arXiv:2304.09053v1 [math.ST] 18 Apr 2023

用于后验近似的贝叶斯希尔伯特空间

George Wynne

布里斯托大学

摘要

在贝叶斯模型中进行推断，需要使用采样算法从后验分布中抽取样本。随着数据集规模增大，这一过程的计算成本会变得高得难以承受。构造易于计算的后验近似，是解决这一问题的一种常用方法。这就引出了一个问题：什么样的空间适合用于近似贝叶斯后验测度？本文研究贝叶斯希尔伯特空间在后验近似问题中的应用。在函数型数据分析中，贝叶斯希尔伯特空间的研究背景是观测函数为概率密度函数的情形；而它们在计算贝叶斯问题中的应用尚处于起步阶段。本文将概述贝叶斯希尔伯特空间及其与贝叶斯计算的联系，尤其是贝叶斯希尔伯特空间、贝叶斯核心集算法和基于核的距离之间的新颖联系。

目录

1 引言.2

2 背景.3

3 测度差异与贝叶斯希尔伯特空间.9

4 贝叶斯核心集与贝叶斯希尔伯特空间.11

5 最大均值差异与贝叶斯希尔伯特空间.13

6 希尔伯特贝叶斯核心集.17

7 Kullback-Leibler 贝叶斯核心集.20

8 结论.22

9 附录.22

1 引言

本手稿旨在倡导并展示贝叶斯希尔伯特空间作为研究计算贝叶斯问题的工具所具有的实用价值。贝叶斯统计学是一种常见的统计建模范式，它将关于待推断未知量的先验知识与观测数据相结合，由此得到关于该未知量的后验测度。后验测度是我们关注的对象，因为它可用于进行预测和不确定性量化。然而在实践中，要将后验测度用于这些任务，就需要对后验测度运行采样算法。实际面临的问题是，当观测数据集较大时，从后验测度中采样的成本很高。通常，当观测到 N 个数据点时 $O(N$，每次采样算法迭代的成本为Gelman 等人 [2013] 。关于贝叶斯建模方法的更多内容，请参见 Kroese 等人 [2011]；关于贝叶斯方法中的采样，请参见手册。

解决这一问题的方法有很多。主要有两种思路：一是设计算法，避免每次迭代都产生 $O(N)$ 的开销；二是对目标后验分布进行近似，并从该近似分布而非目标后验分布中采样。前一种思路的例子包括采用子采样的采样算法 [Bierkens 等人，2019，Dang 等人， 2019],变分推断 [Blei 等人，2017，Hoffman 等人， 2013]以及分而治之方法[Scott 等人，2016，Srivastava 等人，2015，Vyner 等人， 2023]。本文将重点讨论后一种近似方法。这种方法通常保持先验分布不变，并用比 $O(N)$ 计算成本更低的方式来近似似然函数，从而降低采样时每次迭代的开销。本文接下来将重点讨论这类方法中的贝叶斯核心集 [Huggins 等人，2016，Campbell 和 Broderick，2019]。这种方法根据 M 点和权重构造似然函数的近似，其中 $M \ll N$。将这一近似代入采样算法后，每次迭代的开销为 $O(M)$。这种方法有许多变体 [Campbell 和 Broderick，2018，2019，Campbell and Beronov，2019，Manousakas 等人，2020，Naik 等人， 2022]。

在进行这种近似时，事实上，在进行任何后验近似时，都必须思考：什么样的空间适合用来构造近似？这一选择对后验近似的分析与性能起着至关重要的作用。它既会影响近似后验与目标后验分布之间距离的概念，也会影响如何针对这一距离进行优化以构造近似。本文旨在说明，贝叶斯希尔伯特空间适合用于开展此类分析，并为贝叶斯核心集算法和一般意义上的后验近似提供全新的视角。

贝叶斯希尔伯特空间 [van den Boogart 等人，2011年，van den Boogaart 等人，2014年，Barfoot 和 D'Eleuterio， 2023]是测度空间，其中的加法和数乘概念与贝叶斯定理相一致。这使得它们特别适合用于构造后验近似。贝叶斯希尔伯特空间的概念起源于组成数据分析 [Aitchison，1982，Pawlowsky-Glahn 和 Buccianti， 2011]如今已成为分布数据分析中常用的工具 [Mateu 和 Giraldo，2021年，Petersen 等人， 2022]。这是函数型数据分析的一个子领域，用户观测到的数据集由一组概率密度函数构成，并希望对其进行分析。为了处理概率密度函数的特殊结构，贝叶斯希尔伯特空间采用了不同于典型函数空间的加法和数乘形式。在分布数据分析中，贝叶斯希尔伯特空间的应用已经相当成熟，并有许多实际应用 [Mateu 和 Giraldo， 2021]。另一方面，贝叶斯希尔伯特空间在计算贝叶斯问题中的应用尚处于起步阶段。主要参考文献

关于这一应用的是 Barfoot 和 D'Eleuterio [2023] 的创新性贡献，该贡献阐明了如何将变分推断过程视为贝叶斯希尔伯特空间中的投影。这一联系是通过证明以下事实建立的：变分推断方法可以被视为在贝叶斯希尔伯特空间中优化近似函数的系数。

本文稿的主要技术贡献包括：将贝叶斯希尔伯特空间范数与测度上常用的距离联系起来，建立贝叶斯希尔伯特空间与核方法之间的新联系，并在贝叶斯希尔伯特空间的框架下表述多种贝叶斯核心集算法。

第 2 节将介绍本文稿的背景知识。具体而言，第 2.1 节将介绍贝叶斯希尔伯特空间的数学细节，第 2.2 节将说明贝叶斯希尔伯特空间与贝叶斯定理相一致的含义，第 2.3节将概述与贝叶斯希尔伯特空间相关的现有文献。在背景知识之后，第 3 节将给出一项新结果，将贝叶斯希尔伯特空间中的距离与测度的常见差异联系起来；第 4 节将概述如何在贝叶斯希尔伯特空间中表示贝叶斯核心集；第5 节将建立最大均值差异与贝叶斯核心集之间的新联系；第 6节将利用这一联系，建立最大均值差异算法与希尔伯特核心集算法之间的新联系；第 7节将介绍使用 Kullback-Leibler 散度构建的贝叶斯核心集与贝叶斯希尔伯特空间之间的新联系。结论性意见和未来研究方向将在第 8节中概述。

2 背景

本节将介绍背景信息，以说明后文各节将呈现的新颖结果的动机，并为其提供铺垫。第 2.1 节将概述贝叶斯希尔伯特空间，这是本文的核心主题；第 2.2 节将概述贝叶斯定理；第 2.3节将讨论与贝叶斯希尔伯特空间方法相关的研究方向，重点关注将测度映射到希尔伯特空间的方法。

2.1 贝叶斯希尔伯特空间

本节旨在说明贝叶斯希尔伯特空间的动机并介绍其数学细节。这将包括定义贝叶斯希尔伯特空间的元素、加法与数乘的概念，以及希尔伯特内积。

在开始介绍贝叶斯希尔伯特空间的数学细节之前，了解这些思想的发展过程会有所帮助。贝叶斯希尔伯特空间背后的思想可以追溯到组成数据分析 [Aitchison, 1982]。从数学上讲，成分数据是带有和约束的非负多变量数据，它表示相对信息而非绝对信息。当人们希望分析某些对象的组成时，这类数据便十分适用，因此得名成分数据。例如，可以研究一组化学物质，其中每种物质都由固定数量的其他化学物质组成。此时，可以使用组成数据分析来分析各化学物质的相对比例。关于其他大量示例，请参见Aitchison [1986]。

这类数据有许多独特之处。例如，加法和减法的一致概念并不容易确立。这是因为，通常的向量加法概念会导出一种减法概念，而这种减法可能产生负值；但负值无法表示比例。因此，若尝试对这类数据采用通常的数乘概念，也会遇到问题。

Aitchison 的开创性论文Aitchison [1982] 提出了一个数学框架，用以处理上述成分数据的独特之处。这些工具被称为 Aitchison几何，构成了组成数据分析领域的基石。该几何框架围绕着为成分数据确立特定的相等、加法、数乘和减法概念，并通过对数变换将数据从单纯形映射到一个更适宜的空间。这个数学框架的应用领域十分广泛，涵盖地质学、生态学和微生物组研究等领域[Pawlowsky-Glahn 和 Buccianti，2011年，Filzmoser 等人，2018年，Aitchison，1986，van den Boogaart 和 Tolosana-Delgado，2013] 以及一篇颇具启发性的综述论文，纪念 Aitchison 最初发表论文 40 周年 [Greenacre 等人，2022]。

贝叶斯希尔伯特空间的构想，是将组成数据分析中使用的 Aitchison几何推广到研究对象为测度的情形。理解这一转变最简单的方式，是将成分数据视为直方图，其中组成向量的每个第 n 项代表第 n 个事件发生的概率。接着，关键是将概率密度函数视为直方图的连续极限，然后将 Aitchison几何调整到这一连续极限；在实践中，这意味着从对组成向量各项求有限和，转变为关于概率密度函数的积分。最后，利用拉东-尼科迪姆导数，将概率密度函数推广到测度。

本节余下部分将具体阐述这些观点和直觉。所有技术内容均取自现有文献[Hron 等人，2022年，van den Boogaart 等人，2014年，van den Boogart 等人，2011年，Maier 等人，2021]。

设（Θ，$\mathcal{A}$）为一个可测空间，且 μ 是定义在 Θ上的概率测度。该测度 μ 将作为基准测度。下面的内容可以推广到并非概率测度的有限测度，区别仅在于需要增加归一化项。中的元素将记作 θ。定义 M（μ）为定义在（Θ，$\mathcal{A}$）上的测度集合，这些测度 σ-有限，并且相对于 μ相互绝对连续，也就是说，它们与 μ具有相同的零集。关于贝叶斯计算问题中如何选择 μ 的讨论见第 4节。

定义两个测度 $\eta, \nu \in M(\mu)$ 之间的等价关系 $=_B$：当且仅当存在常数 $c > 0$，使得 $\eta(A) = c\nu(A)\ \forall\, A \in \mathsf{A}$。借助相对于 μ 的拉东-尼科迪姆导数，可以更容易地理解这一等价概念。对于测度$\eta \in M(\mu)$，η 相对于 μ 的拉东-尼科迪姆导数是满足 $\eta(A) = \int_A \mathrm{d}_\eta\, \mathrm{d}_\mu \mathrm{d}\mu\ \forall A \in \mathsf{A}$ 的非负函数 $\mathrm{d}_\eta\, \mathrm{d}_\mu$。根据拉东-尼科迪姆定理，此类函数的存在性得到保证，例如参见 Bogachev [2007, 定理 3.2.2]。该函数在μ-几乎处处意义下唯一。该结果的含义很简单：$M(\mu)$ 中的测度相对于 μ 具有可视为密度的对象，从而使后续结果的解释更加直接。为此，对 $\eta \in M(\mu)$，记 $p_{\eta,\mu} := \mathrm{d}_\eta\, \mathrm{d}_\mu$ 为 η 相对于 μ 的拉东-尼科迪姆导数，以方便记号表示。注意，在 $p_{\mu,\mu} = 1$ 的意义下，$p_{\mu,\mu}$ 是恒等于 1 的函数。于是，当且仅当存在常数 $c > 0$，使得 $p_{\eta,\mu} = cp_{\nu,\mu}$ μ-几乎处处成立时，$\eta =_B \nu$ 成立。事实上，每当这些拉东-尼科迪姆导数被写成相等时，都应将其理解为在 μ-几乎处处意义下相等。

定义 B（μ）为关于 $=_B$ 在 M（μ）中的等价类集合。借助这一等同性概念（如第4节所述，这对于研究贝叶斯问题很有帮助），下一步是定义加法和数乘。对于 η，$\nu \in B$（μ）加法定义为

$$(\eta \oplus \nu)(A) := \int_A p_{\eta,\mu}(\theta) \cdot p_{\nu,\mu}(\theta) \mathrm{d}\mu(\theta)$$

对于每个 $A \in \mathsf{A}$，对于 η，$\nu \in B(\mu)$，数乘定义为

$$(\alpha \odot \eta)(A) := \int_A p_{\eta,\mu}(\theta)^\alpha d\mu(\theta)$$

对于每个 $A \in \mathsf{A}$。加法和数乘的结合有助于确定加法逆元，即对于 $\nu \in B(\mu)$ 定义 $\ominus\nu := (-1 \odot \nu)$，并对于 $\eta,\ \nu \in B(\mu)$ 定义$\eta \ominus \nu := \eta \oplus (\ominus\nu)$。

虽然这些定义乍看之下可能显得反常，但所有运算 $\oplus, \odot, \ominus$ 都可以通过拉东-尼科迪姆导数直接解释。具体而言，对于任意 $\eta, \nu \in B(\mu)$，运算 $\oplus$ 可以写作 $p_{\eta,\mu} \oplus p_{\nu,\mu} := p_{\eta,\mu} \cdot p_{\nu,\mu}$，其中 $\cdot$ 表示标准乘法，因此 $p_{\eta\oplus\nu,\mu} = p_{\eta,\mu} \oplus p_{\nu,\mu}$。数乘 $\odot$ 可以写作 $\alpha \odot p_{\eta,\mu} := p_{\eta,\mu}^{\alpha}$，因此 $p_{\alpha\odot\eta,\mu} = \alpha \odot p_{\eta,\mu}$。最后，$\ominus$ 可以写作 $p_{\eta\ominus\nu,\mu} = p_{\eta,\mu}/p_{\nu,\mu}$，因此 $p_{\eta\ominus\nu,\mu} = p_{\eta,\mu} - p_{\eta,\mu}$。直接定义在拉东-尼科迪姆导数上的这些运算，使得对所有 $\eta \in B(\mu)$，自然的对应关系 η 与 $p_{\eta,\mu}$ 是协调一致的：若记 $\eta =_B p_{\eta,\mu}\ \forall \eta \in B(\mu)$，则$\eta \oplus \nu =_B p_{\eta\oplus\nu,\mu}$ 对所有 $\eta, \nu \in B(\mu)$ 都成立；对于 $\odot, \ominus$，情况也类似。因此，为了简化记号，在下文某些地方可以用 $p_{\eta,\mu}$ 替换 η，因为有时关注的重点是拉东-尼科迪姆导数，而非测度本身。

