神经启发式后验近似（NIPA）

Babak Shahbaba*加州大学欧文分校统计学系 babaks@uci.edu

Zahra Moslemi加州大学欧文分校统计学系 zmoslemi@uci.edu

arXiv:2601.22539v1 [cs.LG] 30 Jan 2026

摘要

人类通过调动多个相互作用的神经系统，高效地从环境中学习。这些系统支持彼此不同又相辅相成的多种控制方式，包括基于模型（目标导向）的规划、无模型（习惯性）反应，以及基于情景记忆的学习。基于模型的机制利用环境的内部模型计算前瞻性的行动价值，从而支持灵活但计算成本高昂的规划；无模型机制缓存价值估计并形成启发式方法，从而实现快速、高效的习惯性反应；基于记忆的机制则能够根据个体经验迅速调整。在这项工作中，我们旨在阐明这种生物学效率背后的计算原理，并将其转化为一种采样算法，通过有效探索后验分布，实现可扩展的贝叶斯推断。具体而言，我们提出的算法由三个部分组成：基于模型的模块利用目标分布进行有引导但计算缓慢的采样；无模型模块利用先前的样本学习参数空间中的模式，从而无需直接计算代价高昂的目标分布，即可快速、反射式地采样；情景控制模块则通过回忆特定的过往事件（即样本）来支持快速采样。我们展示了这种方法能够推进贝叶斯方法的发展，并促进其应用于大规模统计机器学习问题。具体而言，我们将所提出的框架应用于贝叶斯深度学习，重点关注恰当且有原则的不确定性量化。

关键词 贝叶斯推断 ·马尔可夫链蒙特卡罗 ·决策

1 引言

贝叶斯推断为不确定性量化和复杂参数空间的探索提供了一个严谨的框架，但精确的后验推断通常难以处理。因此，在实践中我们通常依赖近似方法，其中最著名的是马尔可夫链蒙特卡罗（MCMC）和变分贝叶斯（VB）。MCMC 构造马尔可夫链，使其渐近收敛于目标分布，因此具有无偏性。尽管简单的采样方法（如 Metropolis 算法）[1], 通常能够有效探索低维分布，但对于复杂的高维模型而言，它们可能效率极低且计算成本高昂。为减少 Metropolis 采样的随机游走行为，哈密顿蒙特卡洛（HMC）[2, 3] 利用梯度信息提出与当前状态相距较远、但仍具有较高接受概率的样本。然而，HMC 及相关方法面临严重的可扩展性挑战，因为在海量数据集上进行梯度计算在计算上难以实现。基于随机梯度的马尔可夫链蒙特卡洛变体 [4–7] 旨在通过数据子采样降低这一成本，但由此产生的噪声可能降低准确性，并妨碍有效探索 [8]。

变分贝叶斯（VB）[9, 10] 提供了一种互补的替代方案：它将推断表述为一个优化问题，通过参数化分布近似后验分布，以牺牲无偏性换取速度和可扩展性。为此，变分贝叶斯方法通过调整变分参数，最小化目标分布与参数化近似分布之间的 Kullback–Leibler（KL）散度。随着深度学习中针对变分目标的高效基于梯度的优化方法不断发展，这种方法也得到进一步普及和推进 [11–13]。与马尔可夫链蒙特卡罗方法相比，变分贝叶斯通常速度更快，但会引入近似

*通讯作者。

偏差。此外，变分贝叶斯方法，尤其是常用的平均场公式，往往会低估后验不确定性 [14]。

这两种常见方法之间的差异，自然促使人们探索旨在结合两种范式优势的混合方法 [15, 16]。设计得当时，这类方法能够更好地平衡计算效率与近似精度，从而有效弥合马尔可夫链蒙特卡罗与变分贝叶斯之间的差距。本文提出的方法通过构建一个通用框架来实现这一目标。该框架借鉴了人类如何通过多个相互作用的神经系统高效学习并引导决策；这些系统支持彼此互补的控制形式：基于模型（目标导向）的规划利用环境的内部模型计算前瞻性的行动价值，从而实现灵活但计算开销较大的决策；无模型（习惯性）控制利用价值估计学习行为模式，支持快速、高效的习惯性反应；基于记忆的机制则能够根据个体经验实现快速适应 [17, 18]。

受这一观点启发，我们提出了一个名为神经启发式后验近似（NIPA）的通用框架，由三个相应的组成部分构成：（1）由目标分布引导的基于模型的模块，可实现高质量但计算速度较慢的采样；（2）无模型模块，通过学习参数空间中的模式实现快速、反射式采样；（3）情景控制模块，通过回忆过往经验（即先前的样本）支持快速的单次采样。本文实现了该框架的一种具体且简化的形式。在大脑中，尽管有证据表明基于模型的控制、无模型控制和情景控制分别与前额叶皮层、纹状体和海马体相关联 [18, 19], ，但这些系统的神经实现十分复杂，依赖于横跨海马体、皮层、纹状体和前额叶区域的分布式网络，而且各系统之间存在大量重叠。

在接下来的章节中，我们首先简要概述计算贝叶斯推断的最新进展。随后，我们将详细介绍所提出的方法，并通过与多种最先进的算法进行比较来评估其性能。具体而言，我们重点关注深度学习模型中的不确定性量化。我们提出的方法构成了一个通用框架，可根据基于模型的模块所采用的具体采样算法以及无模型组件所使用的学习算法，进行多种变体。此外，还可以考虑多种策略，以确定在任意时刻启用哪个模块。在讨论部分，我们将概述若干进一步改进和扩展该框架的潜在方向。

2 相关工作

关于提高贝叶斯推断计算方法效率的文献十分丰富。（例如参见 [3, 15, 20–56]。）如上所述，贝叶斯推断通常需要使用 MCMC算法，从难处理分布中模拟样本。这些算法的整体效率取决于它们收敛到目标分布并生成真正具有底层分布代表性的样本的速度。尽管简单的 Metropolis 算法 [1] 通常能有效探索低维分布，但在复杂的高维分布中效率可能很低：由于移动具有随机游走的特性，连续状态之间可能呈现很强的自相关。因此，由于有效样本量（即经过自相关调整后的样本数量）往往很低，收敛到真实分布的速度非常缓慢。例如，使用哈密顿蒙特卡罗（HMC） [2, 3]可以加快探索。HMC及其变体 [44, 48, 50] 通过沿哈密顿动力学所引导的方向，以步长为 ϵ 进行 L 步移动，减少 Metropolis 算法中的随机游走行为；哈密顿动力学利用梯度信息提出远离当前状态、但仍具有较高接受概率的状态。

这些 MCMC算法的主要瓶颈，尤其是在高维或复杂情境下，是反复计算后验密度、其梯度以及其他相关的几何量或统计量，而这些计算通常涉及整个观测（或训练）数据集。一个自然的问题是，如何有效近似这些量，同时在准确性与计算成本之间取得良好平衡。一种常见方法是子采样[41, 54–58], ，即将计算限制在观测数据的一个子集上。这种方法基于这样一种观点：大数据集包含大量冗余信息，因此只需一个小子集便可获取整体信息。然而，在一般应用中，我们不能简单地使用随机子集，因为随机抽样造成的信息损失会导致不可忽视的准确性损失，进而对计算效率产生显著的负面影响 [8]。因此，在实践中，寻找有效子采样的良好准则和策略颇具挑战性，尽管近期已有一些方法旨在缓解这一问题 [59]。

其他方法利用参数空间中的平滑性或正则性，而无论观测数据如何，大多数统计模型都具有这一性质。因此，可以预期参数空间中存在对后验密度及其梯度的准确且紧凑的近似。基于学习的方法可用于发现此类高效表示，随后采样算法便可利用这些表示，大幅降低计算成本并

