最后一图是 10,000 轮迭代之后的结果。

6.5　实验

以下各实验的代码可从网站 “https://github.com/chang-ml-thu/FGH-flow” 下载。

6.5.1　简单模拟实验

首先考察各动力学系统的模拟效果，以展示它们之间的等价性，以及所提 pSGHMC-det 和 pSGHMC-fGH 方法的优势。具体地，考虑二维欧氏空间上的目

标分布 $p(Z) = p(Z_1, Z_2)$ 定义为

$$\log p(Z_1, Z_2) = -0.01 \times \left(\frac{1}{2}(Z_1^2 + Z_2^2) + \frac{0.8}{2}(25Z_1 + Z_2^2)^2 \right) + \text{const.}$$

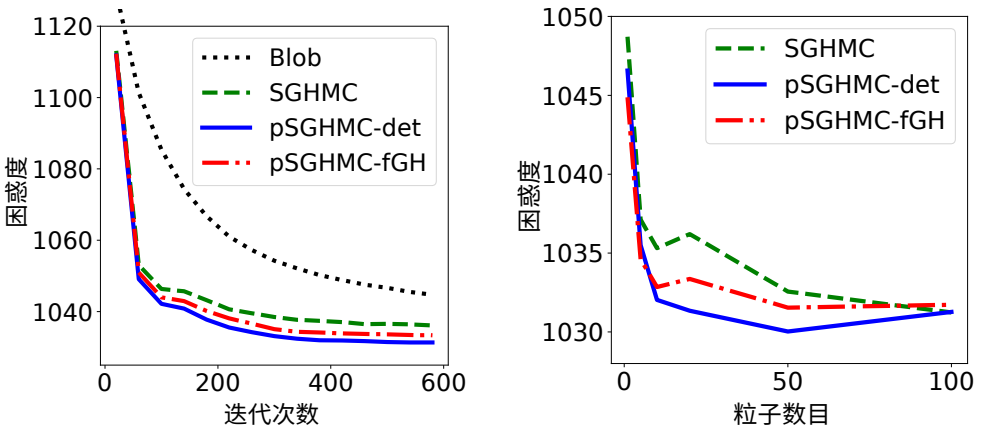
这个分布是受 Girolami 等人的工作[127] 中所使用的实验设定的启发而设计的。针对此目标分布，本实验考察 Blob，SGHMC 以及所提的 pSGHMC-det 和 pSGHMC-fGH 方法使用 50 个粒子对对应动力学系统进行模拟的过程。这 50 个粒子由高斯分布 $\mathcal{N}\big((-2,-7), 0.5^2 I\big)$ 来初始化。为公平比较，所有方法都采用相同的步长 $\varepsilon = 0.01$，而 SGHMC 相关方法（即 SGHMC，pSGHMC-det 和 pSGHMC-fGH）都选择相同的参数 $\Sigma^{-1} = 1.0$，$C = 0.5$。各 ParVI 方法（即 Blob，pSGHMC-det 和 pSGHMC-fGH）使用上一章 5.5 节所提出的 HE 带宽选择方法。所有方法都使用准确梯度进行模拟。图中展示的样本空间区域为 $[-7, 3] \times [-9, 9]$。

图 6.3 中展示了各方法的模拟过程。首先可以看出，所有方法最终都产生了依照目标分布而排列的粒子，这表明这些动力学系统是等价的，即它们都会使粒子的分布收敛到目标分布上。对于 ParVI 方法，所提的 pSGHMC-det（第 3 行）和 pSGHMC-fGH 两方法（第 4 行）都比 Blob 方法（第 1 行）具有更快的收敛速度。这展示了它们使用 SGHMC 动力学系统胜于 LD 的优势，其中 SGHMC 所引入的动量会在模拟过程中不断地在竖直方向积累从而加速粒子的演化过程。对于 SGHMC 动力学系统，可以发现所提的两个 ParVI 形式的模拟方法（第 3，4 行）收敛得比其原初的随机模拟方式（第 2 行）收敛得更快，这体现了确定性更新方法的优势。另外，作为一个 ParVI 方法，pSGHMC-fGH 方法（第 4 行）可明显受益于上一章中所提出的 HE 带宽选择方法，使得其所得粒子分布得十分整齐而规则，成为一组更具有代表性的粒子。pSGHMC-det 方法（第三行）则没有从 HE 方法中受益太多，因为原变量 Z 的粒子所代表的分布 $q(Z)$ 并没有直接在其动力学系统（式 (6-12)）中使用。

6.5.2　隐式狄利克雷分配模型实验

本部分实验考虑以隐式狄利克雷分配模型（latent Dirichlet allocation, LDA）在真实数据集上的后验推理任务考察所提 pSGHMC-det 和 pSGHMC-fGH 方法的优势。本实验使用与 SGNHT 原工作[119] 中相同的实验设定，亦即本文在上一章 5.6.4 节中所使用的设定。

所有方法的模拟中都使用基于在随机子数据集上使用坍缩吉布斯采样（collapsed Gibbs sampling）进行估计的随机梯度。为公平比较，所有方法都采用相同的步长 $\varepsilon = 1 \times 10^{-3}$，而 SGHMC 相关的方法都选择相同的参数 $\Sigma^{-1} = 300$，



(a) 使用 20 个粒子的学习曲线　　(b) 使用多种数目粒子在 600 轮迭代后的结果

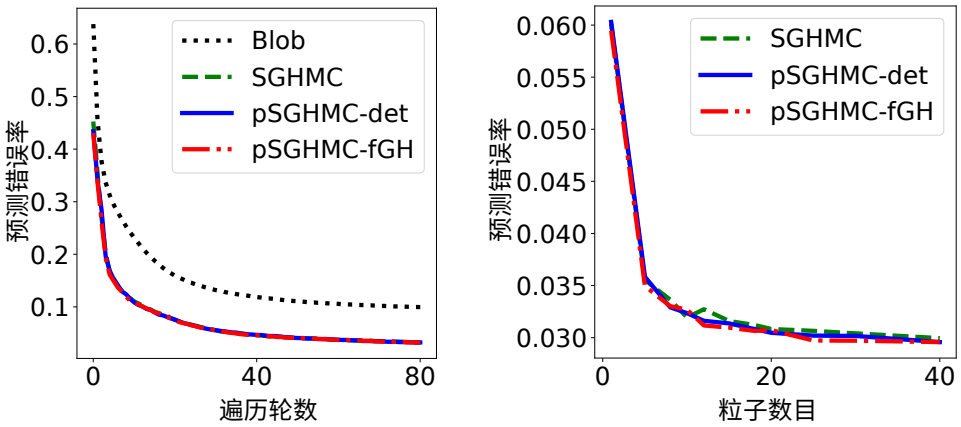
图 6.4　隐式狄利克雷分配模型在 ICML 数据集上的后验推理实验中所提 pSGHMC-det 和 pSGHMC-fGH 方法的表现。

$C = 0.1$。各 ParVI 方法都使用上一章 5.5 节中所提出的 HE 带宽选择方法。为与 ParVI 方法的模拟方式匹配，SGHMC 同时平行地模拟多条链并取各链最后的样本作为整个方法所得样本（与上一节 6.5.1 中的情况相同）。