下一个结果将所有这些运算联系起来，并保证这些运算所得到的是一个有效的实向量空间。

定理 1。*[van den Boogart 等人，2011年，定理 5]* $B(\mu)$ 配备 $\oplus$，$\odot$ 是一个实向量空间，具有 μ 加法零元素。

这个向量空间被称为 贝叶斯线性空间 [van den Boogart 等人，2011] 并为进一步构建贝叶斯希尔伯特空间提供了所需的基本结构。下一步是定义以下集合

$$B^2(\mu) = \left\{ \eta \in B(\mu) \colon \mathbb{E}_\mu \left[\left| \log \frac{\mathrm{d}\eta}{\mathrm{d}\mu} \right|^2 \right] < \infty \right\}. \tag{1}$$

请注意，当且仅当 $\log p_{\eta,\ \mu} \in L^2(\mu)$ 时，$\eta \in B^2(\mu)$，其中 $L^2(\mu)$ 是 Θ 上关于 μ平方可积的函数（等价类）的典型空间。

如前所述，组成数据分析的核心工具是对数变换，用于将成分数据映射到一个良好的空间中。对于当前情形，与之对应的是 中心对数比 （CLR）变换，其定义为 [Egozcue 等人，2006年，van den Boogaart 等人，2014]

$$\Psi_\mu(\eta) = \log p_{\eta,\mu} - \mathbb{E}_\mu[\log p_{\eta,\mu}],$$

其中，期望是针对 $\log p_{\eta,\mu}$ 的输入自变量计算的。这里有几点说明。首先，假设 μ 是有限测度且 $\eta \in B^2(\mu)$，则该期望有限。其次，由于 $p_{\eta,\mu}$ 接受 Θ 中的一个元素作为自变量，CLR $\Psi_\mu(\eta)$ 也同样如此。第三，$\Psi_\mu(\mu) = 0$，即它是恒为零的函数，因为 $p_{\mu,\mu} = 1$，这意味着基准测度（$B^2(\mu)$ 中的加法单位元）被映射为零函数。CLR 映射到 $L_0^2(\mu) = \{f \in L^2(\mu) \colon \mathbb{E}_\mu[f] = 0\}$，即 $L^2(\mu)$ 中均值为零的元素；这与组成数据分析中使用的理想空间相对应。该空间也在信息几何中用作测度流形的切空间，更多讨论见第2.3节。最后，$B^2(\mu)$ 中的可积性条件非常宽松，这意味着 $B^2(\mu)$ 可以包含无限测度。相关讨论见第 8 节。

这些运算 $\oplus, \odot$ 将加法与乘法、数乘与指数运算联系起来；对数保留了这些关系，因此 CLR变换很有用。此外，减去对数的期望值可确保 $=_B$ 定义中的尺度不变性成立。总而言之，这意味着对于任意 $\eta, \nu \in B^2(\mu)$ 和 $\alpha \in \mathbb{R}$，以下线性关系

成立： $\Psi_\mu(\alpha \odot (\eta \oplus \nu)) = \alpha \cdot (\Psi_\mu(\eta) + \Psi_\mu(\nu))$；并且对于任意 $\eta, \nu \in B^2(\mu)$，只要满足 $\eta =_B \nu$，CLR 映射 η, ν 就会映射到相同的输出，即 $\Psi_\mu(\eta) = \Psi_\mu(\nu)$。CLR 映射的反函数就是 exp。为此，任取 $\eta \in B^2(\mu)$，则

$$\exp(\Psi_\mu(\eta)) =_B p_{\eta,\mu} \cdot \exp(-\mathbb{E}_\mu[\log p_{\eta,\mu}]) =_B p_{\eta,\mu}$$

由于 $=_B$ [van den Boogart 等人的尺度不变性， 2011], 其中， η 与p_η之间的自然对应关系μ如定理1之前的说明所述，已在此使用。

利用 CLR 建立 了内积结构。对于 η， $\nu \in B^2(\mu)$，定义

$$\langle \eta, \nu \rangle_{B^2(\mu)} := \langle \Psi_\mu(\eta), \Psi_\mu(\nu) \rangle_{L^2(\mu)} = \int_\Theta \Psi_\mu(\eta)(\theta)\Psi_\mu(\nu)(\theta)\mathrm{d}\mu(\theta). \tag{2}$$

相应的范数为

$$\|\eta - \nu\|_{B^2(\mu)} = \|\Psi_\mu(\eta) - \Psi_\mu(\nu)\|_{L^2(\mu)}. \tag{3}$$

接下来的结果表明， $B^2(\mu)$ 上的内积提供了所需的希尔伯特空间结构。

定理2。*[van den Boogaart 等人，2014年，定理 1]* 映射 Ψ_μ 是 $B^2(\mu)$与 $L_0^2(\mu) = \{f \in L^2(\mu) : \mathbb{E}_\mu[f] = 0\}$之间的等距映射。映射 Ψ_μ 的逆映射为exp，且 $B^2(\mu)$ 是一个希尔伯特空间。

空间 $B^2(\mu)$ 配备内积 (2)，称为贝叶斯希尔伯特空间。定理 2揭示了 $B^2(\mu)$ 的关键结构。这种希尔伯特空间结构将有助于基于函数字典的逼近，以及基于正交投影的优化方法。测度的常见表示通常不具备这种结构，因此贝叶斯希尔伯特空间在测度逼近方面具有巨大潜力。例如，根据 μ 的选择，可以利用正交多项式轻松推导出 $B^2(\mu)$ 的标准正交基。在 μ 为高斯分布并使用厄米特多项式的情形下，Barfoot 和 D'Eleuterio [2023]在高斯变分推断的背景下对此进行了研究。第 8 节将进一步讨论这些基逼近在计算贝叶斯问题中的潜在用途。

2.2 贝叶斯定理与贝叶斯希尔伯特空间

本节旨在说明贝叶斯希尔伯特空间所具有的结构与贝叶斯定理相一致，从而阐明其名称的由来。具体而言，我们将说明，在 $B^2(\mu)$ 中使用的加法和数乘概念，与利用似然函数根据先验信念进行更新以获得后验分布的贝叶斯定理相一致。本节不包含新的技术内容，其讲述内容取自 van den Boogart 等人 [2011], van den Boogaart 等人 [2014]。

回顾组成数据分析，即第 2.1 节所阐述的贝叶斯希尔伯特空间的起源，便可以了解其与贝叶斯定理之间的联系。Aitchison 指出，单纯形作为组成数据分析的规范样本空间，在"统计学的其他领域中也很常见 .....正如贝叶斯公式通过似然函数的扰动影响，将先验概率评估转变为后验概率评估这一操作"[Aitchison， 1986， van den Boogart 等人， 2011]。

现在将这些典故具体化。以下内容摘自van den Boogart 等人 [2011]，有关贝叶斯定理的更多内容请参阅Stuart [2010]。样本空间为 Θ，先验测度为 π_0。假设有 N 个观测值$\{x_n\}_{n=1}^N \subset \mathcal{X}$，即数据空间中的观测值，且它们独立同分布。假设每个观测值的似然为 l 它

以 X 中的一个观测值和 Θ 中的一个参数作为输入。因此，假设这些观测值独立同分布，则所有数据的似然为 $L(\theta) := \prod_{n=1}^{N} l(\theta, x_n)$，或用更简洁的记法表示为 $L = \prod_{n=1}^{N} l_{x_n}$ 其中 $l_{x_n}(\theta) := l(\theta, x_n)$。贝叶斯定理指出，给定先验分布、似然和观测值，关于 Θ 的后验测度 π 满足

$$\frac{\mathrm{d}\pi}{\mathrm{d}\pi_0} = Z^{-1}L, \tag{4}$$

其中，$Z = \mathbb{E}_{\pi_0}[L]$ 是一个常数，称为证据。这里隐含地假设先验分布、似然和数据满足 $Z < \infty$，这是一个非常宽松且常见的假设。

从数学上看，(4)表明，后验分布相对于先验分布的拉东-尼科迪姆导数与似然成正比。因此，利用第 $=_B$ 节中给出的$p_{\pi和\pi_0}$ 的定义2.1

$$p_{\pi,\pi_0} =_B L =_B \bigoplus_{n=1}^{N} l_{x_n}. \tag{5}$$

这意味着，使用贝叶斯希尔伯特空间记法时，贝叶斯定理不过是对似然项求和。这正是贝叶斯希尔伯特空间运算与贝叶斯定理相一致之处。贝叶斯希尔伯特空间的另一个有用性质是将 $=_B$ 定义为相差一个标量常数时相等。这一点很实用，因为在实践中，后验抽样算法不受标量常数的影响，因此，用于后验近似的数学框架也应当对标量常数保持不变。

示例 1。一个反复出现的贝叶斯推断示例是逻辑回归。假设观测到输入输出对 $x_n = \{u_n, y_n\}$，其中 $u_n \in \mathbb{R}^d$ 和 $y_n \in \{0、1\}$。例如，u_n 可能表示总体中第 n个人的医疗信息，而 y_n 表示该人是否患有某种疾病。对于给定u_n 来预测 y_n 的回归问题，逻辑回归是一种模型 [*Gelman 等人，2013年*，第 *3* 章。*7*]。参数空间为 $\Theta = \mathbb{R}^d$，对于参数取值 θ，响应 y_n 建模为伯努利分布($1/1 + e^{-\langle \theta,\ u_n\rangle_{\mathbb{R}^d}}$)，这意味着似然为

$$l_{x_n}(\theta) = \frac{y_n}{1 + e^{-\langle \theta, u_n\rangle_{\mathbb{R}^d}}} + \frac{(1 - y_n)e^{-\langle \theta, u_n\rangle_{\mathbb{R}^d}}}{1 + e^{-\langle \theta, u_n\rangle_{\mathbb{R}^d}}}.$$

θ 的常用先验分布是 $\pi_0 = \mathrm{N}(0、 I_d)$，即标准多元高斯分布。

在假设 $\pi \in B^2(\mu)$ 的条件下，π 的 CLR 为

$$\Psi_\mu(\pi) = \log p_{\pi,\mu} - \mathbb{E}_\mu[\log p_{\pi,\mu}].$$

在附加假设 $\pi_0 \in B^2(\mu)$ 下，利用拉东-尼科迪姆导数的链式法则$p_{\pi,\ \mu} = p_{\pi,\ \pi_0} p_{\pi_0,\ \mu}$ CLR变换可以写成

$$\begin{aligned}\Psi_\mu(\pi) &= \log p_{\pi,\pi_0} - \mathbb{E}_\mu[\log p_{\pi,\pi_0}] + (\log p_{\pi_0,\mu} - \mathbb{E}_\mu[\log p_{\pi_0,\mu}]) \\ &= \mathcal{L} - \mathbb{E}_\mu[\mathcal{L}] + (\log p_{\pi_0,\mu} - \mathbb{E}_\mu[\log p_{\pi_0,\mu}]),\end{aligned} \tag{6}$$

其中 $\mathrm{L} := \log L = \sum_{n=1}^{N} \log\ l_{x_n}$ 为对数似然。注意，若 $\mu = \pi_0$，则括号中的最后一项为零，因为 $p_{\pi_0,\ \pi_0} = 1$，且 CLR 变换等于中心化对数似然函数。

示例 2. 取 $\mu = \pi_0$，接着延续示例 *1*，逻辑回归模型的 *CLR* 为

$$\Psi_{\pi_0}(\pi) = \sum_{n=1}^{N} \log\left(\frac{y_n}{1+e^{-\langle\theta,u_n\rangle_{\mathbb{R}^d}}} + \frac{(1-y_n)e^{-\langle\theta,u_n\rangle_{\mathbb{R}^d}}}{1+e^{-\langle\theta,u_n\rangle_{\mathbb{R}^d}}}\right) - \sum_{n=1}^{N} \mathbb{E}_{\pi_0}\left[\log\left(\frac{y_n}{1+e^{-\langle\theta,u_n\rangle_{\mathbb{R}^d}}} + \frac{(1-y_n)e^{-\langle\theta,u_n\rangle_{\mathbb{R}^d}}}{1+e^{-\langle\theta,u_n\rangle_{\mathbb{R}^d}}}\right)\right].$$

从计算贝叶斯的角度来看，这些表示方式的重要性值得关注。计算贝叶斯统计的主要关注点是后验测度。计算贝叶斯研究者通常希望从后验分布中抽取样本，因此几乎总是需要使用采样算法。采样算法每次迭代的典型成本为 $O(N)$，而通常需要迭代数千次，因此对于大型数据集而言，总成本会变得难以承受。在这种大规模数据场景下，通常会采用以下方法之一：使用能进行低成本近似的采样算法，例如在采样算法运行期间对数据进行子采样；或者尝试对后验分布进行低成本近似，并从该近似中采样。后一种方法的难点在于，通常不会将测度视为处于适用于标准逼近理论和优化理论的空间中。

贝叶斯希尔伯特空间为观察后验分布提供了颇具吸引力的结构。首先，它具有恰当的等同性概念，不受缩放常数影响，这与常见的采样算法颇为相似。其次，希尔伯特空间结构便于运用逼近理论、集中不等式和投影；在本文后续部分，这些工具都将发挥作用，用于构造后验分布的近似。稍后，第 3 节中的定理 3 将表明，贝叶斯希尔伯特空间中的范数可以界定测度之间常用的距离，因此，在贝叶斯希尔伯特空间中进行逼近，能够使这些常用距离最小化。