提升可扩展性。为此，可以构建计算高效的代理函数，以替代原本计算开销较大的目标函数 [60–68]。

近年来，校准–仿真–抽样（CES）框架 [69] 已成为一种有效的大规模不确定性量化方法，尤其适用于逆问题 [70], ：通过校准模型获得参数–后验分布评估，仿真参数到后验分布的映射，并借助马尔可夫链蒙特卡罗以大幅降低的计算成本生成后验样本。标准 CES [69] 采用高斯过程（GP）仿真器（代理模型）；尽管已有各种缓解策略 [65, 71–74]，但其通常仍面临立方级计算复杂度和高维可扩展性方面的限制。为解决这一问题，Lan 等人 [70] 提出以卷积神经网络替代 GP 仿真，使 CES 能够从数百维扩展至数千维。沿着这一研究思路，Moslemi 等人 [75] 提出了一种基于 CES 的神经网络贝叶斯推断方法，该方法采用深度（非贝叶斯）神经网络作为仿真器，利用其在近似高维函数方面出色的计算效率和灵活性。

另一类算法旨在结合马尔可夫链蒙特卡罗和变分贝叶斯方法，以发挥两种方法各自的优势。de Freitas 等人 [15], 提出了这一方向上的早期尝试：他们在分块 Metropolis–Hastings（MH）算法中使用变分近似作为提议分布，以快速识别高概率区域，从而加快收敛。Salimans 等人[76], 也探索了变分推断与马尔可夫链蒙特卡罗的结合方式，即在变分近似中嵌入一个或多个马尔可夫链蒙特卡罗步骤。另一方面，Zhang 等人 [16], 提出了一个将变分近似融入哈密顿蒙特卡罗（HMC）的框架，以减少采样过程中所需的高昂计算开销。他们表明，马尔可夫链蒙特卡罗步骤带来的额外灵活性能够构成丰富的分布族，与常用的变分贝叶斯方法相比，这些分布能够更准确地近似精确后验分布。

本文提出了一个通用框架，将上述所有方法统一为其特例。我们的框架并不依赖单一的精确或近似方法，而是将两种方法作为不同模块纳入其中，并通过一种受人类探索环境和作出决策时调用不同神经系统的方式启发的机制，在参数空间中进行探索，并动态地在两者之间作出选择。

3 预备知识

虽然我们提出的框架可用于一般的贝叶斯推断，但下文将在贝叶斯深度学习的背景下讨论其细节，该领域的计算复杂度通常高得令人望而却步。标准神经网络（NN）由输入层 $\boldsymbol{l}_0 = \boldsymbol{X}$、针对 $l = 1,\ldots,m-1$的一系列隐藏层 $\boldsymbol{l}_l$ 以及输出层 $\boldsymbol{l}_m = \boldsymbol{Y}$组成。每一层先施加仿射变换，再施加非线性激活函数 g_l [77]:

$$\boldsymbol{l}_l = g_l\left(\boldsymbol{W}_l \boldsymbol{l}_{l-1} + \boldsymbol{b}_l\right), \quad l = 1,\ldots,m-1.$$

网络参数记为 $\boldsymbol{\theta} = (\boldsymbol{W}, \boldsymbol{b})$，固定架构定义了由 $\boldsymbol{\theta}$索引的函数类。给定训练数据集 $\{(x_n, y_n)\}_{n=1}^N$，学习涉及估计 $\boldsymbol{\theta}$ 以逼近映射 $\boldsymbol{X} \to \boldsymbol{Y}$。该神经网络定义了前向算子 $\mathcal{G}(\boldsymbol{\theta})$，使得

$$\hat{y}_n = \mathcal{G}(x_n;\boldsymbol{\theta}), \qquad y_n = \hat{y}_n + \varepsilon_n, \quad \varepsilon_n \sim \mathcal{N}(0,\Gamma),$$

其中， $\boldsymbol{Y}$在回归问题中是连续的，或在分类中是连续潜变量。
训练通常涉及求解

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{Y} - \mathcal{G}(\boldsymbol{X};\boldsymbol{\theta}))\,,$$

并使用随机梯度方法。假设存在高斯观测噪声，则得到负对数似然损失

$$\Phi(\boldsymbol{\theta};\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{2}\left\|\boldsymbol{Y} - \mathcal{G}(\boldsymbol{X};\boldsymbol{\theta})\right\|_\Gamma^2.$$

标准深度学习依赖于对 $\boldsymbol{\theta}$的点估计。这种方法计算效率高，但缺乏恰当的不确定性量化（UQ） [78, 79]。为了解决这一问题，贝叶斯神经网络（BNN）通过在参数上设定先验分布 p（$\boldsymbol{\theta}$），并进行贝叶斯推断 [80, 81], 提供了一种有理论依据的不确定性量化框架。由此得到的后验分布为

$$p(\boldsymbol{\theta} \mid X, Y) \propto p(Y \mid X, \boldsymbol{\theta})\, p(\boldsymbol{\theta}).$$

我们假设观测服从高斯观测模型：

$$p(y \mid \theta) \quad = \quad \prod_{i=1}^{N} \mathcal{N}\big(y_i \,;\, f_\theta(x_i), \sigma^2\big);$$

$$\ell(\theta) = \log p(y \mid \theta) \quad = \quad -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \big(y_i - f_\theta(x_i)\big)^2 + C,$$

采用零均值高斯先验， $p(\theta)$。因此，后验分布为： $\pi(\theta) \propto \exp\big(\ell(\theta)\big)\, p(\theta)$。

4 方法

我们提出了一个受人类探索、学习和决策启发的后验近似通用框架。经典框架通过两个计算系统之间的交互来解释此类行为 [19]: 基于模型的控制支持目标导向行为，无模型控制则支持较简单的行为，例如习惯。基于模型的机制利用环境的内部模型计算前瞻性行动价值，从而支持灵活但计算成本高昂的规划；而无模型机制利用价值估计，实现快速高效的响应。尽管这些理论取得了成功，但它们难以解释决策者如何在具有稀疏数据、长程时间依赖关系和丰富潜在结构的环境中迅速适应。为实现快速行为适应，有人提出增加第三个系统——情景记忆控制，通过记忆检索与评估来实现这一目标 [18]。具体而言，情景记忆控制会存储个体经验的详细表示，并能够跨越长程依赖关系 [82]。

基于这些洞见，我们提出了神经启发式后验近似（NIPA）框架，用于实现计算高效的贝叶斯推断。我们的框架包含以下三个组件：

• **基于模型（MB）：** 由目标分布引导，以支持对参数空间进行审慎探索，使其具有灵活性，但计算速度较慢。

• **模型自由（MF）：** 通过基于过去对参数空间的探索构建替代函数（类似于启发式方法和习惯），实现快速、反射式采样。

• **情景控制（EC）：** 通过回忆特定的过去事件（即样本），支持快速的单次采样。

在该框架中，MB 组件（由目标分布引导）探索参数空间，以生成高质量、经过“深思熟虑”的样本。随后，MF 组件利用这些样本学习并近似参数空间上的高成本函数，而 EC 组件则存储记忆，以便回忆过去的函数评估结果，从而进行单次（即基于单个回忆的过去评估结果、无需额外计算）接受—拒绝决策。请注意，前文讨论的许多现有方法都可以视为这一通用框架的特例。具体而言，当仅依赖 MB 组件时，我们的方法退化为一种精确采样算法；当仅依赖 MF 组件时，则退化为 CES。所提出的框架整合了这两个系统，并进一步通过 EC 模块对其进行补充。

下面，我们详细介绍这一后验近似框架的一种简单而具体的实现。对于 MB 组件，我们采用标准的 HMC算法，精确评估对数后验及其梯度。对于 MF 组件，我们将自编码器与前馈深度神经网络（DNN）相结合：自编码器降低参数空间的维度，而 DNN 则在所得的低维空间中构建代理函数。EC 组件则通过一种简单的基于距离的门控机制来获取先前的值。我们指出，该框架还有许多其他可能的实现方式，其中一些将在讨论部分中介绍。

初始样本池构建。我们首先在采用常规贝叶斯神经网络参数化的全量数据后验分布 $\pi(\theta)$ 上运行标准的随机梯度哈密顿蒙特卡洛（SGHMC） [83]采样器。从该 SGHMC 链中，我们保留一组初始 M_0 后验样本（通常为 $M_0 = 100$），并使用全量数据（即不进行小批量处理）计算其对应的精确对数后验值 $\log \pi_{\text{train}}(\theta^{(i)})$:

$$\mathcal{P}_0 \;=\; \big\{(\theta^{(i)}, \log \pi_{\text{train}}(\theta^{(i)}))\big\}_{i=1}^{M_0},$$