图 6.4 展示了实验结果，其中每条曲线是 10 次独立运行结果的平均。其中，由图 6.4(a) 可以发现所提方法 pSGHMC-det 和 pSGHMC-fGH 显著地比 Blob 方法收敛得更快，这得益于 SGHMC 动力学系统的优势。各方法的粒子高效性则在图 6.4(b) 中进行了比较，可发现所提方法 pSGHMC-det 和 pSGHMC-fGH 在各粒子数目下基本上都可比原初随机模拟的 SGHMC 方法取得更好的结果。这体现了 ParVI 方法的粒子高效性，得益于它们直接考虑粒子间的相互作用从而可以充分发挥一组固定数目的粒子的近似能力。

6.5.3　贝叶斯神经网络实验

本部分实验在贝叶斯神经网络（Bayesian neural networks）的后验推理这个有监督的学习任务上考察所提方法。实验采用与 SGHMC 原工作[115] 中相同的设定。数据集选为标准的 MNIST 数据集。所考虑网络结构为三层全链接（fully connected）前馈（feedforward）神经网络，各层隐节点数（或称神经元数）为 784-100-10。激活函数（activation function）选为 sigmoid 函数。所有方法都使用随机梯度进行模拟，对应的随机子数据集大小为 500。SGHMC 相关的方法（即 SGHMC，pSGHMC-det 和 pSGHMC-fGH）使用同样的步长 $\varepsilon = 5 \times 10^{-5}$ 和参数 $\Sigma^{-1} = 1.0$，$C = 1.0$，而 Blob 方法使用步长 $\varepsilon = 5 \times 10^{-8}$（更大的步长会导致结果发散）。



(a) 使用 10 个粒子的学习曲线　　(b) 使用多种数目粒子在遍历轮数为 80 时的结果

图 6.5　贝叶斯神经网络在 MNIST 数据集上的后验推理任务中所提 pSGHMC-det 和 pSGHMC-fGH 方法的表现。

图 6.5 展示了实验结果，其中每条曲线是 10 次独立运行结果的平均。其中，图 6.5(a) 再次印证了所期望的效果：基于 SGHMC 动力学系统的 pSGHMC-det 和 pSGHMC-fGH 方法比基于 LD 的 Blob 方法收敛得更快。而图 6.5(b) 所展示的优势虽然不明显，但也能看出所提 pSGHMC-det 和 pSGHMC-fGH 方法的粒子高效性。

6.6　本章小结与讨论

本章建立了一个将一般 MCMC 动力学系统与沃瑟斯坦空间上的流联系起来的理论框架。通过引入新的概念，本章工作发现常规的 MCMC 动力学系统对应着 fRP 流形的沃瑟斯坦空间上的 fGH 流。其中 fGH 流的性质为一般 MCMC 动力学系统的行为给出了一个明晰的理解，而这个与沃瑟斯坦空间上的流的联系也使 ParVI 形式的 MCMC 动力学系统的模拟成为可能。本章在所提框架下对各现有 MCMC 动力学系统依 3 种类型进行了具体分析，并为 SGHMC 动力学系统开发了两个 ParVI 形式的模拟方法。实验展示了在 ParVI 领域中使用比 LD 更多的 MCMC 动力学系统所带来的更快收敛速度，以及使用 ParVI 方法进行模拟为 MCMC 方法所带来的粒子高效性。

所提 MCMC 动力学系统的理论框架可启发对更多 MCMC 方法的进一步分析和改进，例如将哈密顿流解释为测地流[211] 从而利用“测地流 + （纤维）梯度流”的结构进一步分析 MCMC 方法的收敛性，为更多 MCMC 方法开发对应的 ParVI 方法，以及在各个纤维空间中考虑对梯度流进行加速[123-124] 和方差缩减[156]。此外，一些其他领域中的问题和方法也具有纤维丛的结构，例如强化学习中所关心的策

略（policy）[23] $p^*(a|s) \in \{\, p(\cdot|s) \in \mathscr{P}(\mathscr{A}) \mid s \in \mathscr{S} \,\}$，以及深度神经网络[212] 等。所提理论框架可为分析这类问题提供新的思路，并有望将 MCMC 的理论和方法用于解决这类问题上。

第 7 章　总结与展望

7.1　本文总结

针对当下环境对贝叶斯推理任务所带来的多方面的高效性需求，本文通过利用隐变量和分布的流形结构，建立了理解各类贝叶斯推理方法的表现和联系的理论框架，启发和实现了新的高效方法，从而提高了各类贝叶斯推理方法处理流形隐变量的效率、近似灵活性、粒子高效性以及时间或算力高效性。具体贡献如下：

第 3 章提出了两个随机梯度测地线 MCMC 方法，从而使流形隐变量面对大规模数据时可快速而准确地进行推理。通过利用流形的嵌入空间，所提 MCMC 方法可处理隐变量处于没有全局坐标系的流形上这一难以解决的情况，例如超球面隐变量。两方法都可使用随机梯度进行模拟，从而具有可扩展性这一对数据规模不敏感的高效处理能力，而它们的正确性则使得它们可以给出渐进准确的估计。在球面混合模型上的实验结果表明，所提两方法可取得比已有方法都快且准的结果。

第 4 章提出了黎曼-斯坦因变分梯度下降方法，为处理流形隐变量的推理任务带来了具有粒子高效性和近似灵活性的方法，并提高了基于粒子的变分推理方法（particle-based variational inference, ParVI）的迭代效率。所提方法为解决一般流形与欧氏空间截然不同的性质，使用了新的技术。它在嵌入空间中的形式使得它可用于没有全局坐标系的流形上的隐变量，而它使用信息几何方法的能力可使它在处理欧氏空间上的推理任务时可更准确地利用模型的信息。

第 5 章提出了 ParVI 方法所作假设的理论分析，并基于此分析提出了两个新 ParVI 方法，以及可提高所有 ParVI 方法迭代效率及粒子高效性的加速框架和带宽选择方法。所作理论分析发现 ParVI 方法都需要做一个平滑性假设，并由此发现了各 ParVI 方法之间的等价性以及开发新 ParVI 方法的原则和思路，进而由两个新 ParVI 方法的提出展示了此发现的理论意义。而此理论的实际意义则由加速框架和带宽选择方法体现。加速方法可增强所有 ParVI 方法利用梯度信息的能力，从而提高了它们的迭代效率与算力高效性。所提带宽选择方法基于使用核函数的目标，使得 ParVI 方法可以得到更有代表性的样本。

第 6 章提出了 MCMC 动力学系统作为沃瑟斯坦空间上的流的理论框架，使 MCMC 方法的行为和性质的机理变得清晰，同时也将 MCMC 方法与 ParVI 方法建立了联系。通过利用此理论所发现的联系，ParVI 方法可以使用更高效的动力学系统，而 MCMC 方法则可以拥有更具粒子高效性的实现方法。

7.2　未来工作展望

本文利用流形结构，建立了一些关于贝叶斯推理方法的理论，并实现了一些提高其高效性的方法，但在利用流形结构和解决贝叶斯推理任务两方面都还存在很多有待解决的问题和值得探索的方向。

（1）利用流形结构进行建模。流形可以给出一个思考和解决问题的本质视角，从而可以反映数据或模型的一些根本特征。例如针对图像数据的旋转不变性，已有一些模型[213-215] 利用球面结构或李群（Lie group）结构[216] 进行建模，而在图（graph）的高效表示任务和图的嵌入任务中也有一些方法[217-219] 利用了图的拓扑结构。但在贝叶斯模型领域，类似的工作则少得多。利用数据的流形结构可进一步增强贝叶斯模型的建模能力和实用价值，而本文工作可为这类模型提供多种高效推理方法以完成训练，因而这是一个可行且值得探索的研究方向。