2.3 相关工作

贝叶斯希尔伯特空间的核心思想是构造一个希尔伯特空间，并将测度映射到其中。希尔伯特空间提供了良好的结构，可用于构造距离、建立近似并进行优化。除了用贝叶斯希尔伯特空间运算来表达贝叶斯规则之外，贝叶斯希尔伯特空间在计算贝叶斯问题中的应用[van den Boogart 等人，2011]目前仍较为有限。这是因为迄今为止，贝叶斯希尔伯特空间的应用主要集中于分布数据分析 [Petersen 等人，2022年，Mateu 和 Giraldo，2021]。关于使用贝叶斯希尔伯特空间进行后验近似，目前唯一的相关贡献来自Barfoot 和 D'Eleuterio [2023]，他们从贝叶斯希尔伯特空间的角度构建了变分推断。

还有其他框架将测度嵌入到其他函数空间中。在信息几何中，与之最相似的主题是指数流形[Pistone 和 Sempi，1995；Cena 和 Pistone，2006；Pistone，2013]。简而言之，这种方法在概率测度空间上构造一个流形，使得对于某个给定的测度（称为参考测度），可以构造一个巴拿赫空间，并找到一个子集，该子集与相对于参考测度绝对连续的测度空间同胚。这种方法与贝叶斯希尔伯特空间之间的联系在于，$B^2(\mu)$ 扮演了参考测度空间的角色，而 $L_0^2(\mu)$则扮演了巴拿赫空间的角色，在此背景下称为切空间。事实上，CLR变换也出现在这种方法中，参见 Pistone 和 Sempi [1995, 方程 23]。在 Pistone 和 Sempi [1995] 中，巴拿赫空间是一个奥利奇空间，它施加了比 $L_0^2(\mu)$ 更强的可积性条件，以确保相应的测度都是有限的；而在 $B^2(\mu)$ 中则不然。Erb 和 Ay [2021]已经建立了信息几何与组成数据分析之间的联系，但将这些联系扩展到贝叶斯希尔伯特空间则留待未来研究。

奥利奇空间相当技术化，因此其他研究致力于构造使用希尔伯特空间的类似流形。其中包括 Newton [2012]，其同样使用 $L_0^2(\mu)$ 作为其希尔伯特空间，但使用了与 CLR 变换略有不同的映射，参见 Newton [2012, 方程 6]。再生核希尔伯特空间方法参见第 5节，由 Fukumizu [2009] 完成，其中巴拿赫空间是一个再生核希尔伯特空间。CLR变换再次出现在 Fukumizu [2009, 定理 18。1]这种方法很有吸引力，因为再生核希尔伯特空间是非常适合研究和操作的空间。然而，一个主要区别在于，相对于给定基准测度绝对连续的测度集合，仅与希尔伯特空间的一个子集同胚。这意味着它不会继承希尔伯特空间的向量空间结构。这与贝叶斯希尔伯特空间具有向量空间结构的情况形成对比。

顾名思义，信息几何中的所有这些方法旨在更好地理解测度空间的几何结构，而非直接对其进行近似。最接近解决近似问题的是 Sriperumbudur 等人 [2017]，他们采用 Fukumizu [2009] 的再生核希尔伯特空间方法，根据分布的样本构造密度估计量；但这与后验近似仍是不同的任务。贝叶斯希尔伯特空间方法与指数流形方法的另一个区别在于，后者中的测度空间与某个函数空间的子集一一对应，而非像前者那样对应于一个子空间，因为这些子空间是流形的切空间。关于指数流形在本文所讨论问题中的潜在应用，更多讨论见第 8 节。

另一种将测度映射到希尔伯特空间的方法是核均值嵌入[Muandet 等人，2017年，Gretton 等人，2012年，Berlinet 和 Thomas-Agnan，2004，Guilbart，1978]。第 5 节对这种方法进行了详细讨论。其思路是使用核来表示测度，然后基于测度的表示而非测度本身进行推理，例如计算距离和执行假设检验。这种方法也已应用于贝叶斯计算 [Fukumizu 等人，2013]。这种方法与贝叶斯希尔伯特空间方法及指数流形方法的关键区别在于，测度到希尔伯特空间的映射无论在理论上还是实践中都难以求逆。因此，核均值嵌入方法难以应用于后验近似。定理 4 将使用核均值嵌入距离，把贝叶斯希尔伯特空间中的距离与后验近似中观测数据经验测度之间的距离联系起来。

3 测度差异与贝叶斯希尔伯特空间

本节旨在证明一项新颖的结果，该结果表明，两个测度之间的贝叶斯希尔伯特空间范数为测度之间的典型差异概念提供了上界。具体而言，定理3 以贝叶斯希尔伯特空间范数的形式给出了 Hellinger、Kullback-Leibler 和 Wasserstein-1 距离的上界。如果要将贝叶斯希尔伯特空间用于后验近似，这类上界非常重要，因为它们表明这种近似是有意义的。这些上界通过简单应用贝叶斯逆问题文献中的一项最新结果获得，该结果涉及后验测度的稳定性 [Sprungk，2020]。之所以能够得到这一结果，是因为贝叶斯逆问题文献中所采用的后验近似误差概念，与贝叶斯希尔伯特空间中的距离相一致。

在陈述结果之前，需要先说明一些记号。首先，Θ 被假设为一个带有度量 d_Θ 的可分完备度量空间，并令 $\mathrm{B}(\Theta)$ 表示 Θ 上有限博雷尔测度的集合，$\mathrm{P}(\Theta) \subset \mathrm{B}(\Theta)$ 表示其中概率测度的子集。在以下三个定义中，η，$\nu \in \mathrm{B}(\Theta)$。

假设 η, ν 都关于某个测度 λ 绝对连续，例如 $\lambda = \eta + \nu$，则这样的 λ 总是存在，海林格距离的平方定义为

$$\mathrm{H}(\eta, \nu)^2 = \frac{1}{2} \int_\Theta \left(\sqrt{\frac{\mathrm{d}\eta}{\mathrm{d}\lambda}} - \sqrt{\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}} \right)^2 \mathrm{d}\lambda, \tag{7}$$

其中该定义与 λ的选择无关。

假设 η 相对于 ν 绝对连续，则 Kullback-Leibler（KL）散度定义为

$$\mathrm{KL}(\eta \parallel \nu) = \int_\Theta \log \frac{\mathrm{d}\eta}{\mathrm{d}\nu} \mathrm{d}\eta,$$

有关 KL 散度的更多讨论，请参见第7节。

令 Lip(1) 表示函数集 $f\colon \Theta \to \mathbb{R}$ 其利普希茨常数小于或等于一，则 Wasserstein-1 距离定义为

$$\mathrm{W}_1(\eta, \nu) = \sup_{f \in \mathrm{Lip}(1)} \left| \int_\Theta f \mathrm{d}\eta - \int_\Theta f \mathrm{d}\nu \right|.$$

对于 $\eta \in$P(Θ)，定义

$$\|\eta\|_{\mathcal{P}^2} = \inf_{\theta_0 \in \Theta} \left(\int_\Theta d_\Theta(\theta, \theta_0)^2 \mathrm{d}\eta(\theta) \right)^{1/2}$$
$$\mathcal{P}^2(\Theta) = \{\eta \in \mathcal{P}(\Theta) \colon \|\eta\|_{\mathcal{B}^2} < \infty\},$$

该量将在 Wasserstein-1 界中出现，用于根据测度和空间上的度量刻画样本空间的大小。最后，对于 $f \in L^2(\mu)$，ess sup$_\mu f$ 表示 f 的本质上确界；如果 $\tilde{f}$ 在 $L^2(\mu)$-等价类中存在 f 这样的$\tilde{}$，使得 $f(\theta) \leq B \; \forall \theta \in \Theta$，则 ess sup$_\mu f \leq B$。有关本质上确界的更详细描述，请参阅 Bogachev [2007,第2节。11]。

定理 3. 设 Θ 是一个完备可分度量空间，且 $\mu \in \mathcal{P}^2(\Theta)$ $\eta, \nu \in \mathcal{B}(\Theta)$. 设 $\eta, \nu \in B^2(\mu)$ 满足

$$\frac{\mathrm{d}\eta}{\mathrm{d}\mu} = Z_\eta^{-1} \exp(\Psi_\mu(\eta)) \qquad Z_\eta = \mathbb{E}_\mu[\exp(\Psi_\mu(\eta))] \tag{8}$$
$$\frac{\mathrm{d}\nu}{\mathrm{d}\mu} = Z_\nu^{-1} \exp(\Psi_\mu(\nu)) \qquad Z_\nu = \mathbb{E}_\mu[\exp(\Psi_\mu(\nu))]. \tag{9}$$

假设 esssup$_\mu \Psi_\mu(\eta)$，esssup$_\mu \Psi_\mu(\nu) \leq B$ 则

$$\mathrm{H}(\eta, \nu) \leq \frac{1}{2} \left(e^B + e^{2B} \right)^{1/2} \|\eta - \nu\|_{B^2(\mu)}$$
$$\mathrm{KL}(\eta \parallel \nu) \leq 2e^B \|\eta - \nu\|_{B^2(\mu)}$$
$$\mathrm{W}_1(\eta, \nu) \leq (e^B + e^{2B}) \|\mu\|_{\mathcal{P}^2} \|\eta - \nu\|_{B^2(\mu)}.$$

证明见附录，是对Sprungk [2020]中结果的直接应用。对于 KL 和 Wasserstein-1 情形，使用 $L^1(\mu)$ 范数的上界同样成立，并且将μ 为概率测度 μ 放宽为有限测度，只需额外付出一个常数的代价，正如证明中所述。

只要上界 B 受到控制，定理3 表明，贝叶斯希尔伯特空间范数收敛于零意味着另外三种距离概念也都收敛于零。最相似的结果是Huggins 等人 [2018, 的命题 6。3] 该命题给出了 Wasserstein-1 和 Wasserstein-2 距离的一个上界，不同之处在于，它使用的是一种 L^2 涉及对数似然函数导数的范数，而不是贝叶斯希尔伯特空间范数。第1节中的推论 4 讨论了逼近后验测度的具体情形。

这一切都隐含地取决于基准测度 μ的选择。该选择不会影响界的不等式左侧，但会通过改变 B 以及范数的值影响右侧。定理 3提供了选择 μ 的指导原则，使 B 尽可能小。

例 3。 继续讨论逻辑回归示例，由于 $l_{x_n} \leq 1$ 对数似然有界，如下式所示：$\log l_{x_n} \leq 0$ 因此 *CLR* 有界，如下式所示：

$$\Psi_{\pi_0}(\pi) \leq -\sum_{n=1}^{N} \mathbb{E}_{\pi_0}\left[\log\left(\frac{y_n}{1+e^{-\langle\theta,u_n\rangle_{\mathbb{R}^d}}} + \frac{(1-y_n)e^{-\langle\theta,u_n\rangle_{\mathbb{R}^d}}}{1+e^{-\langle\theta,u_n\rangle_{\mathbb{R}^d}}}\right)\right],$$

该结果可用作定理 3中的界。

在（ Z_η， Z_ν ）中拉东-尼科迪姆导数的各项记法并不常见，因为它们没有提取出所有标量因子。具体而言，由于 CLR 映射 Ψ_μ 涉及用常数进行中心化，见（6），因此仍存在一个可以吸收到归一化项中的乘法因子。在引入后验分布时，已通过（4）中的 Z 项完成了这一处理。定理 3 中没有这样处理，是为了在右侧保留贝叶斯希尔伯特空间范数的完整表达式。

例 4. 继续讨论逻辑回归示例， Z 项在(4)中为

$$Z = \mathbb{E}_{\pi_0}[L] = \mathbb{E}_{\pi_0}[\exp(\mathcal{L})]$$

其中 $L = \prod_{n=1}^{N} l_{x_n}$， $\mathcal{L}$ =对数 L。与此不同，定理 Z_π 中的项3 为

$$Z_\pi = \mathbb{E}_{\pi_0}[\exp(\Psi_{\pi_0}(\pi))] = \mathbb{E}_{\pi_0}[\exp(\mathcal{L} - \mathbb{E}_{\pi_0}[\mathcal{L}])].$$

通常，定理 3 中使用的这类界（参见 Sprungk [2020], ）假设 L ≤ 0 ，并将 Z^{-1} 视为常数，其中 $Z = \mathbb{E}_{\pi_0}[L]$。注意，L ≤ 0 是一个非常宽松的假设；假设首先 L $\leq C$，其中常数为 C，然后利用 L $- C$即可得到该假设，参见 Sprungk [2020,第 2] 节。求 Z^{-1} 上界的任务与定理 3 中关于 Ψ_{π_0}的上界假设密切相关。为此，注意到根据詹森不等式，$Z^{-1} = \mathbb{E}_{\pi_0}[\exp(L)]^{-1} \leq \exp(-\mathbb{E}_{\pi_0}[L])$。当 L ≤ 0 时，上界 B 只需满足 $-\mathbb{E}_{\pi_0}[L] \leq B$。因此 $Z^{-1} \leq e^B$。这说明，尽管拉东-尼科迪姆导数表示中使用了不同的归一化常数，为确保这些界仍涉及 CLR变换，定理 3 得出的上界常数与贝叶斯逆问题文献中常见结果的上界常数一致，例如 Sprungk [2020,定理 5], ；两者的差异在于詹森不等式中的松弛量。

4 贝叶斯核心集与贝叶斯希尔伯特空间

本节旨在介绍贝叶斯核心集后验近似 [Campbell 和 Broderick，2019，Huggins 等人， 2016] 并阐明它们与贝叶斯希尔伯特空间之间的一种新颖联系

希尔伯特空间。这种联系对本文其余部分至关重要，并将在后续章节中进一步分析。

如上一节所述，对于由 N 项相乘构成的似然函数，采样算法每次迭代的典型成本为 $O(N)$。这是因为典型的采样算法在每次迭代中都必须“触及”似然中的每个数据点。贝叶斯核心集的思路是用 M 项的加权乘积来近似似然函数。由此得到的近似成本为 $O(M$，若 $M \ll N$，则可显著降低成本。首篇关于贝叶斯核心集的论文聚焦于逻辑回归[Huggins 等人， 2016]。此后，该领域取得了广泛发展，并在评估与完整后验分布之间的距离以及用于构建核心集的优化方法方面，催生了多项创新[Campbell 和 Broderick，2018年，2019，Manousakas 等人，2020年，Zhang 等人，2021年，Naik 等人， 2022]。