这些样本被标记为 θ_{MB}。

代理建模。 使用初始样本 $\{\theta^{(i)}\}_{i=1}^{M_0}$，我们训练一个自编码器，以获得参数空间的低维潜在嵌入，

$$u \;=\; \mathrm{Encode}(\theta) \in \mathbb{R}^{d_u}, \qquad \theta \approx \mathrm{Decode}(u).$$

接着，我们将所有初始样本嵌入，以形成训练对

$$\big(u^{(i)},\, \log \pi(\theta^{(i)})\big), \qquad u^{(i)} = \mathrm{Encode}(\theta^{(i)}),$$

并利用这些训练对构建代理模型。为此，我们在所得的低维空间中拟合一个 DNN回归模型 $g(\cdot)$，根据潜在编码 u预测对数后验：

$$\widehat{\log \pi(\theta)} = g(u), \quad \text{with } u = \mathrm{Encode}(\theta).$$

神经启发式后验近似（NIPA）。现在，我们介绍 NIPA 框架，该框架由三个组件构成：MB、MF 和 EC。

在第 t 次迭代中，我们首先使用高斯随机游走生成一个门控候选， q_{RW}：

$$\tilde{\theta} = \theta + \sigma\varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I),$$

并利用其与样本池的距离 t o 来确定应激活哪个组件。令 $\mathcal{P}$ 表示当前样本池 l （初始化为 $\mathcal{P}_0$ 和 并在线更新）。我们在标准化参数空间中计算 θ' 到样本池的距离 :

$$d^*(\tilde{\theta}, \mathcal{P}) \;=\; \min_{\theta_i \in \mathcal{P}} d(\tilde{\theta}, \mathcal{P}) \;=\; \min_{\theta_i \in \mathcal{P}} \left\| \frac{\tilde{\theta} - m}{s} - \frac{\theta_i - m}{s} \right\|_2,$$

其中， m 和 s 分别是根据当前样本池估算的逐坐标均值和标准差。然后，我们使用两个阈值 t_1 和 t_2 （ $t_1 < t_2$），从而选择 NIPA 的三个组成部分之一：

• *MB*： 如果 $d^*(\theta', \mathcal{P}) > t_2$，附近没有样本这一情况表明，代理函数在参数空间中识别出的过往记忆和模式可能不可靠。在这种情况下，采样器会继续探索参数空间，并基于目标分布 $\pi(\theta)$ 执行精确梯度 HMC 步骤。如果 HMC 提议 θ' 被接受，我们会将其标记为 θ_{MB} （精确评估），并将相应的样本对（$\theta_{\text{新}}$, $\log \pi(\theta_{\text{新}})$ ）添加到样本池中。

• *MF*： 如果 $t_1 < d^*(\theta', \mathcal{P}) \le t_2$，我们处于一个没有可比记忆片段、但样本数量足以确保代理函数可靠的区域。在这种情况下，我们将门控候选视为提案，$\theta' = \theta$，并使用代理函数近似对数后验密度，而非进行精确评估，以此应用 Metropolis 接受概率：

$$u' = \mathrm{Encode}(\theta'), \qquad \widehat{\log \pi(\theta')} := g(u').$$

若接受， θ' 将成为新状态，我们将其标记为 θ_{MF} （近似评估），并将对应的数据对 $\frown(\theta', \log \pi(\theta'))$ 添加到样本池中。

• 情景控制（*EC*）：如果 $d^*(\theta', \mathcal{P}) \le t_1$，则表明对过往经验的记忆很强，且提案位于某个˜先前样本附近。在这种情况下，我们再次设定 $\theta' = \theta$，但既不计算其精确对数后验，也不使用代理函数计算 Metropolis 接受概率。相反，我们从样本池中检索最近的已存储情景（样本），并将其缓存的对数后验值用作该提案的近似值。接受的 EC 移动会更新当前状态，但不会被加入 $\mathcal{P}$。也就是说，我们不会将其视为新的记忆。

我们每经过 k 次迭代（通常为 100 次）便更新一次代理函数。算法 1 总结了 NIPA 的精确实现。由此得到的流程结合了快速的、基于记忆或学习得到的近似值，以及审慎可靠的决策机制。EC 组件通过跳过先前探索过的状态附近的对数后验评估，大幅提升速度；MF 组件则在样本充足、能够构建可靠启发式方法时，提供成本低廉的基于代理模型的评估；MB 组件则通过定期使用精确后验分布进行稳健更新，确保探索过程可靠。这种组合旨在降低计算成本，同时保持经验上稳定的后验探索。

5 实验

基线方法。 我们将所提出的方法与一系列基线方法进行比较，包括使用 HMC 采样器的贝叶斯神经网络模型（BNN-HMC）、使用 SGHMC 采样器的贝叶斯神经网络模型（BNN-SGHMC）、使用 pCN 采样器的贝叶斯神经网络模型（BNN-pCN）、贝叶斯 Lasso （BNN- LASSO） [84], 变分推断（BNN-VI） [85], MC-Dropout（BNN-MCD） [86],随机权重平均高斯方法（SWAG） [87, 88], 以及随机网络代理（BNN-RNS） [89]。我们还纳入了深度神经网络（DNN）的结果，

算法 1 神经启发式后验近似 (NIPA)

要求： 目标后验分布 $\pi(\theta) \propto \exp(\ell(\theta))p(\theta)$；阈值 $t_1 < t_2$；初始样本池大小 M_0；总迭代次数 T；随机游走提案 $q_{\mathrm{RW}}(\cdot \mid \cdot)$；精确 MB 核 $\mathcal{K}_{\mathrm{MB}}$（哈密顿蒙特卡洛，含 Metropolis-Hastings 校正）；编码器 Enc 与代理模型 g；代理模型更新间隔 k。

定义：

样本池： $\mathcal{P}$ 存储元组 (θ_i, L_i, o_i)，其中 $o_i \in \{\text{MB}, \text{模型自由}\}$；$L_i = \log \pi(\theta_i)$ 用于 MB 以及 $L_i = \widehat{\text{MF}}$ 的 $\log \pi(\theta_i)$。

标准化池距离： $d^*(\tilde{\theta}, \mathcal{P}) = \min_{(\theta_i,\cdot,\cdot)\in\mathcal{P}} \left\| \frac{\tilde{\theta}-m}{s} - \frac{\theta_i-m}{s} \right\|_2$，其中 (m, s) 是逐坐标在 $\{\theta_i$ 上的均值/标准差：$(\theta_i, \cdot, \cdot) \in \mathcal{P}\}$。

运行一个以 $\pi(\theta)$ 为目标的随机梯度哈密顿蒙特卡洛链，并保留 M_0 个状态 $\{\theta^{(i)}\}_{i=1}^{M_0}$。 ▷ 初始样本池构建

$\mathcal{P} \leftarrow \emptyset$。

对于 $i = 1$ 至 M_0 **执行**

  $L^{(i)} \leftarrow$ 对数 $\pi(\theta^{(i)})$ ▷ 精确的完整对数后验

  将（$\theta^{(i)}$，$L^{(i)}$，MB）添加到 $\mathcal{P}$。

结束循环

在 $\{$ 上用 g 和 $\theta^{(i)}$ 训练 Enc（$\text{Enc}(\theta^{(i)})$，$\log \pi(\theta^{(i)})$）$\}_{i=1}^{M_0}$。 ▷ 构建代理函数

初始化 $\theta \leftarrow \theta^{(M_0)}$。

对于 $t = M_0 + 1$ 到 T **执行** ▷ 后验近似

  门控候选（随机游走）： 抽取 $\tilde{\theta} \sim q_{\mathrm{RW}}(\cdot \mid \theta)$。

  计算 $d^* \leftarrow d^*($ $\tilde{\theta}$，$\mathcal{P})$ 并检索最近的 $(\theta_j$，L_j，$o_j) \in \mathcal{P}$。

  如果 $d^* > t_2$ **则** ▷ MB；通过哈密顿蒙特卡洛提出一个新样本 θ'