另外，在考虑模型的流形结构方面，近来也有工作从黎曼流形的角度理解神经网络[220] 并通过流形的纤维丛结构解决了其灾难性遗忘（catastrophic forgetting）问题[212]。本文也注意到强化学习模型中的策略（policy）也具有纤维丛结构[23]。利用模型的流形结构分析和设计模型也将是一个十分有意义的研究方向。

（2）提高在复杂环境中利用流形结构的能力。流形结构虽可提供指导推理的关键信息，但对于复杂模型和结构化数据来说，对应的流形结构可能很复杂，使得一方面获取流形结构的信息变得困难，另一方面在学习过程中利用流形结构变得吃力。例如对于贝叶斯神经网络来说，其费舍尔信息矩阵很难求得闭形式，并且由于参数维度很高，使得利用此矩阵进行计算也会带来很大开销。近来有一些工作[49-50,157] 着力解决这些问题，但对于一般的复杂贝叶斯模型的更有原则性的方法仍有待探索。再如对于具有图（graph）结构的数据，已有工作[219] 利用了树状结构数据与双曲空间结构的相似性，但对于一般的图结构，其对应的流形由于缺乏便于实现的测地线、距离等结构而难以在实际中应用。加强实际任务中利用流形结构的能力可进一步发挥流形结构的优势，从而从更本质的角度解决这些学习问题。

（3）特定场景需求下的贝叶斯推理方法。高维空间总是难以处理的情况。现有的推理方法在处理高维隐变量时都会出现性能下降的表现，特别是基于粒子的方法会表现出粒子聚集这一退化现象[169]。如何在高维情形下更加经济地利用粒子是这类方法要解决的难题。另外，一些特殊任务也为推理方法带来了新的要求，例如使用离散隐变量的模型需要高效处理离散变量，以及在线学习（online learning）需要推理方法可充分利用每一时刻到来的数据并给出快速而可靠的更新。解决这些需求可进一步提高贝叶斯方法的实用性，从而将它的优势在更多机器学习任务上得到发挥。

参考文献

[1] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[2] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems. Lake Tahoe, Nevada USA: NIPS Foundation, 2012: 1097-1105.

[3] Kingma D P, Welling M. Auto-encoding variational Bayes[C]//Proceedings of the International Conference on Learning Representations (ICLR 2014). Banff, Canada: ICLR Committee, 2014.

[4] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems. Montréal, Canada: NIPS Foundation, 2014: 2672-2680.

[5] Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587):484.

[6] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. Las Vegas, Nevada USA: The Computer Vision Foundation, 2016: 770-778.

[7] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. arXiv preprint arXiv:1312.6199, 2013.

[8] Adam K. On the relation between robust and Bayesian decision making[J]. Journal of economic dynamics and control, 2004, 28(10):2105-2117.

[9] Hishinuma T, Senda K. Robust and explorative behavior in model-based Bayesian reinforcement learning[C]//2016 IEEE Symposium Series on Computational Intelligence (SSCI). Athens, Greece: IEEE, 2016: 1-8.

[10] Fei-Fei L, Fergus R, Perona P. One-shot learning of object categories[J]. IEEE transactions on pattern analysis and machine intelligence, 2006, 28(4):594-611.

[11] Schmidhuber J. Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook[D]. München: Technische Universität München, 1987.

[12] Lemke C, Budka M, Gabrys B. Metalearning: a survey of trends and technologies[J]. Artificial intelligence review, 2015, 44(1):117-130.

[13] Erhan D, Bengio Y, Courville A, et al. Visualizing higher-layer features of a deep network[J]. University of Montreal, 2009, 1341(3):1.

[14] Neal R M. Bayesian learning for neural networks: volume 118[M]. New York: Springer Science & Business Media, 2012

[15] Li Y, Gal Y. Dropout inference in Bayesian neural networks with alpha-divergences[C]// Proceedings of the 34th International Conference on Machine Learning (ICML 2017). Sydney, Australia: IMLS, 2017: 2052-2061.

[16] Lake B M, Salakhutdinov R, Tenenbaum J B. Human-level concept learning through probabilistic program induction[J]. Science, 2015, 350(6266):1332-1338.

[17] Yoon J, Kim T, Dia O, et al. Bayesian model-agnostic meta-learning[C]//Advances in Neural Information Processing Systems. Montréal, Canada: NIPS Foundation, 2018: 7343-7353.

[18] Patterson S, Teh Y W. Stochastic gradient Riemannian Langevin dynamics on the probability simplex[C]//Advances in Neural Information Processing Systems. Lake Tahoe, Nevada USA: NIPS Foundation, 2013: 3102-3110.

[19] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation[J]. The Journal of Machine Learning Research, 2003, 3:993-1022.

[20] Mcauliffe J D, Blei D M. Supervised topic models[C]//Advances in Neural Information Processing Systems. Vancouver, Canada: NIPS Foundation, 2008: 121-128.

[21] Zhu J, Ahmed A, Xing E P. MedLDA: maximum margin supervised topic models[J]. Journal of Machine Learning Research, 2012, 13(Aug):2237-2278.

[22] Reisinger J, Waters A, Silverthorn B, et al. Spherical topic models[C]//Proceedings of the 27th International Conference on Machine Learning (ICML 2010). Haifa, Israel: IMLS, 2010: 903-910.

[23] Zhang R, Chen C, Li C, et al. Policy optimization as Wasserstein gradient flows[C]//Proceedings of the 35th International Conference on Machine Learning (ICML 2018). Stockholm, Sweden: IMLS, 2018: 5741-5750.

[24] Haarnoja T, Tang H, Abbeel P, et al. Reinforcement learning with deep energy-based policies[C]//Proceedings of the 34th International Conference on Machine Learning (ICML 2017). Sydney, Australia: IMLS, 2017: 1352-1361.

[25] Liu Y, Ramachandran P, Liu Q, et al. Stein variational policy gradient[C]//Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI 2017). Sydney, Australia: Association for Uncertainty in Artificial Intelligence, 2017.

[26] Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]//Proceedings of the 35th International Conference on Machine Learning (ICML 2018). Stockholm, Sweden: IMLS, 2018: 1856-1865.

[27] Gregor K, Danihelka I, Graves A, et al. DRAW: a recurrent neural network for image generation[C]//Proceedings of the 32nd International Conference on Machine Learning (ICML 2015). Lille, France: IMLS, 2015: 1462-1471.

[28] Burda Y, Grosse R, Salakhutdinov R. Importance weighted autoencoders[J]. arXiv preprint arXiv:1509.00519, 2015.

[29] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1977, 39(1): 1-22.

[30] Hoffman M D, Blei D M, Wang C, et al. Stochastic variational inference[J]. The Journal of Machine Learning Research, 2013, 14(1):1303-1347.

[31] Blei D M, Kucukelbir A, McAuliffe J D. Variational inference: A review for statisticians[J]. Journal of the American Statistical Association, 2017, 112(518):859-877.

[32] Davidson T R, Falorsi L, De Cao N, et al. Hyperspherical variational auto-encoders[J]. arXiv preprint arXiv:1804.00891, 2018.

[33] Taghia J, Ma Z, Leijon A. Bayesian estimation of the von Mises-Fisher mixture model with variational inference[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(9):1701-1715.

[34] Mathieu E, Lan C L, Maddison C J, et al. Hierarchical representations with Poincaré variational auto-encoders[J]. arXiv preprint arXiv:1901.06033, 2019.