沿用第2.2节的设定，先验分布为 π_0，观测数据为 $\{x_n\}_{n=1}^N \subset \mathbf{X}$，单个似然函数为 $l_{x_n}(\theta) := l(\theta, x_n)$，完整似然函数为 $L = \prod_{n=1}^N l_{x_n}$。在本文余下部分，贝叶斯核心集指由点 $\{z_m\}_{m=1}^M \subset \mathbf{X}$ 和非负权重 $\{w_m\}_{m=1}^M \subset \mathbb{R}_{\geq 0}$ 构成的集合，其中 $M \in \mathbb{N}$；与该贝叶斯核心集相对应的贝叶斯核心集后验近似是满足以下条件的测度 $\pi_{w,\,z}$

$$\frac{\mathrm{d}\pi_{w,z}}{\mathrm{d}\pi_0} = Z_{w,z}^{-1} L_{w,z}, \tag{10}$$

其中 $L_{w,\,z} = \prod_{m=1}^M l_{z_m}^{w_m}$ 和 $Z_{w,\,z} = \mathbb{E}_{\pi_0}[L_{w,z}]$。使用非负权重，以确保所得后验分布在具有有限测度的意义下是有效的。这种近似保持先验分布不变，并使用似然 $L_{w,\,z}$ 来近似后验分布，而不是使用 L。注意，如果 $M = N$ 且 $w_m = 1$， $z_m = w_m \, \forall m$ 那么 $L = L_{w,\,z}$；如果 $w_m = 0$，则 l_{z_m} 在近似中不起作用。

这种近似背后的思想是，某些观测数据点可能是冗余的，无需纳入；或者可以根据另一个数据点对似然项重新加权来对其进行处理。$\{z_m\}_{m=1}^M$ 不必是观测数据 $\{x_n\}_{n=1}^N$ 的子集这一事实，有助于实现灵活而高效的近似。此处的典型例子是高斯均值推断任务。在这种情形下，后验分布可以用 $M = 1$精确表示，参见 Manousakas 等人 [2020, 第 3]节。$\{z_m\}_{m=1}^M \not\subset \{x_n\}_{n=1}^N$ 的情形在 Manousakas 等人的文献中被称为伪核心集 [2020]。本文不沿用这一术语，因为本文研究的是一般情形：不对 $\{z_m\}_{m=1}^M$ 作任何假设，因此它可能是也可能不是 $\{x_n\}_{n=1}^N$的子集。

$\pi_{w的z}$ CLR变换可通过使用（6）轻松获得。令 μ为基准测度，并假设$\pi_{w,\,z}$， $\pi_0 \in B^2$（μ），则

$$\Psi_\mu(\pi_{w,z}) = \mathcal{L}_{w,z} - \mathbb{E}_\mu\left[\mathcal{L}_{w,z}\right] + (\log p_{\pi_0,\mu} - \mathbb{E}_\mu[\log p_{\pi_0,\mu}]), \tag{11}$$

其中 $\mathbf{L}_{w,\,z} := \log L_{w,\,z} = \sum_{m=1}^M w_m \log l_{z_m}$ 是基于$\{z_m\}_{m=1}^M$的加权对数似然。

贝叶斯核心集近似利用了贝叶斯方法固有的结构：后验分布相对于先验分布的拉东-尼科迪姆导数，除以一个标量常数后，等于似然函数的乘积。这意味着，通过选择 $M = N$ 和 $w_m = 1$， $z_m = w_m \, \forall m$，即可精确恢复完整的后验分布。因此，精确恢复只涉及有限数量的参数选择。这不同于典型的函数近似方法，在这类方法中，目标函数只有通过无限数量的参数选择才能被精确恢复，例如无限基展开。

在贝叶斯核心集场景中， μ的一些简单选择是 $\mu = \pi_0$，即将 μ 设为基于随机抽取的数据点子集得到的后验分布，以及对该后验分布的拉普拉斯近似

后验分布 [Campbell 和 Broderick， 2019]。目前，如何选择 μ以及这对 μ 成为最佳选择意味着什么，仍是一个开放问题。直观而言，人们希望 μ 在后验分布质量集中的区域也具有最大的质量，但这会导致鸡生蛋、蛋生鸡的问题，因为整个任务就是近似后验分布。正如定理3 之后所述，一个指导原则是选择一个 μ ，使其最小化定理 B 中使用的界3。

将 π、$\pi_{w和z}$ 的表达式代入定理3 ，即可了解在这一特定情况下，贝叶斯希尔伯特空间范数如何对概率测度的通常距离概念给出界。贝叶斯核心集的典型场景是拥有 $M = N$ 和 $z_n = x_n$ ，并且只改变权重，其中一些权重为零以提供稀疏性。下一个引理将定理3应用于这一情况。

引理 1。 在定理3的假设下，如果$\mathcal{L} \leq C$ 对于某个常数 C， $M = N$，$z_n = x_n \, \forall n$，且权重 w_n 为非负并满足$\|w\|_\infty \leq W$ ，则 B 可设为 $B = -\max(1, W)\mathbb{E}_\mu[\mathcal{L}] + C$。

证明。 在这些假设下， $\Psi_\mu(\pi) = \mathbf{L} - C - \mathbb{E}_\mu[\mathbf{L}] + C \leq -\mathbb{E}_\mu[\mathbf{L}] + C$ ，对 $\Psi_\mu(\pi_{w,\ z})$ 采用类似的论证。利用关于 w_m， z_m

$$
\begin{aligned}
-\mathbb{E}_\mu[\mathcal{L}_{w,z}] + C = \sum_{n=1}^{N} -w_m \mathbb{E}_\mu[l_{z_m}] + C &\leq \max(1, W) \sum_{n=1}^{N} -\mathbb{E}_\mu[l_{x_n}] + C \\
&= -\max(1, W)\mathbb{E}_\mu[\mathcal{L}] + C,
\end{aligned}
$$

至此，证明完成。 □

引理 1 表明，如果使用观测数据点的子集作为点来构造贝叶斯核心集（这种做法非常常见）[Campbell 和 Broderick，2018年，2019，Naik 等人， 2022], 那么只要权重有界，项 B 就与权重的选择无关。现有的贝叶斯核心集算法通常不会明确地强制要求权重有界，不过这项要求是合理的。这是因为完整似然的范数为$\|\Psi_\mu(\pi)\|_{L^2(\mu)}$, 因此可以预期，每个权重都不会显著大于这一总体范数值。重要性采样核心集方法中有一个权重有界的例外情况 [Huggins 等人， 2016]。

有了近似方案（10），并且借助推论1 和定理 3，能够用贝叶斯希尔伯特空间范数来界定常见的差异度量，接下来要解决的问题就是如何选择贝叶斯核心集近似中的权重和点。过去五年左右，文献中涌现了许多方法，后续各节将从贝叶斯希尔伯特空间的视角来审视这些方法。这个过程的第一步将在下一节展开，该节将贝叶斯希尔伯特空间距离与最大均值差异距离建立起一种新颖的联系；后者是一种用于测度的基于核的距离。

5 最大均值差异与贝叶斯希尔伯特空间

本节旨在概述最大均值差异（最大均值差异）与贝叶斯核心集之间的一种新颖联系，方法是将最大均值差异距离与贝叶斯希尔伯特空间距离联系起来。二者通过核预像问题建立联系[Sch¨olkopf and Smola，2002年，第18]章。此联系将在第6 节中用于研究一类名为希尔伯特贝叶斯核心集的贝叶斯核心集方法。本节其余部分将先介绍最大均值差异，再介绍核预像问题，最后阐述其与贝叶斯核心集的联系。

最大均值差异是一种测度之间基于核的差异，在统计机器学习领域已被研究近二十年 [Gretton 等人，2006年、2012年、Muandet 等人， 2017] 以及计算统计学领域 [Teymur 等人，2021年，Dwivedi and Mackey，2021年，Mak and Joseph， 2018], 其源头可追溯至 20 世纪 70 年代，当时它在更抽象的数理统计学领域以另一名称出现[Guil-bart， 1978]。最大均值差异兼具理论与实践优势。理论优势包括其可优雅地表示为再生核希尔伯特空间（再生核希尔伯特空间）中的形式（下文将作说明），从而便于将最大均值差异改写为多种有用形式。最大均值差异的主要实践优势在于，只需对来自相关两个测度的样本计算核函数值，便可轻松估计。

要开始介绍最大均值差异，首先需要从核说起。核 $k\colon \mathrm{X} \times \mathrm{X} \to \mathbb{R}$ 是一个对称的正定函数。核的作用是度量其输入之间的相似性。构造核的一种简单方法是取任意函数 $\phi\colon \mathrm{X} \to H$，称为特征映射，其中 H 是某个实希尔伯特空间，称为特征空间，然后令$k(x, y) = \mathrm{h}\phi(x)\ \phi(y)\mathrm{i}_H$。在这种情况下，核通过利用特征空间中的内积度量特征 $\phi(x)$ 和 $\phi(y)$ 的相似性，来度量 x 和 y 之间的相似性。

对于任意核 k，都存在一个称为再生核希尔伯特空间的函数空间，它由该核唯一确定，并具有与核相关的特定性质。特别地，记为 H_k 的再生核希尔伯特空间（RKHS），是唯一一个由 X 到 $\mathbb{R}$ 的函数构成的希尔伯特空间，并满足 $k(x, \cdot) \in \mathrm{H}_k\ \forall\ x \in \mathrm{X}$ 且 $\mathrm{h}f, k(x, \cdot)\mathrm{i}_k = f(x)\ \forall f \in \mathrm{H}_k, x \in \mathrm{X}$，其中 $\mathrm{h}\cdot, \cdot\mathrm{i}_k$ 是 H_k 的内积，参见 Christmann 和 Steinwart [2008, 第4节.2]。后一性质称为再生性质，非常有用，因为对于 H_k 中的函数，它使许多关心的量都能仅用核来表示，而核易于计算。关于核及其再生核希尔伯特空间的更多详情，请参阅 [Christmann 和 Steinwart，2008，Berlinet 和 Thomas-Agnan，2004，Paulsen 和 Raghupathi， 2016]。

在介绍最大均值差异之前，先说明一点符号约定。按照最大均值差异文献中的惯例，拉丁字母用于表示$\mathcal{X}$上的测度和随机变量；在贝叶斯推断问题中，观测数据将取值于这个空间[Sriperumbudur 等人， 2010]。这样便可将它们与 Θ上的测度区分开来；后者是参数空间，使用希腊字母表示，正如第2.1节所采用的记法。

现定义最大均值差异。若需比本文更深入的讨论，请参阅 Sriperumbudur 等人 [2010]。令 $\mathrm{B}_k(\mathrm{X}) = \{P \in \mathrm{B}(\mathrm{X})\colon \mathbb{E}_P[\sqrt{k(X, X)}] < \infty\}$，其中在期望中，$X$ 表示取值于 X 且服从 P分布的随机变量，而 $\mathrm{B}(\mathrm{X})$ 是 X 上有限博雷尔测度的集合。注意，经验测度的所有有限和都属于 B_k。对于 $P, Q \in \mathrm{B}_k(\mathrm{X})$，使用核$k$ 定义的最大均值差异为

$$\mathrm{MMD}_k(P, Q) \coloneqq \sup_{\|f\|_k \leq 1} \left|\mathbb{E}_P[f] - \mathbb{E}_Q[f]\right|, \tag{12}$$

其中，上确界取遍再生核希尔伯特空间的单位球。最大均值差异的关键性质在于，它可以改写成不同形式，从而更易于解释和估计。

其中第一种方法涉及核均值嵌入。这是通过核将测度映射到再生核希尔伯特空间中的一种映射。对于核 k，核均值嵌入 $\Phi_k\colon \mathrm{B}_k(\mathrm{X}) \to \mathrm{H}_k$ 定义为

$$\Phi_k(P) \coloneqq \int_{\mathcal{X}} k(x, \cdot)\mathrm{d}P(x).$$

核均值嵌入的理念是在再生核希尔伯特空间中为测度P提供特征展开，这与标准核方法作用于 $\mathbb{R}^d$ 时对数据进行特征展开的方式非常相似。

类似于再生性质表示再生核希尔伯特空间中函数的逐点求值，核均值嵌入表示再生核希尔伯特空间中函数的积分。对于 $f \in \mathrm{H}_k$ 和

$P \in \mathrm{B}_k(\mathrm{X})$

$$\mathbb{E}_P[f] = \int_{\mathcal{X}} f(x)\mathrm{d}P(x) = \int_{\mathcal{X}} \langle f, k(x,\cdot)\rangle_k \mathrm{d}P(x) = \langle f, \int_{\mathcal{X}} k(x,\cdot)\mathrm{d}P(x)\rangle_k = \langle f, \Phi_k(P)\rangle_k, \tag{13}$$

其中，交换积分与内积的顺序，是基于对k 和 P所作的 $P \in \mathcal{B}_k(\mathcal{X})$ [Sriperumbudur 等人，2010年，定理 1]中的可积性假设得出的结果。

核均值嵌入与最大均值差异之间存在如下优雅的联系

$$\begin{aligned}
\mathrm{MMD}_k(P,Q) &= \sup_{\|f\|_k \le 1} |\mathbb{E}_P[f] - \mathbb{E}_Q[f]| \\
&= \sup_{\|f\|_k \le 1} |\langle f, \Phi_k(P) - \Phi_k(Q)\rangle_k| \qquad (14)\\
&= \|\Phi_k(P) - \Phi_k(Q)\|_k, \qquad (15)
\end{aligned}$$