    提议生成（哈密顿蒙特卡洛）： 采样 $(p_0 \sim \mathcal{N}(0$，$I))$ 并执行蛙跳步以提出 θ'。

    接受概率： $\alpha \leftarrow \min\{1, \exp(H(\theta, p_0) - H(\theta', p'))\}$。

    抽取 $u \sim$ 均匀分布(0,1)。

    如果 $u < \alpha$ **则**

      $\theta \leftarrow \theta'$。

      $(\theta_{\mathrm{MB}}$，$L_{\mathrm{MB}}) \leftarrow (\theta'$，对数 $\pi(\theta'))$。 ▷ 精确评估

      将（θ_{MB}，L_{MB}，MB）添加到 $\mathcal{P}$。

    结束 if

  else if $t_1 < d^* \le t_2$ **then** ▷ 模型自由；将 $\tilde{\theta}$ 作为提案其视为

    $\theta' = \tilde{\theta}$

    近似函数： $L' \leftarrow \widehat{\log \pi}(\theta') = g(\text{Enc}(\theta'))$

    参考： $L_{\mathrm{ref}} \leftarrow L_{\mathrm{MF}}$

    接受概率： $\alpha \leftarrow \min\{1, \exp(L' - L_{\mathrm{ref}})\}$ ▷ Metropolis 更新

    从 $u \sim$ 均匀分布(0,1)中抽取。

    如果 $u < \alpha$ **则**

      $\theta \leftarrow \theta'$。

      （θ_{MF}，L_{MF}）$\leftarrow$（θ'，L'）。

      将（θ_{MF}，L_{MF}，模型自由）添加到 $\mathcal{P}$。

    结束 if

  else ▷ 情景控制；将 $\tilde{\theta}$ 一个提案其视为

    $\theta' = \tilde{\theta}$

    近似函数： $L_{\mathrm{proxy}} \leftarrow L_j$ ▷ 检索缓存值

    参考： $L_{\mathrm{ref}} \leftarrow \begin{cases} L_{\mathrm{MB}} & \text{如果 } o_j = \text{MB}, \\ L_{\mathrm{MF}} & \text{if } o_j = \text{模型自由}. \end{cases}$

    接受概率： $\alpha \leftarrow \min\{1, \exp(L_{\mathrm{proxy}} - L_{\mathrm{ref}})\}$ ▷ Metropolis 更新

    从 $u \sim$ 均匀分布(0，1)中抽取；**如果** $u < \alpha$ **则** $\theta \leftarrow \theta'$ ▷ 不要添加到 $\mathcal{P}$

  结束条件判断

  记录 $\theta^{(t)} \leftarrow \theta$。

  如果 t 取模 $k = 0$**则** ▷ 更新代理函数

    使用当前样本池 $\mathcal{P}$ 重新拟合代理模型

  结束如果

结束循环

表 1：NIPA 与现有方法在回归问题上的比较。在计算效率方面，NIPA 显著优于 BNN-HMC 基线，在合成数据和 YearPrediction 数据上的效率分别是其 8.27 倍和 6.99 倍；同时，其均方根误差与其他方法相当，且 95%覆盖概率更高。

Method	Initial	MB	MF	EC	Time (s)	ESS (Min, Med, Max)	minESS/s	RMSE	CP95	Speedup
DNN	—	—	—	—	81.48	—	—	7.15	—	—
DNN-Ensemble	—	—	—	—	300.35	—	—	7.79	0.45	—
BNN-MCD	—	—	—	—	146.26	—	—	6.89	0.62	—
BNN-VI	—	—	—	—	365.71	—	—	9.38	0.77	—
BNN-LASSO	—	—	—	—	122.59	—	—	5.87	0.56	—
BNN-HMC (Baseline)	—	—	—	—	659.73	(43.42, 656.52, 1544.23)	0.06	6.24	0.72	1.00
BNN-SGHMC	—	—	—	—	596.55	(58.24, 756.33, 1538.62)	0.09	6.72	0.76	1.57
BNN-pCN	—	—	—	—	302.25	(60.68, 855.33, 1389.42)	0.18	7.28	0.40	3.29
BNN-RNS	—	—	—	—	872.38	(39.86, 583.18, 1398.87)	0.05	7.36	0.57	0.82
SWAG	—	—	—	—	1039.54	(47.36, 567.34, 1487.34)	0.05	8.98	0.54	0.82
NIPA (MB only)	100	1900	0	0	626.99	(46.25, 619.62, 1548.71)	0.07	5.76	0.75	1.32
NIPA (MF only)	100	0	1900	0	112.32	(62.45, 820.37, 1542.18)	0.59	9.35	0.33	10.60
NIPA (EC only)	100	0	0	1900	75.75	(66.20, 795.27, 1539.37)	0.48	10.95	0.56	15.66
NIPA (MB/MF/EC)	100	241	1213	446	142.92	(62.45, 820.37, 1542.18)	0.46	6.79	0.82	8.27

(a) 合成回归问题。

Method	Initial	MB	MF	EC	Time (s)	ESS (Min, Med, Max)	minESS/s	RMSE	CP95	Speedup
DNN	—	—	—	—	215.33	—	—	7.83	—	—
DNN-Ensemble	—	—	—	—	2154.87	—	—	7.96	82.34	—
BNN-MCD	—	—	—	—	5054.49	—	—	9.42	81.61	—
BNN-VI	—	—	—	—	10163.66	—	—	9.44	80.63	—
BNN-LASSO	—	—	—	—	12649.78	—	—	8.69	82.42	—
BNN-HMC (Baseline)	—	—	—	—	10875.42	(54.75, 1054.98, 1651.09)	0.0053	8.96	85.89	1.00
BNN-SGHMC	—	—	—	—	8398.18	(49.65, 902.18, 1543.95)	0.0059	9.59	81.53	1.17
BNN-pCN	—	—	—	—	6271.20	(44.87, 896.54, 1536.88)	0.0071	9.29	82.19	1.34
BNN-RNS	—	—	—	—	17872.12	(56.87, 994.56, 1573.32)	0.0031	9.12	82.98	0.60
SWAG	—	—	—	—	9054.87	(49.66, 983.87, 1580.43)	0.0054	9.08	83.64	1.03
NIPA (MB only)	100	1900	—	—	10419.19	(55.62, 1063.87, 1674.92)	0.0049	8.92	82.72	1.01
NIPA (MF only)	100	—	1900	—	1162.48	(49.24, 909.42, 1546.87)	0.0423	8.88	79.31	7.99
NIPA (EC only)	100	—	—	1900	828.91	(12.06, 765.66, 1397.91)	0.0145	9.48	81.14	2.46
NIPA (MB/MF/EC)	100	98	1524	278	1225.08	(45.38, 961.23, 1573.82)	0.0370	8.89	83.15	6.99

(b) Year Prediction MSD。

该方法不提供不确定性量化，但可作为参考。此外，我们报告了集成深度学习方法（DNN 集成）的结果 [90], 。该方法由多个深度神经网络组成，每个网络均使用不同的随机种子进行初始化。尽管这种方法能够进行不确定性量化，但它无法提供用于分析的概率框架，而这正是贝叶斯神经网络及其变体的一项关键优势。

指标。 对于回归基准，我们使用留出测试集上的均方根误差（RMSE）来报告预测性能。我们使用名义覆盖率为 95% 的后验预测区间的经验覆盖概率（CP95）来评估不确定性量化的质量。其定义为落在 95% 区间 $[L_i, U_i]$ 内的测试目标值所占的比例 y_i ，

$$\mathrm{CP95} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{L_i \le y_i \le U_i\},$$

这里， $[L_i, U_i]$ 由 x_i处的后验预测分布构造。

对于分类问题，我们报告测试准确率，并使用期望校准误差（ECE） [91, 92]评估不确定性质量。对于每个测试点 x_i，我们计算后验均值预测概率 $\bar{p}(x_i)$，并以置信度 $c_i = \max_k \bar{p}_k(x_i)$将预测类别指定为$\hat{y}_i = \mathrm{argmax}_k\, \bar{p}_k(x_i)$。根据置信度将测试点分入 B 个分箱；对于每个分箱 b，我们计算经验准确率 $\mathrm{acc}(B_b)$和平均置信度 $\mathrm{conf}(B_b)$。随后，ECE 定义为各分箱中这两个量的绝对差的加权平均值：