[35] Grattarola D, Livi L, Alippi C. Adversarial autoencoders with constant-curvature latent manifolds[J]. arXiv preprint arXiv:1812.04314, 2018.

[36] Ovinnikov I. Poincaré Wasserstein autoencoder[J]. arXiv preprint arXiv:1901.01427, 2019.

[37] Nagano Y, Yamaguchi S, Fujita Y, et al. A differentiable Gaussian-like distribution on hyperbolic space for gradient-based learning[J]. arXiv preprint arXiv:1902.02992, 2019.

[38] Gromov M. Hyperbolic groups[M]//Essays in group theory. New York: Springer, 1987: 75-263

[39] Salakhutdinov R, Mnih A. Bayesian probablistic matrix factorization using Markov chain Monte Carlo[C]//Proceedings of the 25th International Conference on Machine Learning (ICML 2008). Helsinki, Finland: Omnipress, 2008: 880-887.

[40] Song Y, Zhu J. Bayesian matrix completion via adaptive relaxed spectral regularization[C]//The 30th AAAI Conference on Artificial Intelligence (AAAI-16). Phoenix, Arizona USA: AAAI press, 2016: 2044-2050.

[41] Yanush V, Kropotov D. Hamiltonian Monte-Carlo for orthogonal matrices[J]. arXiv preprint arXiv:1901.08045, 2019.

[42] Stiefel E L. Richtungsfelder und fernparallelismus in n-dimensionalen mannigfaltigkeiten[J]. Commentarii Mathematici Helvetici, 1935, 8(1):305–353.

[43] James I M. The topology of Stiefel manifolds: volume 24[M]. New York: Cambridge University Press, 1976

[44] Byrne S, Girolami M. Geodesic Monte Carlo on embedded manifolds[J]. Scandinavian Journal of Statistics, 2013, 40(4):825-845.

[45] Amari S I, Nagaoka H. Methods of information geometry: volume 191[M]. Providence, Rhode Island: American Mathematical Soc., 2007

[46] Amari S I. Information geometry and its applications[M]. Tokyo: Springer, 2016

[47] Amari S I. Natural gradient works efficiently in learning[J]. Neural computation, 1998, 10(2): 251-276.

[48] Ollivier Y. Riemannian metrics for neural networks I: feedforward networks[J]. Information and Inference: A Journal of the IMA, 2015, 4(2):108-153.

[49] Marceau-Caron G, Ollivier Y. Practical Riemannian neural networks[J]. arXiv preprint arXiv:1602.08007, 2016.

[50] Khan M E, Nielsen D. Fast yet simple natural-gradient descent for variational inference in complex models[C]//2018 International Symposium on Information Theory and Its Applications (ISITA). Singapore: IEEE, 2018: 31-35.

[51] Chen Y, Li W. Natural gradient in Wasserstein statistical manifold[J]. arXiv preprint arXiv:1805.08380, 2018.

[52] Otto F. The geometry of dissipative evolution equations: the porous medium equation[J]. 2001.

[53] Villani C. Optimal transport: old and new: volume 338[M]. Berlin: Springer Science & Business Media, 2008

[54] Ambrosio L, Gigli N, Savaré G. Gradient flows: in metric spaces and in the space of probability measures[M]. Berlin: Springer Science & Business Media, 2008

[55] Chen C, Zhang R. Particle optimization in stochastic gradient MCMC[J]. arXiv preprint arXiv:1711.10927, 2017.

[56] Chen C, Zhang R, Wang W, et al. A unified particle-optimization framework for scalable Bayesian sampling[C]//Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI 2018). Monterey, California USA: Association for Uncertainty in Artificial Intelligence, 2018.

[57] Taghvaei A, Mehta P G. Accelerated gradient flow for probability distributions[C]//Proceedings of the 36th International Conference on Machine Learning (ICML 2019). Long Beach, California USA: IMLS, 2019.

[58] Roberts G O, Tweedie R L, et al. Exponential convergence of Langevin distributions and their discrete approximations[J]. Bernoulli, 1996, 2(4):341-363.

[59] Roberts G O, Stramer O. Langevin diffusions and Metropolis-Hastings algorithms[J]. Methodology and computing in applied probability, 2002, 4(4):337-357.

[60] Welling M, Teh Y W. Bayesian learning via stochastic gradient Langevin dynamics[C]// Proceedings of the 28th International Conference on Machine Learning (ICML 2011). Bellevue, Washington USA: IMLS, 2011: 681-688.

[61] Jordan R, Kinderlehrer D, Otto F. The variational formulation of the Fokker-Planck equation [J]. SIAM journal on mathematical analysis, 1998, 29(1):1-17.

[62] Wainwright M J, Jordan M I, et al. Graphical models, exponential families, and variational inference[J]. Foundations and Trends® in Machine Learning, 2008, 1(1–2):1-305.

[63] Gershman S J, Hoffman M D, Blei D M. Nonparametric variational inference[C]//Proceedings of the 29th International Conference on Machine Learning (ICML 2012). Edinburgh, Scotland: IMLS, 2012.

[64] Dai B, He N, Dai H, et al. Provable Bayesian inference via particle mirror descent[C/OL]// Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS-16): volume 51. Cadiz, Spain: PMLR, 2016: 985-994. http://proceedings.mlr.press/v51/dai16.html.

[65] Minka T P. A family of algorithms for approximate Bayesian inference[D]. Cambridge: Massachusetts Institute of Technology, 2001.

[66] Hernández-Lobato J M, Adams R. Probabilistic backpropagation for scalable learning of Bayesian neural networks[C]//Proceedings of the 32nd International Conference on Machine Learning (ICML 2015). Lille, France: IMLS, 2015: 1861-1869.

[67] Li Y, Hernández-Lobato J M, Turner R E. Stochastic expectation propagation[C]//Advances in Neural Information Processing Systems. Montréal, Canada: NIPS Foundation, 2015: 2323-2331.

[68] Blundell C, Cornebise J, Kavukcuoglu K, et al. Weight uncertainty in neural networks[J]. arXiv preprint arXiv:1505.05424, 2015.

[69] Cybenko G. Approximations by superpositions of a sigmoidal function[J]. Mathematics of Control, Signals and Systems, 1989, 2:183-192.

[70] Hornik K. Approximation capabilities of multilayer feedforward networks[J]. Neural networks, 1991, 4(2):251-257.

[71] Csáji B C. Approximation with artificial neural networks[J]. Faculty of Sciences, Etvs Lornd University, Hungary, 2001, 24:48.

[72] Rezende D, Mohamed S. Variational inference with normalizing flows[C]//Proceedings of The 32nd International Conference on Machine Learning (ICML 2015). Lille, France: IMLS, 2015: 1530-1538.

[73] Kingma D P, Salimans T, Jozefowicz R, et al. Improved variational inference with inverse autoregressive flow[C]//Advances in Neural Information Processing Systems. Barcelona, Spain: NIPS Foundation, 2016: 4743-4751.

[74] Makhzani A, Shlens J, Jaitly N, et al. Adversarial autoencoders[J]. arXiv preprint arXiv:1511.05644, 2015.

[75] Mescheder L, Nowozin S, Geiger A. Adversarial variational Bayes: Unifying variational autoencoders and generative adversarial networks[C]//Proceedings of the 34th International Conference on Machine Learning (ICML 2017). Sydney, Australia: IMLS, 2017: 2391-2400.

[76] Polyak B T. Some methods of speeding up the convergence of iteration methods[J]. USSR Computational Mathematics and Mathematical Physics, 1964, 4(5):1-17.