其中，式（14）由式（13）得到，而式（15）则由柯西-施瓦茨定理得到。这表明，最大均值差异等于两个测度的特征展开在再生核希尔伯特空间中的差异。

从式（15）出发，再次利用式（13），可以得到最大均值差异的第三种表示

$$\begin{aligned}
\mathrm{MMD}_k(P,Q)^2 &= \|\Phi_k(P) - \Phi_k(Q)\|_k^2 \\
&= \langle \Phi_k(P), \Phi_k(P)\rangle_k - 2\langle \Phi_k(P), \Phi_k(Q)\rangle_k + \langle \Phi_k(Q), \Phi_k(Q)\rangle_k \qquad (16)\\
&= \mathbb{E}_{P\times P}[k(X,X')] - 2\mathbb{E}_{P\times Q}[k(X,Y)] + \mathbb{E}_{Q\times Q}[k(Y,Y')], \qquad (17)
\end{aligned}$$

其中，式（16）只是展开再生核希尔伯特空间中的范数，而式（17）则是利用式（13）得到的。在上述推导中X服从分布 P，Y服从分布 Q，且 X'，Y，是 X的独立同分布副本，分别对应Y。这个双重期望公式使最大均值差异具有实用性，因为只要给定来自 P，Q [Gretton 等人， 2012]的样本，就能轻松地进行经验估计。

最大均值差异已广泛应用于机器学习和计算统计学，例如双样本检验 [Gretton 等人、2012], 参数推断 [Cherief-Abdellatif 和 Alquier、 2020],训练生成模型 [Li 等人、 2017] 以及分布压缩 [Dwivedi 和 Mackey。2021]其他大量应用和参考文献可见于 Muandet 等人的 [2017, 第 3 节5]。

本节重点讨论的最大均值差异用途是以下原像问题 [Sch¨olkopf 和 Smola，2002年，第 18]章。

问题 1（原像问题）。给定一个核 k，$\{x_n\}_{n=1}^N \subset \mathcal{X}$和 $M \in \mathbb{N}$，求非负权重 $\{w_m\}_{m=1}^M \subset \mathbb{R}_{\ge 0}$ 以及 $\{z_m\}_{m=1}^M \subset \mathcal{X}$，使下式最小

$$\begin{aligned}
\mathrm{MMD}_k(P_{w,z}, P_x)^2 &= \left\| \sum_{m=1}^M w_m k(z_m,\cdot) - \sum_{n=1}^N k(x_n,\cdot) \right\|_k^2 \\
&= \sum_{m,m'=1}^M w_m w_{m'} k(z_m, z_{m'}) - 2\sum_{m=1}^M \sum_{n=1}^N w_m k(z_m, x_n) + \sum_{n,n'=1}^N k(x_n, x_{n'}),
\end{aligned}$$

其中 $P_{w、z} = \sum_{m=1}^M w_m \delta_{z_m}$ 和 $P_x = \sum_{n=1}^N \delta_{x_n}$ 是分别对应于两组数据的经验测度。

原像问题在20世纪90年代和21世纪初的支持向量机背景下得到了深入研究 [Burges，1996，Scholkopf 等人， 1999] 以及核主成分分析[Kwok 和 Tsang，2004，Mika 等人， 1998]。之所以称为原像问题，是因为它旨在寻找权重和点，使其在再生核希尔伯特空间（ $\sum_{n=1}^N k(x_n,\cdot)$）中映射到接近的位置。

这个问题涉及核方法和计算统计学中的许多主题。约简集方法解决了类似的原像问题，旨在为基于核的算法找到稀疏表示 [Burges， 1996]。核主成分分析也解决了类似的原像问题[Sch¨olkopf 等人， 1997]。核驱赶 [Chen 等人，2010年，Bach 等人， 2012] 是一种解决原像问题的方法：通过贪婪地选取点，然后选择权重。主要有两种方法，分别使用均匀权重或线搜索。关于核驱赶的更多内容将在第6节讨论。当$\{z_m\}_{m=1}^M \subset \{x_n\}_{n=1}^N$ 已知时，这种情形称为分布压缩[Dwivedi 和 Mackey， 2021], 或量化 [Teymur 等，2021年，Graf 和 Luschgy， 2000], ，其中权重通常在数据点上保持均匀分布。当用户并非试图近似经验分布，而是使用基于核的方法近似连续分布时，这属于核Stein差异文献的研究范围 [Riabiz 等，2022年，Anastasiou 等， 2023]; 准蒙特卡洛是一个相关领域，旨在为给定测度寻找用于积分近似的离散化表示 [Caflisch，1998年，Dick 等人， 2013]。在统计深度中，h-深度[Wynne 和 Nagy， 2021] 是原像问题的一个实例，其中 $M = 1$ 因此仅使用一个点来表示数据。

原像问题似乎立即与贝叶斯核心集相关，因为二者都涉及从一个数据集出发，并寻找权重和点。微妙之处在于，贝叶斯核心集的有效性是通过其相应后验近似的质量来衡量的。这意味着，位于$\mathcal{X}$ 中的权重和点的质量概念，是通过 Θ上的测度之间的距离来表达的。这与对原像问题及上述相关问题的大多数研究形成对比。造成这种差异的原因在于，人们通常使用的核所涉及的表达式，仅依据$\mathcal{X}$的几何结构来衡量其输入之间的距离。将原像问题（1）与贝叶斯核心集联系起来的关键，是使用一种将数据映射到贝叶斯希尔伯特空间的核。这样便可以根据相应的后验分布，对$\mathcal{X}$ 中的点进行比较。

这是通过使用形如 $k(x, y) = \mathrm{h}\phi_\mu(x)\ \phi_\mu(y)\mathrm{i}_H$ 的核来实现的；本节开头讨论了此类核。沿用第2.2节中使用的贝叶斯定理设置和记号。以点 $x \in \mathrm{X}$ 为基准的似然函数为 l_x。对于某个基准测度 μ ，特征映射为 $\phi_\mu(x) = \log l_x - \mathbb{E}_\mu[\log\ l_x]$ ，特征空间为 $H = L^2(\mu)$，核定义为

$$k(x, y) = \langle \phi_\mu(x), \phi_\mu(y) \rangle_{L^2(\mu)}. \tag{18}$$

该核通过比较以这些点为基础的对数似然的中心化版本之间的相似性，来衡量 X 中两个点之间的相似性。这为数据空间与样本空间之间提供了关键联系。

定理 4。设 X 是一个配备其 *Borel* σ-代数的度量空间。设 μ 是可测空间 Θ 上的有限测度，且 l 是一个似然函数，使得 $\phi_\mu\colon X \to L^2(\mu)$， $\phi_\mu(x) = \log\ l_x - \mathbb{E}_\mu[\log l_x]$ 有明确定义且可测。令 $\pi_0 \in B^2(\mu)$ 和 $\pi \in B^2(\mu)$ 分别为以后验分布，其先验分布为π_0，似然函数为 l，观测值为 $\{x_n\}_{n=1}^N \subset X$。令 $M \in \mathbb{N}$、$\{w_m\}_{m=1}^M \subset \mathbb{R}$ 和 $\{z_m\}_{m=1}^M \subset X$。若 k是核（18），则

$$\mathrm{MMD}_k(P_{w,z}, P_x) = \|\pi_{w,z} - \pi\|_{B^2(\mu)}\,, \tag{19}$$

其中 $\pi_{w,\ z}$ 是基于加权似然得到的后验分布 (10)。

证明。 这些假设确保 k 定义良好且可测，并且在 X 上的经验测度

是可测的。

$$\mathrm{MMD}_k(P_{w,z}, P_x)^2 = \left\| \sum_{m=1}^{M} w_m k(z_m, \cdot) - \sum_{n=1}^{N} k(x_n, \cdot) \right\|_k^2 \tag{20}$$

$$= \sum_{m,m'=1}^{M} w_m w_{m'} k(z_m, z_{m'}) - 2 \sum_{m=1}^{M} \sum_{n=1}^{N} w_m k(z_m, x_n) + \sum_{n,n'=1}^{N} k(x_n, x_{n'}) \tag{21}$$

$$= \left\| \sum_{m=1}^{M} \phi_\mu(z_m) - \sum_{n=1}^{N} \phi_\mu(x_n) \right\|_{L^2(\mu)}^2 \tag{22}$$

$$= \|\Psi_\mu(\pi_{w,z}) - \Psi_\mu(\pi)\|_{L^2}^2 (\mu) \tag{23}$$

$$= \|\pi_{w,z} - \pi\|_{B^2(\mu)}^2 , \tag{24}$$

其中（20）根据（15），（21）根据再生性质，（22）根据 k的定义，（23）根据近似后验（11）和完整后验（6）的 CLR变换表达式，其中各表达式方括号中的项相互抵消，最后（24）根据贝叶斯希尔伯特空间范数（3）的定义。

□

定理 4 表明，在使用核（18）时，原像问题，即问题 1，等价于最小化贝叶斯核心集后验与目标后验分布之间的贝叶斯希尔伯特空间距离。该结果的一个直接推论是，可以将最大均值差异的界转换为贝叶斯希尔伯特空间距离的界，因此，根据定理 3，还可以得到测度之间常用距离的界。下面的推论给出了一个简单示例，讨论如何通过使用基于观测数据子集的似然来近似后验，其中该子集通过无放回均匀抽样获得。这种随机均匀无放回抽样常被用作评估近似后验方法的基本基准 [Campbell 和 Broderick，2019]。

推论 1。在与定理 4相同的假设下，令 $M \leq N$，并令 π_M 为满足 $p_{\pi_M,\ \pi_0} = B(N/M) \odot \bigoplus_{m=1}^{M} l_{z_m}$ 的近似后验分布，其中 $\{z_m\}_{m=1}^{M}$ 是从$\{x_n\}_{n=1}^{N}$中随机均匀地无放回抽取的子样本。假设 $\|\phi_\mu(x_n)\|_{L^2(\mu)} \leq \gamma \ \forall\, 1 \leq n \leq N$，则其成立的概率至少为 $1 - \delta$

$$\|\pi - \pi_M\|_{B^2(\mu)} \leq \sqrt{\frac{8\gamma^2 N(N - M + 1)}{M}} \sqrt{2 \log\left(\frac{2}{\delta}\right)}.$$

该证明直接应用了Schneider [2016]的核均值嵌入集中不等式。该界与贝叶斯核心集的其他集中不等式类似，例如 Campbell 和 Broderick [2019,定理 4。1]。不同之处在于，推论 1 关注简单随机子样本的情形。由于贝叶斯希尔伯特空间具有希尔伯特空间结构，像 Campbell 和 Broderick [2019, 定理 4。1] 中采用的重要性抽样等更复杂的子抽样方法，也可以轻松应用于贝叶斯希尔伯特空间的情形。

下一节将再次利用最大均值差异与贝叶斯希尔伯特空间之间的联系，阐述贝叶斯核心集算法与用于解决原像问题的基于核的方法之间的一种新颖联系。

6 希尔伯特贝叶斯核心集

本节旨在揭示希尔伯特贝叶斯核心集算法与贝叶斯希尔伯特空间之间的新颖联系。前者是指围绕以下做法构建的贝叶斯核心集算法：使用

希尔伯特范数和内积来定义目标后验分布与其近似之间的距离概念。本节主要将说明，Frank-Wolfe 贝叶斯核心集算法 [Campbell 和 Broderick， 2019] 等价于 Frank-Wolfe 核函数驱赶 [Chen 等人，2010年，Bach 等人， 2012] 在使用核函数（18）时成立。这种新颖的关系将被证明是定理4的推论。另一个重点是迭代硬阈值核心集算法[Zhang 等人， 2021] 并将其与最大均值差异下降方法联系起来。

6.1 Frank-Wolfe 贝叶斯核心集

Frank-Wolfe 贝叶斯核心集算法和 Frank-Wolfe 核函数驱赶的目标本质上都是使用候选集来近似希尔伯特空间中的一个元素。正如其名称所示，在这两种情况下，都会使用 Frank-Wolfe 优化方法来寻找集合上的最佳近似。首先，仅以建立所需联系所需的一般性，概述相关的近似问题和 Frank-Wolfe 算法的解。如需进一步了解该方法，请参阅 Clarkson [2010]。

对于某个希尔伯特空间 H，固定某个近似目标 $f \in H$，并令 $\{g_n\}_{n=1}^N \subset H$ 为一组将用于近似该目标的元素。令 $\{\sigma_n\}_{n=1}^N \subset \mathbb{R}_{>0}$ 为一组正实数，设定 $\sigma = \sum_{n=1}^N \sigma_n$，并定义多面体 $\mathsf{W} = \{w \in \mathbb{R}^N \colon w \geq 0, \ \sum_{n=1}^N w_n \sigma_n \leq \sigma\}$，注意到它的顶点为 $v_n = (\sigma/\sigma_n)\mathbf{1}_n$，其中 $\mathbf{1}_n$ 是除第 n 个条目为 1 外其余条目均为 0 的独热向量。目标是找到

$$\arg\min_{w \in \mathcal{W}} \frac{1}{2} \|f - g_w\|_H^2, \tag{25}$$

其中 $g_w = \sum_{n=1}^N w_n g_n$。求解该问题的 Frank-Wolfe 方法涉及迭代的条件梯度更新。它会考察当前近似的残差，找到与其最对齐的方向，然后沿该方向进行线搜索。具体而言，在第 t 次迭代中，按如下方式更新当前估计 $u_t \in \mathsf{W}$

$$\begin{aligned} \overline{u}_t &= \arg\max_{v_n} \langle f - g_{u_t}, g_{v_n} \rangle_H \\ u_{t+1} &= (1 - \rho_t) u_t + \rho_t \overline{u}_t, \end{aligned} \tag{26}$$

其中 ρ_t 是利用 $\overline{u}_t$、u_t、f [Bach 等人、 2012] 和 $v_n = (\sigma/\sigma_n)\mathbf{1}_n$ 以闭式计算得到的标量值，且它们是$\mathcal{W}$的顶点，这意味着 $g_{v_n} = (\sigma/\sigma_n) g_n$。在 (26) 中只需对顶点进行优化，因为这等价于对整个多面体$\mathcal{W}$ 进行优化，因为 (26) 中要优化的函数是线性的。这意味着每次迭代只会添加近似字典中的一个成员，因此经过 T 次迭代后，当前解是 T稀疏的。