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|B_b|}{n} \left|\mathrm{acc}(B_b) - \mathrm{conf}(B_b)\right|,$$

数值越小表示校准效果越好。

除了预测性能和不确定性量化外，我们还使用实际运行时间 T和有效样本量（ESS） [93, 94]评估计算效率与采样质量。每种采样算法都会生成$S = 2000$ 个后验样本 $\{\theta^{(t)}\}_{t=1}^{S}$，其中 $\theta^{(t)} \in \mathbb{R}^D$ 将所有权重和偏置连接起来。我们在单个标量参数层面评估混合情况：对于每个 $j \in 1, \ldots, D$，我们将 $\{\theta_j^{(t)}\}_{t=1}^{S}$ 视为单变量马尔可夫链，并计算其 ESS。随后，以最小值、中位数和最大值汇总逐参数有效样本量。最小有效样本量（minESS）反映混合最慢的方向。为综合考量采样质量与计算成本，我们报告每秒最小有效样本量 minESS/s，并据此计算采样算法相对于 BNN-HMC 基线的加速比。

回归。我们首先基于一个具有 ReLU 激活的双隐藏层前馈神经网络构建合成回归基准。该网络将随机生成的输入映射为标量输出，并叠加加性高斯噪声。令 d 表示输入特征维度， N 表示样本量。网络架构包含两个隐藏层，其宽度分别为 h_1 和 h_2。我们将所有权重和偏置拼接成一个参数向量 $\theta \in \mathbb{R}^D$，其中 D 由层维度决定。

从元素独立同分布且服从均匀分布的样本中抽取设计矩阵 $X \in \mathbb{R}^{N\times d}$ ，其中每一行均视为一个预测变量。随后，从有界均匀分布中抽取元素独立同分布的真实参数向量 $\theta^\star$ ，该分布

表 2：NIPA 与现有方法在分类问题上的比较。在计算效率方面，NIPA 显著优于 BNN-HMC 基线方法：在合成数据和 MNIST 上分别快 8.44 倍和 8.65 倍，同时准确率相当，且相较于其他方法，期望校准误差（ECE）更低。

Method	Initial	MB	MF	EC	Time (s)	ESS (Min, Med, Max)	minESS/s	Accuracy	ECE	Speedup
DNN	—	—	—	—	483.21	—	—	76.85	—	—
DNN-Ensemble	—	—	—	—	1865.83	—	—	77.03	0.40	—
BNN-MCD	—	—	—	—	2279.94	—	—	76.95	0.42	—
BNN-VI	—	—	—	—	6673.74	—	—	76.83	0.39	—
BNN-LASSO	—	—	—	—	3581.42	—	—	77.32	0.54	—
BNN-HMC (Baseline)	—	—	—	—	7327.24	(51.82, 881.99, 1545.60)	0.0070	75.10	0.44	1.00
BNN-SGHMC	—	—	—	—	3676.01	(19.44, 195.76, 1540.76)	0.0052	78.17	0.39	0.76
BNN-pCN	—	—	—	—	1327.54	(43.24, 783.28, 1495.30)	0.0325	71.62	0.46	4.65
BNN-RNS	—	—	—	—	9321.76	(39.54, 852.13, 1493.43)	0.0042	69.40	0.42	0.61
SWAG	—	—	—	—	8954.39	(41.75, 847.32, 1503.53)	0.0046	72.39	0.41	0.67
NIPA (MB only)	100	1900	0	0	7452.92	(51.87, 886.64, 1545.20)	0.0069	77.22	0.38	0.99
NIPA (MF only)	100	0	1900	0	719.26	(36.31, 854.20, 1543.41)	0.0504	64.17	0.27	7.25
NIPA (EC only)	100	0	0	1900	540.70	(5.22, 167.27, 702.06)	0.0096	65.38	0.23	1.39
NIPA (MB/MF/EC)	100	355	1059	486	778.87	(45.77, 912.19, 1542.88)	0.0587	76.31	0.35	8.44

(a) 合成分类问题。

Method	Initial	MB	MF	EC	Time (s)	ESS (Min, Med, Max)	minESS/s	Accuracy	ECE	Speedup
DNN	—	—	—	—	801.63		—	98.82	—	—
DNN-Ensemble	—	—	—	—	4632.51	—	—	98.04	0.49	—
BNN-MCD	—	—	—	—	4113.04	—	—	93.12	0.48	—
BNN-VI	—	—	—	—	8759.58	—	—	94.08	0.47	—
BNN-LASSO	—	—	—	—	7416.91	—	—	95.63	0.46	—
BNN-HMC (Baseline)	—	—	—	—	12042.83	(62.54, 964.61, 1652.20)	0.0051	98.07	0.51	1.00
BNN-SGHMC	—	—	—	—	7694.74	(35.91, 843.96, 1503.42)	0.0045	98.19	0.48	0.88
BNN-pCN	—	—	—	—	4921.38	(45.36, 804.31, 1489.05)	0.0092	95.45	0.46	1.80
BNN-RNS	—	—	—	—	9628.62	(33.80, 829.14, 1493.22)	0.0035	94.28	0.52	0.68
SWAG	—	—	—	—	7432.94	(45.61, 893.66, 1583.91)	0.0061	96.33	0.49	1.20
NIPA (MB only)	100	1900	0	0	12203.57	(65.96, 969.32, 1669.43)	0.0054	98.12	0.52	1.06
NIPA (MF only)	100	0	1900	0	1013.06	(47.38, 905.82, 1604.71)	0.047	96.94	0.34	9.21
NIPA (EC only)	100	0	0	1900	824.53	(42.84, 826.57, 1404.33)	0.039	95.21	0.31	7.81
NIPA (MB/MF/EC)	100	137	1149	714	1579.43	(69.74, 995.36, 1632.59)	0.028	98.01	0.39	8.65

(b) MNIST奇偶数字。

定义了数据生成机制。无噪声潜在信号通过前向传播获得，即 $f(X;\theta^\star) \in \mathbb{R}^N$，观测响应则通过添加高斯噪声生成，

$$y = f(X;\theta^\star) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}\big(0, \mathrm{diag}(\sigma^2)\big),$$

其中，噪声方差 σ^2 的选择用于控制噪信比水平。所得数据集包含5,000个观测值、100个特征和3,505个模型参数。

我们还基于Year Prediction MSD数据集 [95], 评估了我们的方法，其目标是根据音频特征预测歌曲的发行年份。这些歌曲主要是西方商业音乐，发行年份从1922年到2011年不等，其中2000年代的歌曲数量达到峰值。该数据集包含515,345个观测值、90个特征和309,761个模型参数。
每个数据集都被划分为训练集和测试集，并且对所有方法保持固定。测试集上的结果见表1。如我们所见，与BNN-HMC基线相比，NIPA的效率显著更高，同时具有相当的均方根误差，并且其95%覆盖概率高于其他方法。

分类。 接下来，我们构建一个高维合成二分类基准，同样使用具有两个隐藏层和 ReLU 激活的前馈神经网络。输入采样自标准化高斯设计矩阵 $X \in \mathbb{R}^{N\times d}$。该网络的隐藏层宽度分别为 h_1 和 h_2，所有权重和偏置均连接成参数向量 $\theta \in \mathbb{R}^D$，从而得到一个维度极高的参数空间。

我们采样真实参数向量 $\theta^\star \sim \mathcal{N}(0,\ \Sigma)$，其中 Σ 按权重组和偏置组进行分块指定。给定 $\theta^\star$，我们计算对数几率 $\ell = f(X;\theta^\star) \in \mathbb{R}^N$，并根据伯努利分布采样二分类标签：

$$y_i \sim \mathrm{Bernoulli}\big(\mathrm{logit}^{-1}(\ell_i)\big), \qquad i = 1,\dots,N.$$

所得数据集包含 20,000 个观测值、512 个特征和 147,841 个模型参数。
我们还在 MNIST 数据集上评估了我们的方法，该数据集通常用作手写数字分类的基准 [96]。在这里，我们专注于区分奇数和偶数。该数据集包含 70,000 个观测值、784 个特征和 217,538 个模型参数