[77] Nesterov Y. A method of solving a convex programming problem with convergence rate $O(1/k^2)$[C]//Soviet Mathematics Doklady: volume 27. Moscow: Russian Academy of Sciences, 1983: 372-376.

[78] Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization[J]. Journal of Machine Learning Research, 2011, 12(Jul):2121-2159.

[79] Zeiler M D. AdaDelta: an adaptive learning rate method[J]. arXiv preprint arXiv:1212.5701, 2012.

[80] Kingma D P, Ba J. Adam: A method for stochastic optimization[C]//Proceedings of the International Conference on Learning Representations (ICLR 2015). San Diego, California USA: ICLR Committee, 2015.

[81] Liu Q, Wang D. Stein variational gradient descent: A general purpose Bayesian inference algorithm[C]//Advances in Neural Information Processing Systems. Barcelona, Spain: NIPS Foundation, 2016: 2370-2378.

[82] Gorham J, Mackey L. Measuring sample quality with kernels[J]. arXiv preprint arXiv:1703.01717, 2017.

[83] Chen W Y, Mackey L, Gorham J, et al. Stein points[J]. arXiv preprint arXiv:1803.10161, 2018.

[84] Futami F, Cui Z, Sato I, et al. Frank-Wolfe Stein sampling[J]. arXiv preprint arXiv:1805.07912, 2018.

[85] Liu Q. Stein variational gradient descent as gradient flow[C]//Advances in Neural Information Processing Systems. Long Beach, California USA: NIPS Foundation, 2017: 3118-3126.

[86] Metropolis N, Rosenbluth A W, Rosenbluth M N, et al. Equation of state calculations by fast computing machines[J]. The journal of chemical physics, 1953, 21(6):1087-1092.

[87] Hastings W K. Monte Carlo sampling methods using Markov chains and their applications[J]. Biometrika, 1970, 57(1):97-109.

[88] Veach E, Guibas L J. Optimally combining sampling techniques for Monte Carlo rendering[C]// Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques. Los Angeles, California USA: ACM, 1995: 419-428.

[89] Neal R M. Annealed importance sampling[J]. Statistics and computing, 2001, 11(2):125-139.

[90] Del Moral P. Non-linear filtering: interacting particle resolution[J]. Markov processes and related fields, 1996, 2(4):555-581.

[91] Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images[M]//Readings in Computer Vision. Los Altos, California USA: Elsevier, 1987: 564-584

[92] Griffiths T L, Steyvers M. Finding scientific topics[J]. Proceedings of the National academy of Sciences, 2004, 101(suppl 1):5228-5235.

[93] Li A Q, Ahmed A, Ravi S, et al. Reducing the sampling complexity of topic models[C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, New York USA: ACM, 2014: 891-900.

[94] Yuan J, Gao F, Ho Q, et al. LightLDA: Big topic models on modest computer clusters[C]// Proceedings of the 24th International Conference on World Wide Web. Florence, Italy: International World Wide Web Conferences Steering Committee, 2015: 1351-1361.

[95] Chen J, Li K, Zhu J, et al. WarpLDA: a cache efficient $o(1)$ algorithm for latent Dirichlet allocation[J]. Proceedings of the VLDB Endowment, 2016, 9(10):744-755.

[96] Duane S, Kennedy A D, Pendleton B J, et al. Hybrid Monte Carlo[J]. Physics Letters B, 1987, 195(2):216-222.

[97] Neal R M. MCMC using Hamiltonian dynamics[J]. Handbook of Markov Chain Monte Carlo, 2011, 2.

[98] Betancourt M. A conceptual introduction to Hamiltonian Monte Carlo[J]. arXiv preprint arXiv:1701.02434, 2017.

[99] Beskos A, Pinski F J, Sanz-Serna J M, et al. Hybrid Monte Carlo on Hilbert spaces[J]. Stochastic Processes and their Applications, 2011, 121(10):2201-2230.

[100] Hoffman M D, Gelman A. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo.[J]. Journal of Machine Learning Research, 2014, 15(1):1593-1623.

[101] Strathmann H, Sejdinovic D, Livingstone S, et al. Gradient-free Hamiltonian Monte Carlo with efficient kernel exponential families[C]//Advances in Neural Information Processing Systems. Montréal, Canada: NIPS Foundation, 2015: 955-963.

[102] Zhang Y, Wang X, Chen C, et al. Towards unifying Hamiltonian Monte Carlo and slice sampling [C]//Advances in Neural Information Processing Systems. Barcelona, Spain: NIPS Foundation, 2016: 1741-1749.

[103] Lu X, Perrone V, Hasenclever L, et al. Relativistic Monte Carlo[J]. arXiv preprint arXiv:1609.04388, 2016.

[104] Livingstone S, Betancourt M, Byrne S, et al. On the geometric ergodicity of Hamiltonian Monte Carlo[J]. arXiv preprint arXiv:1601.08057, 2016.

[105] Dalalyan A S. Theoretical guarantees for approximate sampling from smooth and log-concave densities[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2017, 79(3):651-676.

[106] Cheng X, Bartlett P. Convergence of Langevin MCMC in KL-divergence[J]. arXiv preprint arXiv:1705.09048, 2017.

[107] Durmus A, Moulines E, et al. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm[J]. The Annals of Applied Probability, 2017, 27(3):1551-1587.

[108] Durmus A, Moulines E, Saksman E. On the convergence of Hamiltonian Monte Carlo[J]. arXiv preprint arXiv:1705.00166, 2017.

[109] Mangoubi O, Smith A. Rapid mixing of Hamiltonian Monte Carlo on strongly log-concave distributions[J]. arXiv preprint arXiv:1708.07114, 2017.

[110] Cheng X, Chatterji N S, Bartlett P L, et al. Underdamped Langevin MCMC: A non-asymptotic analysis[J]. arXiv preprint arXiv:1707.03663, 2017.

[111] Robbins H, Monro S. A stochastic approximation method[J]. The annals of mathematical statistics, 1951:400-407.

[112] Sato I, Nakagawa H. Approximation analysis of stochastic gradient Langevin dynamics by using Fokker-Planck equation and Ito process[C]//Proceedings of the 31st International Conference on Machine Learning (ICML 2014). Beijing, China: IMLS, 2014: 982-990.

[113] Chen C, Ding N, Carin L. On the convergence of stochastic gradient MCMC algorithms with high-order integrators[C]//Advances in Neural Information Processing Systems. Montréal, Canada: NIPS Foundation, 2015: 2269-2277.

[114] Teh Y W, Thiery A H, Vollmer S J. Consistency and fluctuations for stochastic gradient Langevin dynamics[J]. The Journal of Machine Learning Research, 2016, 17(1):193-225.

[115] Chen T, Fox E, Guestrin C. Stochastic gradient Hamiltonian Monte Carlo[C]//Proceedings of the 31st International Conference on Machine Learning (ICML 2014). Beijing, China: IMLS, 2014: 1683-1691.

[116] Betancourt M. The fundamental incompatibility of scalable Hamiltonian Monte Carlo and naive data subsampling[C]//Proceedings of the 32nd International Conference on Machine Learning (ICML 2015). Lille, France: IMLS, 2015: 533-540.

[117] Gan Z, Chen C, Henao R, et al. Scalable deep Poisson factor analysis for topic modeling[C]// Proceedings of the 32nd International Conference on Machine Learning (ICML 2015). Lille, France: IMLS, 2015.

[118] Zhang Y, Chen C, Gan Z, et al. Stochastic gradient monomial Gamma sampler[J]. arXiv preprint arXiv:1706.01498, 2017.