由 Campbell 和 Broderick [2019]推导出的 Frank-Wolfe 希尔伯特核心集算法符合这一模板。下面给出一个包含对数似然中心化的版本；Campbell 和 Broderick [2018, 2019] 并未采用该方法，但Campbell and Beronov[2019],Zhang 等人 [2021] 后来倡导了这种方法，因为它能确保似然近似具有所需的尺度不变性。设 $H = L^2(\mu)$、$f = \Psi_\mu(\pi)$ $g_n = \phi_\mu(x_n)$ 和 $\sigma_n = \mathbf{k}\phi_\mu(x_n)\mathbf{k}_{L^2(\mu)}$。经过这些替换后，目标变为求解

$$\arg\min_{w \in \mathcal{W}} \frac{1}{2} \left\| \Psi_\mu(\pi) - \sum_{n=1}^N w_n \phi_\mu(x_n) \right\|_{L^2(\mu)}^2 = \arg\min_{w \in \mathcal{W}} \frac{1}{2} \left\| \Psi_\mu(\pi) - \Psi_\mu(\pi_w) \right\|_{L^2(\mu)}^2, \tag{27}$$

其中 π_w 表示权重为 10的贝叶斯核心集近似 w 和 $z_n = x_n \, \forall \, n$。Frank-Wolfe 迭代为

$$\overline{u}_t = \frac{\sigma}{\sigma_{x_n}} \mathbf{1}_n \text{ where } n = \arg\min_{n \in [N]} \left\langle \Psi_\mu(\pi) - g_{u_t}, \frac{\sigma}{\sigma_{x_n}} \phi_\mu(x_n) \right\rangle_{L^2(\mu)} \tag{28}$$
$$u_{t+1} = (1 - \rho_t) u_t + \rho_t \overline{u}_t,$$

其中，更新 $\overline{u}_t$的完整表达式（见（26））为清晰起见已写出，并且 $[N] = \{1, \ldots, N\}$。注意，Campbell 和 Broderick [2019, 公式 4。2] 与（25）相符，而 Campbell 和 Broderick [2019, 公式 4。4] 在将对数似然替换为中心化对数似然后，与（26）相符。由于（27）和（28）是用 CLR 变换表示的，因此 $L^2(\mu)$ 内积可以用 $B^2(\mu)$ 内积表示，见（2）。这表明，Frank-Wolfe 希尔伯特核心集算法可以用贝叶斯希尔伯特空间的语言来表述。

现在将 Frank-Wolfe 核心集算法与 Frank-Wolfe 核函数驱赶联系起来。核函数驱赶是指将希尔伯特空间 H 设为再生核希尔伯特空间，并在求解核预像问题的背景下使用该算法，见问题 1。令 k 为核（18），$H = \mathcal{H}_k$ 的近似目标为 $f = \sum_{n=1}^N k(x_n, \cdot)$，并沿用上文对 g_n, σ_n所作的相同选择。核函数驱赶旨在寻找

$$\arg\min_{w \in \mathcal{W}} \frac{1}{2} \left\| \sum_{n=1}^N k(x_n, \cdot) - \sum_{n=1}^N w_n k(x_n, \cdot) \right\|_k^2, \tag{29}$$

根据定理 4，其等于 (27)。核函数驱赶 Frank-Wolfe 更新为

$$\overline{u}_t = \frac{\sigma}{\sigma_{x_n}} \mathbf{1}_n \text{ where } n = \arg\min_{n \in [N]} \left\langle \sum_{n=1}^N k(x_n, \cdot) - g_{u_t}, \frac{\sigma}{\sigma_{x_n}} k(x_n, \cdot) \right\rangle_k, \tag{30}$$
$$= \arg\min_{n \in [N]} \left\langle \Psi_\mu(\pi) - g_{u_t}, \frac{\sigma}{\sigma_{x_n}} \phi_\mu(x_n) \right\rangle_{L^2(\mu)}, \tag{31}$$
$$u_{t+1} = (1 - \rho_t) u_t + \rho_t \overline{u}_t,$$

其中，(31) 同样由定理 4 得出。

由于优化目标 (29) 与 (27)，以及更新 (30) 与 (28) 等价，因此可以得出结论：使用核 (18) 时，Frank-Wolfe 希尔伯特核心集算法等价于 Frank-Wolfe 核函数驱赶；该核的特征映射使用 CLR 将数据映射到贝叶斯希尔伯特空间中的似然函数。这一对应关系立即引出了一些问题。例如，通常使用核函数驱赶时，并不要求使用 Frank-Wolfe 贝叶斯核心集方法所规定的数据子集，而是选择定义域中任意位置的点；Chen 等人将这些点称为“超样本”Chen 等人 [2010]。这引出了一个问题：在不限制核心集中的点必须是观测数据点子集的情况下，使用核函数驱赶构建贝叶斯核心集的效果如何？Manousakas 等人在使用 *KL* 散度而非希尔伯特范数研究核心集构建时，将这种方法称为 “伪核心集”，见 Manousakas 等人 [2020]。

6.2 迭代硬阈值贝叶斯核心集

迭代硬阈值（IHT）核心集算法 [Zhang 等人， 2021] 提出使用梯度下降求解贝叶斯核心集的权重，然后对权重进行阈值处理，以确保它们

为非负且稀疏。更具体地说，所考虑的问题是

$$\arg\min_{w\in\mathcal{W}_M} \frac{1}{2}\left\|\Psi_\mu(\pi) - \sum_{n=1}^{N} w_n \phi_\mu(x_n)\right\|^2_{L^2(\mu)}, \tag{32}$$

其中 $\mathcal{W}_M = \{w \in \mathbb{R}^N \colon\ w \geq 0,\ \|w\|_0 \leq M\}$ 是由 $\mathbb{R}^N$ 中权重非负且非零值至多为 M 的向量构成的集合。参数 M 决定了逼近的稀疏程度。IHT 技术 [Zhang 等人， 2021]通过在每次迭代中对 w 执行梯度下降，然后将权重投影到 $\mathcal{W}_M$，从而维持权重选择的有效性和稀疏性。

根据定理 4 可知，(32) 等价于寻找一个稀疏、非负的权重解，以解决核预像问题，即问题 1。梯度下降方法是最早用于解决核预像问题的方法之一 [Burges， 1996]，尽管由于当时的计算方面的考量，这些方法并未得到深入研究。IHT 方法利用加速优化方法的进展，高效地找到能使后验分布得到良好逼近的权重选择。

通过认识到 IHT 方法实际上是在优化最大均值差异判据，可以进一步深入探讨原像问题与 IHT 方法之间的联系。因此，IHT 可视为对与所选点和权重相对应的经验测度执行最大均值差异下降。最大均值差异下降是一种经验测度逼近方法，它利用梯度最小化最大均值差异目标 [Arbel 等人， 2019]。还有一些相关方法采用其他基于核的不一致性[Korba 等人，2021年，Xu 等人， 2022]。这引出了一个问题：在使用核（18）构造贝叶斯核心集时，这些其他基于核的下降方法表现如何?

7 Kullback-Leibler 贝叶斯核心集

本节旨在建立 Kullback-Leibler（KL）贝叶斯核心集算法与贝叶斯希尔伯特空间之间的新颖联系。前者是使用 Kullback-Leibler 散度来衡量目标后验分布与贝叶斯核心集后验之间差异的贝叶斯核心集算法。本节将重点说明，拟牛顿 KL 贝叶斯核心集算法 [Naik 等人， 2022] 除超参数选择外，与贝叶斯希尔伯特空间中的变分推断等价。这一新颖关系将被证明是以下事实的结果：变分推断技术可以用贝叶斯希尔伯特空间来描述，而这一描述由Barfoot 和 D'Eleuterio [2023] 给出。

KL 散度，见第 2.1 节，是变分推断中经常使用的一种散度[Blei 等人， 2017]。这是一种测度逼近方法，在用户选定的一组测度（称为变分族）上最小化 KL 散度。变分族通常使用给定分布的参数进行参数化。例如，变分族可以是高斯混合，而变分参数则是各个高斯分布的混合权重、均值和协方差。寻找变分族内最佳逼近的优化通常涉及梯度下降更新。这不需要来自目标测度的样本；当目标测度的采样成本很高时，例如目标测度是观测大量数据点后的后验分布时，这是一个至关重要的性质。

Barfoot 和 D'Eleuterio 以贝叶斯希尔伯特空间为基础阐述了变分推断，这一视角现概述如下。设[2023]令 $B^2(\mu)$ 为一个贝叶斯希尔伯特空间， $\nu \in B^2(\mu)$ 为

所要近似的目标测度，而 $\{b_n\}_{n=1}^N \subset B^2(\mu)$ 是构成近似字典的一组 N 元素。变分族为 $\{\oplus_{n=1}^N w_n \odot b_n : w \in \mathbb{R}^N\}$，相应的优化问题为

$$\arg\min_{w\in\mathbb{R}^N} \mathrm{KL}\left(\bigoplus_{n=1}^N w_n \odot b_n \parallel \nu\right), \tag{33}$$

其中符号 $\oplus$, $\odot$ 定义见第 2.1节。

Barfoot 和 D'Eleuterio 的关键洞见是 [2023]，相对于 w 使用拟牛顿梯度下降方法求解该问题，等价于在贝叶斯希尔伯特空间中进行迭代投影。具体而言，Barfoot 和 D'Eleuterio [2023, 在方程 41] 中给出了拟牛顿更新

$$w_{t+1} = w_t + G_{\nu_t}(\mathbf{b})^{-1}\langle \mathbf{b}, \nu \ominus \nu_t\rangle_{B^2(\nu_t)}, \tag{34}$$

其中，ν_t 表示第 t 次迭代时使用权重 w_t得到的近似，$G_{\nu_t}(\mathbf{b})$ 是由近似字典 $\{b_n\}_{n=1}^N$ 构成的 $N \times N$ Gram矩阵，其 (n, m) 项为 $\mathrm{h}b_n, b_m\mathrm{i}_{B^2(\nu_t)}$，而 $\mathrm{h}\mathbf{b}\ \nu \ominus \nu_t\mathrm{i}_{B^2(\nu_t)}$ 是$\mathbb{R}^N$ 向量，其 n项为 $\mathrm{h}b_n, \nu \ominus \nu_t\mathrm{i}_{B^2(\nu_t)}$。有关 $\ominus$ 的定义以及贝叶斯希尔伯特空间内积，请参见第2.1节。有人指出，该拟牛顿梯度更新可以写成

$$w_{t+1} = G_{\nu_t}(\mathbf{b})^{-1}\langle \mathbf{b}, \nu\rangle_{B^2(\nu_t)},$$

参见 Barfoot 和 D'Eleuterio [2023, 方程 43], 这意味着在每次迭代中，目标 ν 都会投影到以下对象的张成空间: $\{b_n\}_{n=1}^N$，所依据的是 $B^2(\nu_t)$ 内积。这表明贝叶斯希尔伯特空间的几何结构与变分推断方法之间的联系。

现在来说明 Naik 等人[2022]推导的准牛顿贝叶斯核心集算法与变分推断的贝叶斯希尔伯特空间视角之间的联系。Naik 等人 [2022]提出的方法首先将目标测度设为 π，即后验分布，并选择某个$M \in \mathbb{N}$。然后从 M 中均匀抽取一个大小为 $\{x_n\}_{n=1}^N$ 的子集，对其索引重新排序后记为$\{x_m\}_{m=1}^M$。随后，使用准牛顿迭代优化对数似然项$\{l_{x_m}\}_{m=1}^M$ 的权重，并在每次迭代中通过阈值处理使权重保持非负。

与 Barfoot 和 D'Eleuterio [2023] 的联系通过在（$\nu = \pi$、$N = M$ 和$b_m = l_{x_m}$）中设置 33 得以建立。这意味着优化问题为

$$\arg\min_{w\in\mathbb{R}^M_{\geq 0}} \mathrm{KL}\left(\bigoplus_{m=1}^M w_m \odot b_m \parallel \pi\right) = \arg\min_{w\in\mathbb{R}^M_{\geq 0}} \mathrm{KL}(\pi_{w,x} \parallel \pi),$$

这等同于（33），但额外加入了非负权重约束。

从准牛顿迭代式 [Naik 等人，2022年，方程 13]可以看出，除了超参数选择之外，它与更新式（34）相同，因为 Naik 等人 [2022, 的方程13] 与 B^2（ν_t）内积一致。随后，Naik 等人 [2022] 的方法会对权重进行阈值处理，以确保其非负。

总体而言，这表明准牛顿 KL 核心集方法可以通过贝叶斯希尔伯特空间中的迭代投影来理解。这种视角的交融颇有成效，因为它为准牛顿贝叶斯核心集算法提供了贝叶斯希尔伯特空间几何解释。反过来，Naik 等人 [2022]的数值和理论研究也进一步支持了这一几何视角。鉴于贝叶斯希尔伯特空间在信息几何意义上是切空间 [Fukumizu，2009年，Pistone， 2013] 准牛顿核心集方法 [Naik 等人，2022] 可以视为在切空间中沿某一方向进行投影的迭代更新，随后再

更新切空间的基点位置。这也与自然梯度下降有关，正如Barfoot 和 D'Eleuterio [2023]所指出的。这一视角最初针对Campbell and Beronov [2019] 提出的贪心 KL 核心集方法，而本节的结果表明，借助贝叶斯希尔伯特空间视角，它在更广泛的意义上同样适用。

8 结论

本文介绍了贝叶斯希尔伯特空间，并说明了它们为何适用于后验测度近似。文中还给出了贝叶斯希尔伯特空间范数与常见差异度之间的一种新界，并在后验近似的背景下，揭示了基于核的差异度与贝叶斯希尔伯特空间之间的新联系。本文还将多种贝叶斯核心集方法表述为贝叶斯希尔伯特空间中的方法，从而深入揭示了这些方法的运作原理，并提出了新的解读。