对于每个问题，我们将数据划分为训练集和测试集，并在所有方法中保持划分不变。表 2 给出了各种模型在测试集上的结果。如我们所见，与 BNN-HMC 基线相比，NIPA 的计算效率更高，准确率相当；与其他方法相比，其 ECE 更低，表现更佳。

6 讨论

我们提出了一种神经启发式贝叶斯推理框架，旨在通过模仿人类利用多种机制从环境中有效学习并做出决策的方式，提高计算效率。该框架由三个组件构成，分别类似于大脑中的基于模型、模型自由和情景控制。将这些系统整合到后验近似算法中，便形成了一个通用框架，其中涵盖了许多现有贝叶斯计算方法作为特例。此外，我们还表明，所得框架能够提高贝叶斯推断的计算效率，尤其是在计算开销传统上令人望而却步的深度学习场景中。

本文所述算法的具体设置简单、直观且在实践中有效。尽管如此，未来的工作仍可进一步优化这些选择，以提高算法在更广泛模型和数据集上的泛化能力与鲁棒性。例如，当前版本的 NIPA 依赖于阈值 t_1 和 t_2，这些阈值被视为类似于哈密顿蒙特卡洛中步长的调优参数。然而，选择其最优值并非易事。有趣的是，大脑中的证据表明，控制会根据不同机制的相对不确定性，在多种机制之间动态分配，从规范性角度解释了何时灵活规划应优先于习惯性效率 [17]。依照这一思路，我们的门控方法可以进一步扩展，根据模型自由组件或 EC 组件对估计值的置信度，决定使用哪一种效率更高的组件。

如前所述，这种方法有许多变体，具体取决于用于生成初始样本池、执行基于模型的采样、为无模型采样构建替代函数以及检索情景记忆的具体算法。例如，可以不使用随机梯度哈密顿蒙特卡洛，而是采用随机梯度优化方法，并沿优化轨迹收集样本（与 SWAG 类似，但并不完全相同）。对于基于模型的采样，可以采用预条件 Crank–Nicolson（pCN）算法 [97]，因为它更擅长探索众数附近的分布 [75]。此外，可以对 MB 组件中使用的梯度函数进行近似，以实现更快的计算 [98]。对于代理模型，可以采用比深度神经网络更简单的替代方案，例如极限学习机 [67], 。此外，在情景控制组件中，不必依赖单个情景；可以使用 k 近邻模型，利用多个邻近情景来近似对数后验。

最后，尽管本文聚焦于贝叶斯深度学习的应用，所提出的方法也能轻松扩展到其他计算密集型贝叶斯场景，例如高斯过程模型以及针对机理模型的贝叶斯逆问题。虽然这些方法为复杂数据建模提供了灵活且有理论依据的框架，但其高昂的计算成本往往限制了它们的广泛应用。

致谢

我们谨向 Norbert Fortin 博士（加州大学欧文分校神经生物学系）和 Aaron Bornstein 博士（加州大学欧文分校认知科学系）致谢，感谢他们就启发本研究的神经科学概念所提供的真知灼见和指导。

参考文献

[1] N. Metropolis、A. W. Rosenbluth、M. N. Rosenbluth、A. H. Teller 和 E. Teller。《快速计算机的状态方程计算》。《化学物理学杂志》，21(6):1087–1092，1953年。[2] S. Duane、A. D. Kennedy、B J. Pendleton 和 D. Roweth。混合蒙特卡罗。《物理快报*B*》，195(2):216 – 222，1987年。[3] R.M. Neal。使用哈密顿动力学的MCMC。载于 S. Brooks、A. Gelman、G. Jones 和 X. L. Meng 编，《马尔可夫链蒙特卡洛手册，第113–162页。Chapman and Hall/CRC，2011年。[4] M. Welling 和 Y. W. Teh。通过随机梯度朗之万动力学进行贝叶斯学习。载于国际机器学习会议论文集，2011年。[5] Sungjin Ahn、Anoop Korattikara 和 Max Welling。通过随机梯度费舍尔评分进行贝叶斯后验采样。载于第*29*届国际机器学习会议（*ICML*）论文集，第1591–1598页，2012年。[6] T. Chen、E. B. Fox 和 C. Guestrin。随机梯度哈密顿蒙特卡罗。预印本，2014年。[7] Nan Ding、Youhan Fang、Ryan Babbush、Changyou Chen、Robert D Skeel 和 Hartmut Neven。使用随机梯度恒温器进行贝叶斯采样。载于 Z. Ghahramani、M. Welling、C. Cortes、N. D. Lawrence 和 K. Q. Weinberger 编，神经信息处理系统进展 *27*，第3203–3211页。Curran Associates公司，2014年。[8] Michael Betancourt。可扩展哈密顿蒙特卡罗与朴素数据子采样之间的根本不相容性。载于第*32*届国际机器学习会议论文集（*ICML 2015*），法国里尔，*2015*年*7*月*6*日至*11*日，第533–540页，2015年。[9] M. I. Jordan、Z. Ghahramani、T. S. Jaakkola 和 L. K. Saul。图方法变分方法导论。载于《机器学习》，第183–233页。MIT出版社，1999年。[10] M. Wainwright 和 M. Jordan。图模型、指数族与变分推断。《机器学习基础与趋势》，1(1-2):1–305，2008年。[11] Diederik P. Kingma 和 Max Welling。自动编码变分贝叶斯。载于国际学习表征会议论文集（*ICLR*），2014年。

[12] Danilo Jimenez Rezende、Shakir Mohamed 和 Daan Wierstra。深度生成模型中的随机反向传播与近似推断。载于国际机器学习会议，页码1278–1286。PMLR，2014年。[13] Charles Blundell、Julien Cornebise、Koray Kavukcuoglu 和 Daan Wierstra。神经网络中的权重不确定性。载于国际机器学习会议，页码1613–1622。PMLR，2015年。[14] David M. Blei、Alp Kucukelbir 和 Jon D. McAuliffe。变分推断：统计学家综述。《美国统计学会会刊》，112(518):859–877，2017年。[15] N. de Freitas、P. Højen-Sørensen、M. Jordan 和 R. Stuart。变分马尔可夫链蒙特卡罗。载于第17届人工智能中的不确定性会议论文集（UAI '01），第120–127页，美国加利福尼亚州旧金山，2001年。Morgan Kaufmann出版社。ISBN 1-55860-800-1。[16] C. Zhang、B. Shahbaba 和 H. Zhao。通过得分匹配实现变分哈密顿蒙特卡罗。已提交，2016年。URL https://arxiv.org/abs/1602.02219。[17] Nathaniel D. Daw、Yael Niv 和 Peter Dayan。基于不确定性的前额叶与背外侧纹状体系统之间的行为控制竞争。《自然·神经科学》，8(12):1704–1711，2005年。数字对象唯一标识符:10.1038/nn1560。[18] Máté Lengyel 和 Peter Dayan。海马对控制的贡献：第三种途径。载于 J. C. Platt、D. Koller、Y. Singer 和 S. T. Roweis 编，《神经信息处理系统进展 20》（NeurIPS '07），第889–896页，美国纽约州雷德胡克，2007年。Curran Associates公司。[19] Nathaniel D. Daw 和 Peter Dayan。基于模型的评估的算法解剖学。《皇家学会哲学学报B辑：生物科学》，369(1655):20130478，2014年。数字对象唯一标识符:10.1098/rstb.2013.0478。[20] R.M. Neal。使用温度跃迁从多峰分布中采样。《统计与计算》，6(4):353，1996年。[21] R.M. Neal。使用马尔可夫链蒙特卡罗方法进行概率推断。技术报告 CRG-TR-93-1，统计学系，多伦多大学，1993年。[22] C. J. Geyer。实用马尔可夫链蒙特卡罗。《统计科学》，7(4):473–483，1992年。[23] P. Mykland、L. Tierney 和 B. Yu。马尔可夫链采样器中的再生。《美国统计学会会刊》，90(429):233–241，1995年。[24] James G. Propp 和 David B. Wilson。使用耦合马尔可夫链进行精确采样及其在统计力学中的应用。《随机结构与算法》，9:223–252，1996年。数字对象唯一标识符:10.1002/(SICI)1098- 2418(199608/09)9:1/2<223::AID-RSA14>3.0.CO;2-O。[25] G. O. Roberts 和 S. K. Sahu。吉布斯采样器的更新方案、相关结构、分块与参数化。《皇家统计学会会刊》B辑，59:291–317，1997年。[26] W. R. Gilks、G. O. Roberts 和 S. K. Sahu。通过再生实现自适应马尔可夫链蒙特卡罗。《美国统计学会会刊》，93(443):1045–1054，1998年。[27] G. R. Warnes。正态核耦合器：一种用于从多模态分布中高效采样的自适应马尔可夫链蒙特卡罗方法。技术报告第395号，华盛顿大学，2001年。[28] A. E. Brockwell。通过预取实现并行马尔可夫链蒙特卡罗模拟。《计算与图形统计学杂志》，第246–261页，2006年。[29] R.M. Neal。捷径梅特ropolis方法。技术报告0506，统计学系，多伦多大学，2005年。[30] R.M. Neal。切片采样。《统计学年鉴》，31(3):705–767，2003年。[31] M. J. Beal。近似贝叶斯推断的变分算法。博士论文，伦敦大学学院，英国伦敦，2003年。[32] J. Møller、A. Pettitt、K. Berthelsen 和 R. Reeves。一种适用于具有难处理归一化常数的分布的高效马尔可夫链蒙特卡罗方法。《Biometrica》，93，2006年。即将发表。[33] C. Andrieu 和 E. Moulines。关于某些自适应MCMC算法的遍历性。《应用概率年鉴》，16(3):1462–1505，2006年。[34] K. Kurihara、M. Welling 和 N. Vlassis。加速变分狄利克雷过程混合模型。载于《神经信息处理系统进展》（NIPS），第19卷，2006年。[35] Olivier Cappé、Randal Douc、Arnaud Guillin、Jean-Michel Marin 和 Christian P. Robert。一般混合类中的自适应重要性采样。《统计与计算》，18(4):447–459，2008年。