[119] Ding N, Fang Y, Babbush R, et al. Bayesian sampling using stochastic gradient thermostats[C]// Advances in Neural Information Processing Systems. Montréal, Canada: NIPS Foundation, 2014: 3203-3211.

[120] Ma Y A, Chen T, Fox E. A complete recipe for stochastic gradient MCMC[C]//Advances in Neural Information Processing Systems. Montréal, Canada: NIPS Foundation, 2015: 2899-2907.

[121] Udriste C. Convex functions and optimization methods on Riemannian manifolds: volume 297 [M]. Udriste: Springer Science & Business Media, 1994

[122] Bonnabel S. Stochastic gradient descent on Riemannian manifolds[J]. IEEE Transactions on Automatic Control, 2013, 58(9):2217-2229.

[123] Liu Y, Shang F, Cheng J, et al. Accelerated first-order methods for geodesically convex optimization on Riemannian manifolds[C]//Advances in Neural Information Processing Systems. Long Beach, California USA: NIPS Foundation, 2017: 4875-4884.

[124] Zhang H, Sra S. An estimate sequence for geodesically convex optimization[C]//Proceedings of the 31st Annual Conference on Learning Theory (COLT 2018). Stockholm, Sweden: IMLS, 2018: 1703-1723.

[125] Brubaker M A, Salzmann M, Urtasun R. A family of MCMC methods on implicitly defined manifolds[C]//Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS-12). La Palma, Canary Islands: AISTATS Committee, 2012: 161-172.

[126] Lan S, Zhou B, Shahbaba B. Spherical Hamiltonian Monte Carlo for constrained target distributions[C]//Proceedings of the 31st International Conference on Machine Learning (ICML 2014). Beijing, China: IMLS, 2014: 629-637.

[127] Girolami M, Calderhead B. Riemann manifold Langevin and Hamiltonian Monte Carlo methods [J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2011, 73(2): 123-214.

[128] Lan S, Stathopoulos V, Shahbaba B, et al. Markov chain Monte Carlo from lagrangian dynamics [J]. Journal of Computational and Graphical Statistics, 2015, 24(2):357-378.

[129] Do Carmo M P. Riemannian geometry[M]. Boston: Birkhäuser, 1992

[130] 陈维桓, 萧树铁. 流形上的微积分[M]. 北京: 高等教育出版社, 2003

[131] Abraham R, Marsden J E, Ratiu T. Manifolds, tensor analysis, and applications: volume 75[M]. New York: Springer Science & Business Media, 2012

[132] Nicolaescu L I. Lectures on the geometry of manifolds[M]. Singapore: World Scientific, 2007

[133] Romano G. Continuum mechanics on manifolds[J]. Lecture notes University of Naples Federico II, Naples, Italy, 2007:1-695.

[134] Carroll S M. Spacetime and geometry: An introduction to general relativity[M]. San Francisco: Addison Wesley, 2004

[135] Hopf H, Rinow W. Über den begriff der vollständigen differential geometrischen fläche[J]. Commentarii Mathematici Helvetici, 1931, 3(1):209-225.

[136] Whitney H. The self-intersections of a smooth n-manifold in 2n-space[J]. Annals of Math, 1944, 45(220-446):180.

[137] Persson M. The Whitney embedding theorem[Z]. Umeå, Sweden: Umeå University, 2014.

[138] Nash J. The imbedding problem for Riemannian manifolds[J]. Annals of Mathematics, 1956: 20-63.

[139] Betancourt M, Byrne S, Livingstone S, et al. The geometric foundations of Hamiltonian Monte Carlo[J]. Bernoulli, 2017, 23(4A):2257-2298.

[140] Li C, Chen C, Fan K, et al. High-order stochastic gradient thermostats for Bayesian learning of deep models[J]. arXiv preprint arXiv:1512.07662, 2015.

[141] Goldstein H. Classical mechanics[M]. Delhi: Pearson Education India, 1965

[142] Marsden J E, Ratiu T S. Introduction to mechanics and symmetry: a basic exposition of classical mechanical systems: volume 17[M]. Berlin: Springer Science & Business Media, 2013

[143] Abraham R, Marsden J E, Marsden J E. Foundations of mechanics[M]. Providence, Rhode Island: Benjamin/Cummings Publishing Company Reading, Massachusetts, 1978

[144] Sherman J, Morrison W J. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix[J]. Annals of Mathematical Statistics, 1950, 21(1):124-127.

[145] Ahn S, Korattikara A, Welling M. Bayesian posterior sampling via stochastic gradient Fisher scoring[J/OL]. arXiv preprint arXiv:1206.6380, 2012. http://arxiv.org/abs/1206.6380.

[146] Mardia K V, Jupp P E. Distributions on spheres[J]. Directional Statistics, 2000:159-192.

[147] Du C, Zhu J, Zhang B. Learning deep generative models with doubly stochastic MCMC[J]. arXiv preprint arXiv:1506.04557, 2015.

[148] Zhang A, Zhu J, Zhang B. Sparse online topic models[C]//Proceedings of the 22nd International Conference on World Wide Web. Rio de Janeiro, Brazil: International World Wide Web Conferences Steering Committee, 2013: 1489-1500.

[149] Banerjee A, Dhillon I S, Ghosh J, et al. Clustering on the unit hypersphere using von Mises-Fisher distributions[J]. Journal of Machine Learning Research, 2005, 6:1345-1382.

[150] Gopal S, Yang Y. Von Mises-Fisher clustering models[C]//Proceedings of the 31st International Conference on Machine Learning (ICML 2014). Beijing, China: IMLS, 2014.

[151] Ghosh K, Jammalamadaka R, Tiwari R C. Semiparametric Bayesian techniques for problems in circular data[J]. Journal of Applied Statistics, 2003, 30(2):145-161.

[152] Anh N K, Tam N T, Linh N V. Document clustering using Dirichlet process mixture model of von Mises-Fisher distributions[C]//The 4th International Symposium on Information and Communication Technology, SoICT 2013. Danang, Vietnam: ACM, 2013: 131–138.

[153] Straub J, Chang J, Freifeld O, et al. A Dirichlet process mixture model for spherical data [C]//Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS-15). San Diego, California USA: AISTATS Committee, 2015: 930-938.

[154] Feng Y, Wang D, Liu Q. Learning to draw samples with amortized Stein variational gradient descent[C]//Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI 2017). Sydney, Australia: Association for Uncertainty in Artificial Intelligence, 2017.

[155] Pu Y, Gan Z, Henao R, et al. Stein variational autoencoder[J]. arXiv preprint arXiv:1704.05155, 2017.

[156] Zhang H, Reddi S J, Sra S. Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds[C]//Advances in Neural Information Processing Systems. Barcelona, Spain: NIPS Foundation, 2016: 4592-4600.

[157] Li C, Chen C, Carlson D E, et al. Preconditioned stochastic gradient Langevin dynamics for deep neural networks[C]//The 30th AAAI Conference on Artificial Intelligence (AAAI-16): volume 2. Phoenix, Arizona USA: AAAI press, 2016: 1788-1794.

[158] Gemici M C, Rezende D, Mohamed S. Normalizing flows on Riemannian manifolds[J]. arXiv preprint arXiv:1611.02304, 2016.

[159] Liu Q. Stein variational gradient descent: Theory and applications[J/OL]. 2017. http://approximateinference.org/accepted/Liu2016.pdf.