关于贝叶斯希尔伯特空间、贝叶斯核心集和后验近似的未来研究方向不胜枚举。下面概述三个主要方向。首先，定义贝叶斯希尔伯特空间时提到的一个问题是，从某种意义上说，它们的范围过大，因为其中可能包含无限测度。信息几何领域的相关工作 [Fukumizu，2009]定义了一组相关测度，这组测度与希尔伯特空间的一个子集同胚，并且具有更严格的可积性条件；然而，这组测度并非向量空间，而贝叶斯希尔伯特空间是向量空间。因此，一个问题是：如何找到一个具有更严格可积性条件的合适空间，既能确保测度有限，又能保有向量空间结构。其次，正如 van den Boogaart 等人 [2014] 所指出的，利用正交多项式可以直接构造贝叶斯希尔伯特空间的一组基。Barfoot 和 D'Eleuterio [2023]研究了高斯基测度和厄米特多项式这一特定情形。近年来，利用稀疏多项式进行稀疏高维近似的研究取得了诸多进展 [Adcock 等人，2022]。这引出了一个问题：如何将这些方法应用于贝叶斯希尔伯特空间？Huggins 等人 [2017]研究了在贝叶斯计算中使用多项式进行似然近似；若将这一方法与上述进展相结合，或能带来进一步优势。最后，经验测度的分布压缩与量化领域近期出现了许多创新成果 [Riabiz 等人, 2022年, Dwivedi 和 Mackey，2021]。鉴于定理 4 将这一问题与贝叶斯核心集问题联系起来，一个直接浮现的问题是：如何在后验近似中利用这些创新，尤其是在构成似然的点 $\{z_m\}_{m=1}^M$ 并非观测数据子集的情况下。迄今为止，这种伪核心集方法只在 KL 核心集设定下的贝叶斯核心集中得到探索 [Manousakas 等人，2020]，但在分布压缩设定中采用这一方法也很自然 [Chen 等人，2010]。

9 附录

9.1 定理3 的证明

该结果的证明只是对Sprungk [2020]中的定理 5、定理 11 和定理 14 进行调整。所有证明都利用了局部 Lipschitz 连续性，而这种连续性源于涉及 B CLR 变换的界的假设。

回顾一下，相关测度 η，ν 可写为

$$\frac{\mathrm{d}\eta}{\mathrm{d}\mu} = Z_\eta^{-1}\exp(\Psi_\mu(\eta)) \qquad Z_\eta = \mathbb{E}_\mu[\exp(\Psi_\mu(\eta))] \tag{35}$$

$$\frac{\mathrm{d}\nu}{\mathrm{d}\mu} = Z_\nu^{-1}\exp(\Psi_\mu(\nu)) \qquad Z_\nu = \mathbb{E}_\mu[\exp(\Psi_\mu(\nu))]. \tag{36}$$

根据詹森不等式，这意味着 $Z_\eta \geq \exp(\mathbb{E}_\mu[\Psi_\mu(\eta)]) = 1$，由于 $\Psi_\mu(\eta) \in L_0^2(\mu)$，相同的下界也适用于 Z_ν。因此，Z_η^{-1}，$Z_\nu^{-1} \leq 1$。

海林格距离

首先推导海林格距离界。从定义（7）出发，并使用 μ作为基准测度

$$\begin{aligned}
2\mathrm{H}(\eta,\nu)^2 &= \int_\Theta \left(e^{\frac{1}{2}\Psi_\mu(\eta)}Z_\eta^{-\frac{1}{2}} - e^{\frac{1}{2}\Psi_\mu(\nu)}Z_\nu^{-\frac{1}{2}}\right)^2 \mathrm{d}\mu \\
&\leq 2\int_\Theta \left(e^{\frac{1}{2}\Psi_\mu(\eta)}Z_\eta^{-\frac{1}{2}} - e^{\frac{1}{2}\Psi_\mu(\nu)}Z_\eta^{-\frac{1}{2}}\right)^2 \mathrm{d}\mu \\
&\quad + 2\int_\Theta \left(e^{\frac{1}{2}\Psi_\mu(\nu)}Z_\eta^{-\frac{1}{2}} - e^{\frac{1}{2}\Psi_\mu(\nu)}Z_\nu^{-\frac{1}{2}}\right)^2 \mathrm{d}\mu \\
&=: I_1 + I_2,
\end{aligned}$$

其中使用了不等式 $(a-b)^2 \leq 2(a-c)^2 + 2(c-b)^2$，这是三角不等式的结果。现在将对 I_1、I_2 给出上界。

$$\begin{aligned}
I_1 &= 2Z_\eta^{-1}\int_\Theta \left(e^{\frac{1}{2}\Psi_\mu(\eta)} - e^{\frac{1}{2}\Psi_\mu(\nu)}\right)^2 \mathrm{d}\mu \\
&\leq 2e^B \int_\Theta \left(\frac{\Psi_\mu(\eta)}{2} - \frac{\Psi_\mu(\nu)}{2}\right)^2 \mathrm{d}\mu \qquad (37) \\
&= \frac{1}{2}e^B\|\Psi_\mu(\eta) - \Psi_\mu(\nu)\|_{L^2(\mu)}^2 = \frac{1}{2}e^B\|\eta-\nu\|_{B^2(\mu)}^2,
\end{aligned}$$

其中（37）使用了 $|e^x - e^y| \leq e^{\max(x,y)}|x-y| \forall x$、$y$、$\in \mathbb{R}$ 以及假设的上界 B，其作用于 Ψ_μ（η）Ψ_μ（ν）。为界定 I_2，首先注意到根据归一化常数的定义，

$$\begin{aligned}
I_2 &= 2\int_\Theta \left(e^{\frac{1}{2}\Psi_\mu(\nu)}Z_\eta^{-\frac{1}{2}} - e^{\frac{1}{2}\Psi_\mu(\nu)}Z_\nu^{-\frac{1}{2}}\right)^2 \mathrm{d}\mu \\
&= 2Z_\eta^{-1}\left(Z_\eta^{\frac{1}{2}} - Z_\nu^{\frac{1}{2}}\right)^2 \leq 2\left(Z_\eta^{\frac{1}{2}} - Z_\nu^{\frac{1}{2}}\right)^2.
\end{aligned}$$

现在，$|x^{1/2} - y^{1/2}| \leq \frac{1}{2}\min(x,\ y)^{-1/2}|x-y| \,\forall\, x,\ y > 0$ 得到

$$I_2 \leq \frac{1}{\min(Z_\eta, Z_\nu)}|Z_\eta - Z_\nu|^2 \leq \frac{1}{2}|Z_\eta - Z_\nu|^2.$$

最后，

$$|Z_\eta - Z_\nu|^2 \leq \int_\Theta \left|e^{\Psi_\mu(\eta)} - e^{\Psi_\mu(\nu)}\right|^2 \mathrm{d}\mu \leq e^{2B}\|\Psi_\mu(\eta) - \Psi_\mu(\nu)\|_{L^2(\mu)}^2 = e^{2B}\|\eta-\nu\|_{B^2(\mu)}^2,$$

其中采用了与（37）相同的界定技术。综合起来可得

$$\mathrm{H}(\eta,\nu)^2 \leq \frac{1}{4}\left(e^B + e^{2B}\right)\|\eta-\nu\|_{B^2(\mu)}^2,$$

如愿。

KL 散度

KL 散度的情况现已涵盖。首先注意

$$\frac{\mathrm{d}\eta}{\mathrm{d}\nu} = \frac{\mathrm{d}\eta}{\mathrm{d}\mu}\frac{\mathrm{d}\mu}{\mathrm{d}\nu} = Z_\eta^{-1} Z_\nu e^{\Psi_\mu(\eta)-\Psi_\mu(\nu)},$$

因此

$$\begin{aligned}\mathrm{KL}(\eta \,\|\, \nu) = \int_\Theta \log \frac{\mathrm{d}\eta}{\mathrm{d}\nu}\mathrm{d}\eta &\le |\log Z_\eta - \log Z_\nu| + \int_\Theta |\Psi_\mu(\eta) - \Psi_\mu(\nu)|\, Z_\eta^{-1} e^{\Psi_\mu(\eta)}\mathrm{d}\mu \\ &=: I_1 + I_2.\end{aligned}$$

利用 $|\log x - \log y| \le \min(x,y)^{-1}|x-y|\ \forall x, y > 0$ 可得

$$\begin{aligned}I_1 &\le \min(Z_\eta, Z_\nu)^{-1}|Z_\eta - Z_\nu| \\ &\le e^B \|\Psi_\mu(\eta) - \Psi_\mu(\nu)\|_{L^1(\mu)} \\ &\le e^B \|\Psi_\mu(\eta) - \Psi_\mu(\nu)\|_{L^2(\mu)} = e^B \|\eta - \nu\|_{B^2(\mu)},\end{aligned}$$

其中，利用局部Lipschitz界 $|e^x - e^y| \le e^{\max(x,y)}|x-y|\ \forall\, x, y, \in \mathbb{R}$ 得到 $L^1(\mu)$ 范数的上界，其用法与式 (37) 类似。由于 $L^2(\mu)$ 范数上界了 $L^1(\mu)$ 范数，且μ 是一个概率测度。如果仅假设 μ 是一个有限测度，那么上界中还会多出一个 $\mu(\Theta)$ 因子。对 I_2 的界可直接得到

$$\begin{aligned}I_2 = \int_\Theta |\Psi_\mu(\eta) - \Psi_\mu(\nu)|\, Z_\eta^{-1} e^{\Psi_\mu(\eta)}\mathrm{d}\mu &\le e^B \|\Psi_\mu(\eta) - \Psi_\mu(\nu)\|_{L^1(\mu)} \\ &\le e^B \|\Psi_\mu(\eta) - \Psi_\mu(\nu)\|_{L^2(\mu)} = e^B \|\eta - \nu\|_{B^2(\mu)}.\end{aligned}$$

将 I_1, I_2 合并即可得到 KL 散度所需的上界。

Wasserstein-1 距离

最后，我们处理 Wasserstein-1 情形。任取 $\theta_0 \in \Theta$ 则只需平移函数值，即可将上确界限制在满足以下条件的 Lip(1) 函数上：$f(\theta_0) = 0$。对于这样的函数$|f(\theta)| \le d_\Theta(\theta, \theta_0)$。根据 Sprungk [2020, 定理 14],

$$\begin{aligned}\left|\int_\Theta f\mathrm{d}\eta - \int_\Theta f\mathrm{d}\nu\right| &= \left|\int_\Theta f \cdot \left(Z_\eta^{-1} e^{\Psi_\mu(\eta)} - Z_\nu^{-1} e^{\Psi_\mu(\nu)}\right)\mathrm{d}\mu\right| \\ &\le \left|Z_\eta^{-1} - Z_\nu^{-1}\right| \left|\int_\Theta f e^{\Psi_\mu(\eta)}\mathrm{d}\mu\right| + \left|Z_\nu^{-1}\int_\Theta f \cdot \left(e^{\Psi_\mu(\eta)} - e^{\Psi_\mu(\nu)}\right)\mathrm{d}\mu\right| \\ &=: I_1 + I_2,\end{aligned}$$

其中不等式由三角不等式得出。为了界定 I_1 注意

$$\left|Z_\eta^{-1} - Z_\nu^{-1}\right| = \frac{|Z_\eta - Z_\nu|}{Z_\eta Z_\nu} \le |Z_\eta - Z_\nu| \le e^B \|\Psi_\mu(\eta) - \Psi_\mu(\nu)\|_{L^1(\mu)},$$

其中，第一个不等式由归一化常数的下界 1 得出，第二个不等式则再次利用局部 Lipschitz 结论，做法与式（37）中的相同。由此可得

$$I_1 \le \int_\Theta \left|f e^{\Psi_\mu(\eta)}\right| \mathrm{d}\mu \cdot e^B \|\Psi_\mu(\eta) - \Psi_\mu(\nu)\|_{L^1(\mu)} \le e^{2B} \|\mu\|_{\mathcal{P}^1} \|\Psi_\mu(\eta) - \Psi_\mu(\nu)\|_{L^1(\mu)},$$

其中，第二个不等式由 $|f(\theta)|$, $\Psi_\mu(\eta)$ 的界以及 θ_0 的选取具有任意性这一事实得出，因此可得到下确界界 $norm\mu_{\mathcal{P}^1}$。

对 I_2 进行界定在很大程度上类似，

$$I_2 \leq \int_\Theta |d_\Theta(\theta,\theta_0)| \left| e^{\Psi_\mu(\eta)} - e^{\Psi_\mu(\nu)} \right| \mathrm{d}\mu \leq e^B \|\mu\|_{\mathcal{P}^2} \int_\Theta \|\Psi_\mu(\eta) - \Psi_\mu(\nu)\|_{L^2(\mu)},$$

其中，第二个不等式再次由柯西-施瓦茨不等式和局部Lipschitz界得到。总之，利用柯西-施瓦茨不等式，并根据 μ 是一个概率测度这一性质，用 L^2 范数界定 L^1 范数，可得

$$\mathrm{W}_1(\eta,\nu) \leq \left(e^B + e^{2B}\right) \|\mu\|_{\mathcal{P}^2} \|\Psi_\mu(\eta) - \Psi_\mu(\nu)\|_{L^2(\mu)} = \left(e^B + e^{2B}\right) \|\mu\|_{\mathcal{P}^2} \|\eta - \nu\|_{B^2(\mu)}.$$

9.2 推论的证明1

该证明是对Schneider [2016, 定理 2]的简单应用。首先，注意到根据定理4

$$\|\pi_M - \pi\|_{B^2(\mu)} = \mathrm{MMD}_k(P_M, P_N) = \|\Phi_k P_M - \Phi_k P_N\|_k,$$