[36] R. V. Craiu、Jeffrey R. 和 Chao Y.。《向邻居学习：并行链和区域自适应马尔可夫链蒙特卡罗》。《美国统计学会会刊》，104(488):1454–1466，2009。[37] M. Welling。《通过集成动态权重进行学习》。《国际机器学习会议论文集》，2009。[38] A. Gelfand、L. van der Maaten、Y. Chen 和 M. Welling。《关于集成与循环感知机定理》。《神经信息处理系统进展》第23卷，694–702页，2010。[39] R. Randal、G. Arnaud、M. Jean-Michel 和 R. P. Christian。《通过总体蒙特卡罗实现最小方差重要性采样》。《ESAIM：概率与统计》，11:427–447，2007。[40] D. Randal 和 Christian R. P.。《Metropolis-Hastings算法的朴素Rao-Blackwell化》。《统计学年鉴》，39(1):261–277，2011。[41] M. Welling 和 Y.W. Teh。《通过随机梯度朗之万动力学进行贝叶斯学习》。《第28届国际机器学习会议（ICML）论文集》，681–688页，2011。[42] Y. Zhang 和 C. Sutton。《拟牛顿马尔可夫链蒙特卡罗》。《神经信息处理系统进展》，2011。[43] Y. Ahmadian、J. W. Pillow 和 L. Paninski。《用于解码神经尖峰序列的高效马尔可夫链蒙特卡罗方法》。《神经计算》，23(1):46–96，2011。[44] M. Girolami 和 B. Calderhead。《黎曼流形朗之万与哈密顿蒙特卡罗方法》。《皇家统计学会会刊》，B辑（含讨论），73(2):123–214，2011。[45] M. Hoffman 和 A. Gelman。《无U形转向采样器：在哈密顿蒙特卡罗中自适应设置路径长度》. arxiv.org/abs/1111.4246，2011。[46] A. Beskos、F. J. Pinski、J. M. Sanz-Serna 和 A. M. Stuart。《希尔伯特空间上的混合蒙特卡罗》。《随机过程及其应用》，121:2201–2230，2011。[47] Ben Calderhead 和 Mátyás A. Sustik。《用于微分几何马尔可夫链蒙特卡罗的稀疏近似流形》。《神经信息处理系统进展》第25卷，2888–2896页。Curran Associates公司，2012。[48] B. Shahbaba、S. Lan、W.O. Johnson 和 R.M. Neal。《分裂哈密顿蒙特卡罗》。《统计与计算》，24(3):339–349，2014。[49] S. Ahn、Y. Chen 和 M. Welling。《通过再生实现分布式自适应跳跃蒙特卡罗》。《第16届人工智能与统计学国际会议（AI Stat）论文集》，2013。[50] S. Lan、V. Stathopoulos、B. Shahbaba 和 M. Girolami。《基于拉格朗日动力学的马尔可夫链蒙特卡罗》。《计算物理学杂志》，24(2):357–378，2015。[51] S. Lan、B. Zhou 和 B. Shahbaba。《用于受约束目标分布的球面哈密顿蒙特卡罗》。《第31届国际机器学习会议（ICML）论文集》，（2014）。[52] S. Lan、J. Streets 和 B. Shahbaba。《虫洞哈密顿蒙特卡罗》。《第二十八届AAA人工智能会议论文集》，2014。[53] S. Ahn、B. Shahbaba 和 M. Welling。《分布式随机梯度马尔可夫链蒙特卡罗》。《国际机器学习会议》，2014。[54] Christopher Nemeth 和 Paul Fearnhead。《随机梯度马尔可夫链蒙特卡罗》。《美国统计学会会刊》，116(533):433–450，2021。[55] Antonios Alexos、Alex J. Boyd 和 Stephan Mandt。《结构化随机梯度马尔可夫链蒙特卡罗》。载于 Kamalika Chaudhuri、Stefanie Jegelka、Le Song、Csaba Szepesvari、Gang Niu 和 Sivan Sabato（编），《第39届国际机器学习会议论文集》，《机器学习研究论文集》第162卷，414–434页。PMLR，2022。URL https://proceedings.mlr.press/v162/alexos22a.html。[56] Jeremie Coullon、Leah F. South 和 Christopher Nemeth。《随机梯度马尔可夫链蒙特卡罗的高效且可泛化调优策略》。《统计与计算》，33(3):66，2023。数字对象唯一标识符:10.1007/s11222-023-10233-3。URL https://doi.org/10.1007/s11222-023-10233-3。[57] Matthew D. Hoffman、David M. Blei 和 Francis R. Bach。《潜在狄利克雷分配的在线学习》。载于 John D. Lafferty、Christopher K. I. Williams、John Shawe-Taylor、Richard S. Zemel 和 Aron Culotta（编），NIPS，856–864页。Curran Associates公司，2010。[58] Tianqi Chen、Emily Fox 和 Carlos Guestrin。《随机梯度哈密顿蒙特卡罗》。《国际机器学习会议》，1683–1691页，2014。[59] Naitong Chen 和 Trevor Campbell。《核心集马尔可夫链蒙特卡罗》。载于 Sanjoy Dasgupta、Stephan Mandt 和 Yingzhen Li（编），《第27届国际人工智能与统计学会议论文集》，《机器学习研究论文集》第238卷，4438–4446页。PMLR，2024年5月2–4日。