[160] Frankel T. The geometry of physics: an introduction[M]. Cambridge: Cambridge University Press, 2011

[161] Aronszajn N. Theory of reproducing kernels[J]. Transactions of the American mathematical society, 1950, 68(3):337-404.

[162] Steinwart I, Christmann A. Support vector machines[M]. New York: Springer Science & Business Media, 2008

[163] Micchelli C A, Pontil M. On learning vector-valued functions[J]. Neural computation, 2005, 17(1):177-204.

[164] Liu Q, Lee J D, Jordan M I. A kernelized Stein discrepancy for goodness-of-fit tests[C]// Proceedings of the 33rd International Conference on Machine Learning (ICML 2016). New York, New York USA: IMLS, 2016.

[165] Zhou D X. Derivative reproducing properties for kernel methods in learning theory[J]. Journal of computational and Applied Mathematics, 2008, 220(1):456-463.

[166] Chwialkowski K, Strathmann H, Gretton A. A kernel test of goodness of fit[C]//Proceedings of the 33rd International Conference on Machine Learning (ICML 2016). New York, New York USA: IMLS, 2016: 2606-2615.

[167] Müller A. Integral probability metrics and their generating classes of functions[J]. Advances in Applied Probability, 1997, 29(2):429-443.

[168] Mika S, Ratsch G, Weston J, et al. Fisher discriminant analysis with kernels[C]//Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop. Madison, Wisconsin USA: IEEE, 1999: 41-48.

[169] Zhuo J, Liu C, Shi J, et al. Message passing Stein variational gradient descent[C/OL]//Dy J, Krause A. Proceedings of Machine Learning Research: volume 80 Proceedings of the 35th International Conference on Machine Learning. Stockholmsmässan, Stockholm Sweden: PMLR, 2018: 6018-6027. http://proceedings.mlr.press/v80/zhuo18a.html.

[170] Pu Y, Gan Z, Henao R, et al. VAE learning via Stein variational gradient descent[C]//Advances in Neural Information Processing Systems. Long Beach, California USA: NIPS Foundation, 2017: 4239-4248.

[171] Zhang R, Wen Z, Chen C, et al. Scalable Thompson sampling via optimal transport[C]// Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS-19). Naha, Okinawa Japan: AISTATS Committee, 2019.

[172] Sutskever I, Martens J, Dahl G, et al. On the importance of initialization and momentum in deep learning[C]//Proceedings of the 30th International Conference on Machine Learning (ICML 2013). Atlanta, Georgia USA: IMLS, 2013: 1139-1147.

[173] Wibisono A, Wilson A C, Jordan M I. A variational perspective on accelerated methods in optimization[J]. Proceedings of the National Academy of Sciences, 2016, 113(47):E7351-E7358.

[174] Detommaso G, Cui T, Marzouk Y, et al. A Stein variational Newton method[C]//Advances in Neural Information Processing Systems. Montréal, Canada: NIPS Foundation, 2018: 9187-9197.

[175] Benamou J D, Brenier Y. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem[J]. Numerische Mathematik, 2000, 84(3):375-393.

[176] Erbar M, et al. The heat equation on manifolds as a gradient flow in the Wasserstein space[C]// Annales de l'Institut Henri Poincaré, Probabilités et Statistiques: volume 46. Paris: Institut Henri Poincaré, 2010: 1-23.

[177] Risken H. Fokker-Planck equation[M]//The Fokker-Planck Equation. Berlin: Springer, 1996: 63-95

[178] Kováčik O, Rákosník J. On spaces $L^p(x)$ and $W^{k,p}(x)$[J]. Czechoslovak Mathematical Journal, 1991, 41(4):592-618.

[179] Li Y, Turner R E. Gradient estimators for implicit models[C/OL]//Proceedings of the International Conference on Learning Representations (ICLR 2018). Vancouver, Canada: ICLR Committee, 2018. https://openreview.net/forum?id=SJi9WOeRb.

[180] Pele O, Werman M. Fast and robust earth mover's distances.[C]//Proceedings of the 12th International Conference on Computer Vision (ICCV-09): volume 9. Kyoto, Japan: IEEE, 2009: 460-467.

[181] Cuturi M. Sinkhorn distances: Lightspeed computation of optimal transport[C]//Advances in Neural Information Processing Systems. Lake Tahoe, Nevada USA: NIPS Foundation, 2013: 2292-2300.

[182] Xie Y, Wang X, Wang R, et al. A fast proximal point method for computing Wasserstein distance [J]. arXiv preprint arXiv:1802.04307, 2018.

[183] Lott J. Some geometric calculations on Wasserstein space[J]. Communications in Mathematical Physics, 2008, 277(2):423-437.

[184] Lott J. An intrinsic parallel transport in Wasserstein space[J]. Proceedings of the American Mathematical Society, 2017, 145(12):5329-5340.

[185] Ehlers J, Pirani F, Schild A. The geometry of free fall and light propagation, in the book "General Relativity" (papers in honour of JL Synge), 63–84[M]. Oxford: Clarendon Press, 1972.

[186] Kheyfets A, Miller W A, Newton G A. Schild's ladder parallel transport procedure for an arbitrary connection[J]. International Journal of Theoretical Physics, 2000, 39(12):2891-2898.

[187] Gabay D. Minimizing a differentiable function over a differential manifold[J]. Journal of Optimization Theory and Applications, 1982, 37(2):177-219.

[188] Qi C, Gallivan K A, Absil P A. Riemannian BFGS algorithm with applications[M]//Recent advances in optimization and its applications in engineering. Berlin: Springer, 2010: 183-192

[189] Yuan X, Huang W, Absil P A, et al. A Riemannian limited-memory BFGS algorithm for computing the matrix geometric mean[J]. Procedia Computer Science, 2016, 80:2147-2157.

[190] Dua D, Graff C. UCI machine learning repository[EB/OL]. University of California, Irvine, School of Information and Computer Sciences, 2017. http://archive.ics.uci.edu/ml.

[191] Shi J, Sun S, Zhu J. A spectral approach to gradient estimation for implicit distributions[C]// Proceedings of the 35th International Conference on Machine Learning (ICML 2018). Stockholm, Sweden: IMLS, 2018: 4651-4660.

[192] Chen C, Ding N, Li C, et al. Stochastic gradient MCMC with stale gradients[C]//Advances in Neural Information Processing Systems. Barcelona, Spain: NIPS Foundation, 2016: 2937-2945.

[193] Li C, Chen C, Pu Y, et al. Communication-efficient stochastic gradient MCMC for neural networks[C]//The 33rd AAAI Conference on Artificial Intelligence (AAAI-19). Honolulu, Hawaii USA: AAAI press, 2019.

[194] Langevin P. Sur la théorie du mouvement Brownien[J]. Compt. Rendus, 1908, 146:530-533.

[195] Wibisono A. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem[J]. arXiv preprint arXiv:1802.08089, 2018.

[196] Bernton E. Langevin Monte Carlo and JKO splitting[J]. arXiv preprint arXiv:1802.08671, 2018.

[197] Durmus A, Majewski S, Miasojedow B. Analysis of Langevin Monte Carlo via convex optimization[J]. arXiv preprint arXiv:1802.09188, 2018.

[198] Kondratyev S, Vorotnikov D. Nonlinear Fokker-Planck equations with reaction as gradient flows of the free energy[J]. arXiv preprint arXiv:1706.08957, 2017.

[199] Bruna M, Burger M, Ranetbauer H, et al. Asymptotic gradient flow structures of a nonlinear Fokker-Planck equation[J]. arXiv preprint arXiv:1708.07304, 2017.