其中， k 是核（18），$P_M = \frac{N}{M}\sum_{m=1}^M z_m$ 且 $P_N = \sum_{n=1}^N \delta_{x_n}$。因此，需要界定的量与Schneider [2016,定理 2] 中的相同，只是额外缩放了 N，因为所引用的结果涉及归一化经验测度 $\widetilde{P}_M = \frac{1}{N}P_M$， $\widetilde{P}_N = \frac{1}{N}P_N$。这意味着，Schneider [2016,定理 2] 中的界对任意 ε_0

$$\mathbb{P}\left(\left\|\Phi_k\widetilde{P}_M - \Phi_k\widetilde{P}_N\right\|_k \geq \varepsilon_0\right) = \mathbb{P}(\|\Phi_k P_M - \Phi_k P_N\|_k \geq N\varepsilon_0) \leq 2\exp\left(-\frac{M\varepsilon_0^2}{8\gamma^2(1-(M-1)/N)}\right),$$

然后，最后一步只需代入 $\varepsilon = N\varepsilon_0$ 并重新整理，即可得到完全用 δ表示的界。

参考文献

B. Adcock、S. Brugiapaglia 和 C. G. Webster。《高维函数的稀疏多项式逼近》。工业与应用数学学会，2022年。

J. Aitchison。组成数据的统计分析。《皇家统计学会会刊：B辑（方法论）》，44(2):139–177, 1982年。附讨论。

J. Aitchison。《组成数据的统计分析》。Chapman and Hall，1986年。

A. Anastasiou、A. Barp、F.-X. Briol、B. Ebner、R. E. Gaunt、F. Ghaderinezhad、J. Gorham、A. Gret- ton、C. Ley、Q. Liu、L. Mackey、C. J. Oates、G. Reinert 和 Y. Swan。Stein方法与计算统计学相遇：近期若干进展综述。《Statistical Science》，38(1):120 – 139，2023。

M. Arbel、A. Korba、A. SALIM 和 A. Gretton。最大均值差异梯度流。载于神经信息处理系统进展，第 32 卷，2019 年。

F. Bach、S. Lacoste-Julien 和 G. Obozinski。关于集群算法与条件梯度算法之间等价性的研究。载于第*29*届国际机器学习会议论文集，第1355–1362页，2012年。

T. D. Barfoot 和 G. M. T. D'Eleuterio。将变分推断视为贝叶斯希尔伯特空间中的迭代投影：在机器人状态估计中的应用。*Robotica*，41(2):632–667，2023。A. Berlinet 和 C. Thomas-Agnan。概率论与统计学中的再生核希尔伯特空间。Springer US，2004。J. Bierkens、P. Fearnhead 和 G. Roberts。Zig-Zag 过程与用于大数据贝叶斯分析的超高效抽样。统计年刊，47(3):1288–1320，2019。D. M. Blei、A. Kucukelbir 和 J. D. McAuliffe。变分推断：面向统计学家的综述。美国统计学会会刊，112(518):859–877，2017年。V. I. Bogachev。测度论。Springer Berlin Heidelberg，2007。C. J. C. Burges。简化的支持向量决策规则。载于国际机器学习会议，1996。R. E. Caflisch。蒙特卡洛与拟蒙特卡洛方法。*Acta Numerica*，7:1–49，1998。T. Campbell 和 B. Beronov。稀疏变分推断：从零开始构造贝叶斯核心集。载于神经信息处理系统进展，第32卷。Curran Associates, Inc.，2019。T. Campbell 和 T. Broderick。通过贪婪迭代测地上升构造贝叶斯核心集。载于第*35*届国际机器学习会议论文集，第80卷，机器学习研究论文集，第698–706页，2018年。T. Campbell 和 T. Broderick。通过希尔伯特核心集实现自动化可扩展贝叶斯推断。《机器学习研究杂志》，20(15):1–38，2019。A. Cena 和 G. Pistone。指数统计流形。统计数学研究所年刊，59(1):27–56，2006年。Y. Chen、M. Welling 和 A. Smola。通过核集群获得超样本。载于第二十六届人工智能中的不确定性会议论文集，第109–116页，2010年。B.-E. Cherief-Abdellatif 和 P. Alquier。MMD-Bayes：通过最大均值差异实现稳健的贝叶斯估计。载于第*2*届近似贝叶斯推断进展研讨会论文集，第118卷，第1–21页，2020年。A. Christmann 和 I. Steinwart。支持向量机。Springer New York，2008。K. L. Clarkson。核心集、稀疏贪婪近似与 Frank-Wolfe 算法。*ACM* 算法汇刊，6(4)，2010年。K.-D. Dang、M. Quiroz、R. Kohn、M.-N. Tran 和 M. Villani。采用能量守恒子采样的哈密顿蒙特卡洛。《机器学习研究杂志》，20(100):1–31，2019。J. Dick、F. Y. Kuo 和 I. H. Sloan。高维积分：拟蒙特卡洛方法。*Acta Numerica*，22:133–288，2013年。R. Dwivedi 和 L. Mackey。核稀疏化。载于第三十四届学习理论会议论文集，第134卷，第1753–1753页，2021年。

J. J. Egozcue、J. L. D´ıaz-Barrero 和 V. Pawlowsky-Glahn。基于 Aitchison几何的概率密度函数希尔伯特空间。《数学学报（英文版）》，22(4):1175–1182，2006年。I. Erb 和 N. Ay。成分数据分析的信息几何视角。载于《组合数据分析进展》，第 21–43 页。施普林格国际出版社，2021年。P. Filzmoser、K. Hron 和 M. Templ。应用组合数据分析。施普林格国际出版社，2018年。K. Fukumizu。基于再生核希尔伯特空间的指数流形，第 291–306 页。剑桥大学出版社，2009年。K. Fukumizu、L. Song 和 A. Gretton。核贝叶斯规则：基于正定核的贝叶斯推断。《机器学习研究杂志》，14(118):3753–3783，2013年。A. Gelman、J. B. Carlin、H. S. Stern、D. B. Dunson、A. Vehtari 和 D. B. Rubin。贝叶斯数据分析。Chapman and Hall/CRC，2013年。S. Graf 和 H. Luschgy。概率分布量化基础。Springer Berlin Heidelberg，2000年。M. Greenacre、E. Grunsky、J. Bacon-Shone、I. Erb 和 T. Quinn。Aitchison 的组成数据分析四十年：重新审视。*arXiv:2201.05197*，2022年。A. Gretton、K. Borgwardt、M. Rasch、B. Sch¨olkopf 和 A. Smola。两样本问题的核方法。载于神经信息处理系统进展，第 19 卷。麻省理工学院出版社，2006年。A. Gretton、K. M. Borgwardt、M. J. Rasch、B. Sch¨olkopf 和 A. Smola。核两样本检验。《机器学习研究杂志》，13(1):723–773，2012年。C. Guilbart。测度空间上的标量积研究：投影估计。博士论文，里尔科学技术大学，1978年。M. D. Hoffman、D. M. Blei、C. Wang 和 J. Paisley。随机变分推断。《机器学习研究杂志》，14(40):1303–1347，2013年。K. Hron、J. Machalov´a 和 A. Menafoglio。贝叶斯空间中的二元密度：正交分解与样条表示。《统计论文》，2022年。J. Huggins、T. Campbell 和 T. Broderick。可扩展贝叶斯逻辑回归的核心集。载于神经信息处理系统进展，第 29 卷，2016年。J. Huggins、R. P. Adams 和 T. Broderick。PASS-GLM：用于可扩展贝叶斯广义线性模型推断的多项式近似充分统计量。载于神经信息处理系统进展，第 30 卷，2017年。J. H. Huggins、T. Campbell、M. Kasprzak 和 T. Broderick。贝叶斯后验近似误差的实用界：一种非渐近方法。*arXiv:1809.09505*，2018年。

A. Korba、P.-C. Aubin-Frankowski、S. Majewski 和 P. Ablin。核 Stein 差异下降。载于第 *38* 届国际机器学习会议论文集，第 139 卷，第 5719–5730 页，2021年。D. P. Kroese、T. Taimre 和 Z. I. Botev。蒙特卡罗方法手册。Wiley，2011年。J.-Y. Kwok 和 I.-H. Tsang。核方法中的原像问题。*IEEE* 神经网络汇刊，15(6):1517–1525，2004年。C.-L. Li、W.-C. Chang、Y. Cheng、Y. Yang 和 B. Poczos。MMD GAN：深入理解矩匹配网络。载于神经信息处理系统进展，第 30 卷，2017年。E.-M. Maier、A. St¨ocker、B. Fitzenberger 和 S. Greven。贝叶斯希尔伯特空间中的加性密度对标量回归及其在性别经济学中的应用。*arXiv:2110.11771*，2021年。S. Mak 和 V. R. Joseph。支撑点。统计年刊，46(6A):2562–2592，2018年。D. Manousakas、Z. Xu、C. Mascolo 和 T. Campbell。贝叶斯伪核心集。载于神经信息处理系统进展，第 33 卷，第 14950–14960 页，2020年。J. Mateu 和 R. Giraldo，编。地统计函数型数据分析。Wiley，2021年。S. Mika、B. Sch¨olkopf、A. Smola、K.-R. M¨uller、M. Scholz 和 G. R¨atsch。特征空间中的核主成分分析与去噪。载于神经信息处理系统进展，第 11 卷，1998年。K. Muandet、K. Fukumizu、B. Sriperumbudur 和 B. Sch¨olkopf。分布的核均值嵌入：综述与拓展。机器学习基础与趋势®，10(1-2):1–141，2017年。C. Naik、J. Rousseau 和 T. Campbell。通过子采样和拟牛顿优化快速构建贝叶斯核心集。载于神经信息处理系统进展，2022年。N. J. Newton。一个以希尔伯特空间为模型的无限维统计流形。函数分析杂志，263(6):1661–1681，2012年。V. I. Paulsen 和 M. Raghupathi。再生核希尔伯特空间理论导论，计算机科学讲义第 152 卷之剑桥高等数学研究。剑桥大学出版社，剑桥，2016年。V. Pawlowsky-Glahn 和 A. Buccianti，编。组成数据分析。Wiley，2011年。A. Petersen、C. Zhang 和 P. Kokoszka。将概率密度函数建模为数据对象。计量经济学与统计学，21:159–178，2022年。G. Pistone。非参数信息几何。载于计算机科学讲义，第 5–36 页。Springer Berlin Heidelberg，2013年。G. Pistone 和 C. Sempi。在与给定概率测度等价的所有概率测度空间上构造无限维几何结构。统计年刊，23(5):1543–1561，1995年。

M. Riabiz、W. Y. Chen、J. Cockayne、P. Swietach、S. A. Niederer、L. Mackey 和 C. J. Oates。MCMC 输出的最优稀疏化。英国皇家统计学会杂志*B*辑：统计方法学，84(4):1059–1081，2022年。M. Schneider。无放回抽样中核嵌入的概率不等式。载于第*19*届国际人工智能与统计学会议论文集，第51卷，第66–74页，2016年。B. Sch¨olkopf 和 A. J. Smola。核方法学习。麻省理工学院出版社，2002年。B. Sch¨olkopf、A. Smola 和 K.-R. M¨uller。核主成分分析。载于计算机科学讲义，第583–588页。Springer Berlin Heidelberg，1997年。B. Scholkopf、S. Mika、C. Burges、P. Knirsch、K.-R. Muller、G. Ratsch 和 A. Smola。基于核的方法中的输入空间与特征空间。*IEEE* 神经网络汇刊，10(5): 1000–1017，1999年。S. L. Scott、A. W. Blocker、F. V. Bonassi、H. A. Chipman、E. I. George 和 R. E. McCulloch。贝叶斯与大数据：共识蒙特卡罗算法。国际管理科学与工程管理杂志，11(2):78–88，2016年。B. Sprungk。贝叶斯逆问题的局部Lipschitz稳定性。《逆问题》，36(5)，2020年。B. Sriperumbudur、K. Fukumizu、A. Gretton、A. Hyv¨arinen 和 R. Kumar。无限维指数族中的密度估计。《机器学习研究杂志》，18(57):1–59，2017年。B. K. Sriperumbudur、A. Gretton、K. Fukumizu、B. Sch¨olkopf 和 G. R. Lanckriet。希尔伯特空间嵌入与概率测度上的度量。《机器学习研究杂志》，11: 1517–1561，2010年。S. Srivastava、V. Cevher、Q. Dinh 和 D. Dunson。WASP：通过子集后验的重心实现可扩展贝叶斯推断。载于第十八届国际人工智能与统计学会议论文集，第38卷，第912–920页，2015年。A. M. Stuart。逆问题：贝叶斯视角。*Acta Numerica*，19:451–559，2010年。O. Teymur、J. Gorham、M. Riabiz 和 C. Oates。使用最大均值差异对概率测度进行最优量化。载于第*24*届国际人工智能与统计学会议论文集，第130卷，第1027–1035页，2021年。K. G. van den Boogaart 和 R. Tolosana-Delgado。使用 *R* 分析成分数据。Springer Berlin Heidelberg，2013。K. G. van den Boogaart、J. J. Egozcue 和 V. Pawlowsky-Glahn。贝叶斯希尔伯特空间。澳大利亚与新西兰统计学杂志，56(2):171–194，2014年。K.-G. van den Boogart、E. Juan Jos´e 和 P.-G. Vera。贝叶斯线性空间。*SORT*-统计与运筹学汇刊，34(2):201–222，2011年。C. Vyner、C. Nemeth 和 C. Sherlock。SwISS：一种可扩展的马尔可夫链蒙特卡罗分而治之策略。*Stat*，12(1)，2023。

G. Wynne 和 S. Nagy。统计深度与机器学习相遇：函数型数据分析中的核均值嵌入与深度。*arXiv:2105.12778*，2021年。L. Xu、A. Korba 和 D. Slepcev。通过基于相互作用粒子的优化实现测度的精确量化。载于第*39*届国际机器学习会议论文集，第162卷，第24576–24595页，2022年。J. Zhang、R. Khanna、A. Kyrillidis 和 S. Koyejo。贝叶斯核心集：重新审视非凸优化视角。载于第*24*届国际人工智能与统计学会议论文集，第130卷，第2782–2790页，2021年。