[60] J. S. Liu。《科学计算中的蒙特卡罗策略》。施普林格，2001。[61] C. E. Rasmussen。《利用高斯过程加速昂贵贝叶斯积分的混合蒙特卡罗》。《贝叶斯统计》，651–659页，2003。[62] C. Zhang、B. Shahbaba 和 H. Zhao。《基于参数空间正则性的哈密顿蒙特卡罗方法预计算策略》。arXiv预印本 arXiv:1504.01418，2015。[63] Edward Meeds 和 Max Welling。《GPs-ABC：高斯过程代理近似贝叶斯计算》。《第三十届人工智能中的不确定性会议（UAI）论文集》，593–602页，加拿大魁北克市，2014。AUAI出版社。URL https://arxiv.org/abs/1401.2838。arXiv:1401.2838。[64] J. Li 和 Y. M. Marzouk。《用于贝叶斯求解逆问题的代理模型自适应构建》。《SIAM科学计算杂志》，36(3):A1163–A1186，2014。[65] S. Lan、J. A. Palacios、M. Karcher、V. Minin 和 B Shahbaba。《基于共祖理论的非参数系统发育动力学的高效贝叶斯推断框架》。《生物信息学》，31(20):3282–3289，2015。[66] H. Strathmann、D. Sejdinovic、S. Livingstone、Z. Szabo 和 A. Gretton。《利用高效核指数族的无梯度哈密顿蒙特卡罗》。arXiv预印本 arXiv:1506.02564，2015。[67] C. Zhang、B. Shahbaba 和 H. Zhao。《利用神经网络代理函数加速哈密顿蒙特卡罗》。《统计与计算》（待刊），2015。URL http://arxiv.org/abs/1506.05555。[68] Lingge Li、Andrew Holbrook、Babak Shahbaba 和 Pierre Baldi。《神经网络梯度哈密顿蒙特卡罗》。《计算统计学》，34(1):281–299，2019。数字对象唯一标识符:10.1007/s00180-018-00861-x。[69] Emmet Cleary、Alfredo Garbuno-Inigo、Shiwei Lan、Tapio Schneider 和 Andrew M. Stuart。《校准、仿真、采样》。《计算物理学杂志》，424:109716，2021。数字对象唯一标识符:10.1016/j.jcp.2020.109716。[70] Shiwei Lan、Shuyi Li 和 Babak Shahbaba。《利用深度神经网络扩展逆问题的贝叶斯不确定性量化》。《SIAM/ASA不确定性量化杂志》，10(4):1684–1713，2022。数字对象唯一标识符:10.1137/21M1439456。[71] Haitao Liu、Yew-Soon Ong、Xiaobo Shen 和 Jianfei Cai。《当高斯过程遇上大数据：可扩展高斯过程综述》。《IEEE神经网络与学习系统汇刊》，31(11):4405–4423，2020。[72] Edwin V Bonilla、Kian Chai 和 Christopher Williams。《多任务高斯过程预测》。《神经信息处理系统进展》，20，2007。[73] Jacob Gardner、Geoff Pleiss、Kilian Q Weinberger、David Bindel 和 Andrew G Wilson。《GPyTorch：借助GPU加速实现黑盒矩阵-矩阵高斯过程推断》。《神经信息处理系统进展》，31，2018。[74] Matthias W. Seeger、Christopher K. I. Williams 和 Neil D. Lawrence。《通过快速前向选择加速稀疏高斯过程回归》。载于《人工智能与统计国际研讨会》，254–261页。PMLR，2003。[75] Zahra Moslemi、Yang Meng、Shiwei Lan 和 Babak Shahbaba。《利用神经网络扩展贝叶斯神经网络》。《机器学习研究汇刊》，2024。ISSN 2835-8856。URL https://openreview. net/forum?id=cD209UgOX7。[76] Tim Salimans、Diederik Kingma 和 Max Welling。《马尔可夫链蒙特卡罗与变分推断：弥合差距》。载于 Francis Bach 和 David Blei（编），《第32届国际机器学习会议论文集》，第37卷，1218–1226页。PMLR，2015。[77] Laurent Valentin Jospin、Hamid Laga、Farid Boussaid、Wray Buntine 和 Mohammed Bennamoun。《贝叶斯神经网络实践指南——面向深度学习用户的教程》。《IEEE计算智能杂志》，17(2):29–48，2022。[78] Chuan Guo、Geoff Pleiss、Yu Sun 和 Kilian Q Weinberger。《现代神经网络的校准》。载于《第34届国际机器学习会议论文集》，1321–1330页。PMLR，2017年8月。[79] Jeremy Nixon、Michael W Dusenberry、Linchuan Zhang、Ghassen Jerfel 和 Dustin Tran。《深度学习中的校准度量》。载于《IEEE/CVF计算机视觉与模式识别会议（CVPR）研讨会论文集》，2019。[80] David JC MacKay。《反向传播网络的实用贝叶斯框架》。《神经计算》，4(3):448–472，1992。[81] R.M. Neal。《神经网络的贝叶斯学习》。《统计学讲义》第118号，纽约：施普林格出版社，1996。

[82] Samuel J. Gershman 和 Nathaniel D. Daw。人类与动物的强化学习和情景记忆：一个整合框架。《心理学年度评论》，68:101–128，2017。数字对象唯一标识符:10.1146/annurev-psych-122414-033625。[83] Tianqi Chen、Emily Fox 和 Carlos Guestrin。随机梯度哈密顿蒙特卡洛。载于国际机器学习会议，第1683–1691页，2014。[84] T. Park 和 G. Casella。贝叶斯 Lasso。《美国统计学会会刊》，103(482):681–686，2008年6月。[85] Alex Graves。神经网络的实用变分推断。载于第*24*届神经信息处理系统国际会议论文集，NIPS'11，第2348–2356页，美国纽约州雷德胡克，2011。Curran Associates Inc. ISBN 9781618395993。[86] Yarin Gal 和 Zoubin Ghahramani。将丢弃法作为贝叶斯近似：表示深度学习中的模型不确定性。载于国际机器学习会议，第1050–1059页。PMLR，2016。[87] Pavel Izmailov、Dmitrii Podoprikhin、Timur Garipov、Dmitry P. Vetrov 和 Andrew Gordon Wilson。权重平均带来更宽的最优点和更好的泛化。载于 Amir Globerson 和 Ricardo Silva（编），《第三十四届人工智能中的不确定性会议论文集》，*UAI 2018*，美国加利福尼亚州蒙特雷，*2018*年*8*月*6*日至*10*日，第876–885页。AUAI出版社，2018。[88] Wesley J Maddox、Pavel Izmailov、Timur Garipov、Dmitry P Vetrov 和 Andrew Gordon Wilson。深度学习中贝叶斯不确定性的简单基线。神经信息处理系统进展，32，2019。[89] C. Zhang、B. Shahbaba 和 H. Zhao。利用随机基和替代函数加速哈密顿蒙特卡洛。《统计与计算》，27(6)，2017。ISSN 15731375。数字对象唯一标识符:10.1007/s11222-016-9699-1。[90] Balaji Lakshminarayanan、Alexander Pritzel 和 Charles Blundell。使用深度集成进行简单且可扩展的预测不确定性估计。神经信息处理系统进展，30，2017。[91] Mahdi Pakdaman Naeini、Gregory F. Cooper 和 Milos Hauskrecht。利用贝叶斯分箱获得校准良好的概率。载于第二十九届*AAAI*人工智能会议论文集，2015。[92] Chuan Guo、Geoff Pleiss、Yu Sun 和 Kilian Q. Weinberger。关于现代神经网络的校准。载于 Doina Precup 和 Yee Whye Teh（编），第*34*届国际机器学习会议论文集，第70卷，《机器学习研究论文集》，第1321–1330页。PMLR，2017。URL https://proceedings.mlr.press/v70/guo17a.html。[93] C. J. Geyer。马尔可夫链蒙特卡洛实用指南。《统计科学》，7(4):473–483，1992。[94] Charles J. Geyer。马尔可夫链蒙特卡洛简介。载于 Steve Brooks、Andrew Gelman、Galin L. Jones 和 Xiao-Li Meng（编），《马尔可夫链蒙特卡洛手册》，第3–48页。Chapman & Hall/CRC，2011。[95]Thierry Bertin-Mahieux。Year Prediction MSD。UCI机器学习资料库，2011。DOI: https://doi.org/10.24432/C50K61。[96] Li Deng。用于机器学习研究的手写数字图像MNIST数据库。《*IEEE*信号处理杂志》，29(6):141–142，2012。[97] Giuseppe Da Prato 和 Jerzy Zabczyk。无限维随机方程。剑桥大学出版社，2014。[98] Lingge Li、Andrew Holbrook、Babak Shahbaba 和 Pierre Baldi。神经网络梯度哈密顿蒙特卡洛。《计算统计学》，34:281–299，2019。