[200] Gallego V, Insua D R. Stochastic gradient MCMC with repulsive forces[J]. arXiv preprint arXiv:1812.00071, 2018.

[201] Fernandes R L, Marcut I. Lectures on Poisson geometry[M]. Basel: Springer, 2014

[202] Da Silva A C. Lectures on symplectic geometry: volume 3575[M]. Boston: Springer, 2001

[203] Gangbo W, Kim H K, Pacini T. Differential forms on Wasserstein space and infinite-dimensional Hamiltonian systems[M]. Providence, Rhode Island: American Mathematical Soc., 2010

[204] Ambrosio L, Gangbo W. Hamiltonian ODEs in the Wasserstein space of probability measures [J]. Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences, 2008, 61(1):18-53.

[205] Barbour A D. Stein's method for diffusion approximations[J]. Probability theory and related fields, 1990, 84(3):297-322.

[206] Stein C. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables[C]//Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory. Oakland: The Regents of the University of California, 1972.

[207] Gorham J, Mackey L. Measuring sample quality with Stein's method[C]//Advances in Neural Information Processing Systems. Montréal, Canada: NIPS Foundation, 2015: 226-234.

[208] Gorham J, Duncan A B, Vollmer S J, et al. Measuring sample quality with diffusions[J]. arXiv preprint arXiv:1611.06972, 2016.

[209] Santambrogio F. Euclidean, metric, and Wasserstein gradient flows: an overview[J]. Bulletin of Mathematical Sciences, 2017, 7(1):87-154.

[210] Durmus A, Moulines E. High-dimensional Bayesian inference via the unadjusted Langevin algorithm[J]. arXiv preprint arXiv:1605.01559, 2016.

[211] McCord C, Meyer K, Offin D. Are Hamiltonian flows geodesic flows?[J]. Transactions of the American Mathematical Society, 2003, 355(3):1237-1250.

[212] Cao Z. Realizing continual learning through modeling a learning system as a fiber bundle[J]. arXiv preprint arXiv:1903.03511, 2019.

[213] Cohen T S, Geiger M, Köhler J, et al. Spherical CNNs[J]. arXiv preprint arXiv:1801.10130, 2018.

[214] Coors B, Paul Condurache A, Geiger A. SphereNet: Learning spherical representations for detection and classification in omnidirectional images[C]//Proceedings of the 15th European Conference on Computer Vision (ECCV-18). Munich, Germany: IEEE, 2018: 518-533.

[215] Schonsheck S C, Dong B, Lai R. Parallel transport convolution: A new tool for convolutional neural networks on manifolds[J]. arXiv preprint arXiv:1805.07857, 2018.

[216] Falorsi L, de Haan P, Davidson T R, et al. Explorations in homeomorphic variational auto-encoding[J]. arXiv preprint arXiv:1807.04689, 2018.

[217] Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering[C]//Advances in Neural Information Processing Systems. Barcelona, Spain: NIPS Foundation, 2016: 3844-3852.

[218] Bronstein M M, Bruna J, LeCun Y, et al. Geometric deep learning: going beyond Euclidean data[J]. IEEE Signal Processing Magazine, 2017, 34(4):18-42.

[219] Nickel M, Kiela D. Poincaré embeddings for learning hierarchical representations[C]//Advances in Neural Information Processing Systems. Long Beach, California USA: NIPS Foundation, 2017: 6338-6347.

[220] Hauser M, Ray A. Principles of Riemannian geometry in neural networks[C]//Advances in Neural Information Processing Systems. Long Beach, California USA: NIPS Foundation, 2017: 2807-2816.

致 谢

衷心感谢我的导师朱军教授。朱老师是我的第一位学术引路人。他纯粹的学术追求，充沛的科研热情和活跃的创新思想都深深引领和激励着我在学术道路上勇敢前行，而他在我遇到瓶颈和低谷时的启发和鼓励也为我带来了莫大的帮助与动力。

感谢张钹院士的言传身教。张院士高屋建瓴的学术思想，严谨求实的治学精神以及高尚谦和的为人风格都让我看到了长远的奋斗榜样。

感谢人工智能国家重点实验室胡晓林老师、李建民老师和陈宁老师在各方面的指导和帮助。

感谢实验室卓靖炜、李崇轩、陈键飞、杜超、石佳欣、许堃等各位同学。作为同行者，我在博士生涯中深受他们的鼓励和支持，感受到了集体的温暖，并且在相互讨论中得到了许多收获与启发。

感谢美国杜克大学 Lawrence Carin 教授、Ricardo Henao 助理教授以及其课题组程鹏宇、陶陈旸、张瑞祎、李春元、王栋、陈立群等各位同学在我访学期间在学习科研上的指导和启发以及在生活上的帮助。

感谢我的父母刘新惠和刘文娟以及各位家人长期的理解和无条件的支持，使我能够专心科研。感谢我的女朋友黄盈同学一直以来的陪伴和鼓励，我们相互支持，一起欢笑，一起成长。

本人在赴美国公派交换期间受到了国家留学基金委的资助，特此致谢。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____ 日　期：_____

个人简历、在学期间发表的学术论文与研究成果

个人简历

1992 年 03 月 31 日出生于河北省邯郸市。

2010 年 7 月考入清华大学物理系基础数理科学专业，2014 年 7 月本科毕业并获得理学学士学位。

2014 年 9 月免试进入清华大学计算机科学与技术系攻读工学博士学位至今。

2017 年 10 月至 2018 年 10 月赴美国杜克大学公派留学一年。

发表的学术论文

[1] **Chang Liu**, Jun Zhu, and Yang Song. Stochastic Gradient Geodesic MCMC Methods. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, pp. 3009–3017, 2016. （CCF A 类会议）

[2] **Chang Liu**, and Jun Zhu. Riemannian Stein Variational Gradient Descent for Bayesian Inference. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI 2018)*, pp. 3627–3634, 2018. （CCF A 类会议）

[3] **Chang Liu**, Jingwei Zhuo, Pengyu Cheng, Ruiyi Zhang, Jun Zhu, and Lawrence Carin. Understanding and Accelerating Particle-Based Variational Inference. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, pp. 4082–4092, 2019. （CCF A 类会议）

[4] **Chang Liu**, Jingwei Zhuo, and Jun Zhu. Understanding MCMC Dynamics as Flows on the Wasserstein Space. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, pp. 4093–4103, 2019. （CCF A 类会议）

[5] Jingwei Zhuo, **Chang Liu**, Jiaxin Shi, Jun Zhu, Ning Chen, and Bo Zhang. Message Passing Stein Variational Gradient Descent. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, pp. 6013–6022, 2018. （CCF A 类会议）

[6] Chenyang Tao, Shuyang Dai, Liqun Chen, Ke Bai, Junya Chen, **Chang Liu**, Ruiyi Zhang, Georgiy Bobashev, and Lawrence Carin. Variational Annealing of GANs: A Langevin Perspective. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, pp. 6176–6185, 2019. （CCF A 类会议）

[7] Pengyu Cheng, **Chang Liu**, Chunyuan Li, Dinghan Shen, Ricardo Henao, and Lawrence Carin. Straight-Through Estimator as Projected Wasserstein Gradient

Flow. In *NeurIPS 2018 Bayesian Deep Learning Workshop*, 2018.

发表的专利

[1] 朱军, **刘畅**, 宋飏. 随机梯度测地线马尔可夫链蒙特卡罗方法及装置: 中国, CN106599909A. (中国专利公开号)