arXiv:2601.13272v1 [cs.LG] 19 Jan 2026

用于高效不确定性量化的多层蒙特卡洛 Dropout

AARON PIM 和 TRISTAN PRYER

Abstract. 我们提出一种基于蒙特卡洛 Dropout 的多层蒙特卡洛（MLMC）不确定性量化框架。我们将丢弃掩码视为认知随机性的来源，并根据用于估计预测矩的随机前向传播次数定义保真度层级。通过在不同保真度间重复使用丢弃掩码，我们构造了耦合粗细估计量，从而得到用于预测均值和预测方差的望远镜式 MLMC 估计量。这些估计量对于相应的随机失活诱导量保持无偏，并能在固定评估预算下减少抽样方差。我们推导出明确的偏差、方差和有效成本表达式，以及跨层级的样本分配规则。针对正向与逆向 PINNs–Uzawa 基准的数值实验验证了预测的方差收敛率，并表明在成本相当时，该方法比单层蒙特卡洛 Dropout 更有效。

1. Introduction

在许多科学计算和机器学习应用中，神经网络日益被用作复杂模拟的快速代理模型 [RPK19]。与传统机理模型不同，这些神经代理模型默认不会量化预测不确定性。它们的输出是确定性的。实际上，这会带来问题：深度网络通常能够很好地插值训练数据，但在进行外推时可能会给出过度自信且错误的预测 [Gaw+23；HW21]。在实际的代理模型流程中，这种不确定性的缺失可能成为可靠部署的实际障碍，包括将代理模型用于重复评估和决策的场景 [PP25]。

因此，亟需为神经代理模型配套不确定性量化（UQ）。贝叶斯深度学习方法通过对网络输出建立分布来解决这一问题。例如，蒙特卡洛 Dropout（蒙特卡洛随机失活）将测试时的随机失活视为近似贝叶斯推断，无需改变网络架构即可得到认知不确定性估计 [GG16]。类似地，深度集成等方法会训练多个网络，并利用集成方差进行不确定性量化 [LPB17]。在科学机器学习中，物理信息神经网络（PINNs）通过在物理约束下训练来融入先验知识，但除非扩展为贝叶斯表述，否则通常也无法提供校准的不确定性输出。对于逆问题和约束训练设置，这一问题尤其突出，因为推断同时受数据失配和建模假设的影响 [Cox+24；MPP24]。在这类情形下，可能同时存在偶然不确定性（数据中的固有噪声）和认知不确定性（模型不确定性） [KG17]。本文重点关注降低通过蒙特卡洛随机失活估计认知不确定性的成本。

一种标准方法（如蒙特卡洛随机失活或集成采样）是进行多次随机前向传播，以估计预测统计量。然而，计算成本会随样本数量线性增长。在大规模场景中，通过增加样本数量来降低估计量方差，可能会带来难以承受的开销。例如，蒙特卡洛随机失活通常会针对每个输入进行数十次前向传播来估计不确定性，而每次传播都需要对网络进行完整的前向评估。这使得大型或实时模型的不确定性估计成本高昂；对于需要反复评估不确定性的代理模型流程而言，这已成为实际瓶颈 [PP25]。更一般地说，如果每个样本的计算成本都很高，朴素的蒙特卡洛估计量往往代价过高 [Gil15]。

多层蒙特卡洛（MLMC）通过利用保真度层级和望远镜分解，恰好解决了这一问题 [Hei01；Gil08]。大多数样本取自低保真度（成本低）的层级，少数样本取自高保真度（成本高）的层级，从而在给定误差下减少总成本 [Gil08；Gil15]。在随机模拟（例如随机微分方程和偏微分方程）中，多层蒙特卡洛已被证明非常有效 [Cli+11]。近期研究开始将多层蒙特卡洛的思想引入机器学习；例如， [FS21] 将多层蒙特卡洛应用于贝叶斯模型中的变分推断

以及 [Ger+21] 提出了深度网络的多层训练策略。与我们的设定更为接近的是，[Bla+23] 提出了一种无偏的多层蒙特卡洛方案，以加速分类器的丢弃训练。然而，据我们所知，多层蒙特卡洛尚未用于神经代理模型的推理时不确定性量化。

本工作的目标是将实用的基于随机失活的不确定性估计与多层蒙特卡洛（MLMC）中的方差缩减工具相结合。我们构建了一个 MLMC 框架，其中蒙特卡洛 Dropout 提供随机性，将一次包含随机失活的前向评估视为一次随机模型评估，并通过对丢弃掩码进行采样来估计预测统计量。我们对保真度的主要定义是预测估计量中所使用的随机前向传播次数 T。更高的保真度对应更多的丢弃掩码，因而也意味着更高的计算成本。在此层次结构的基础上，我们通过使用共享的丢弃掩码耦合粗略与精细评估，构建多层估计量，从而降低层级修正项的方差。由此，我们分别为丢弃引起的预测期望和丢弃引起的预测方差构建了多层估计量，同时保持了针对这些随机失活诱导目标的无偏性。其他保真度定义，例如改变丢弃概率或采用不同网络分辨率的层次结构，也可以从多层视角加以理解，但这需要额外的建模选择，并非本文核心论述所必需。

我们对由此得到的估计量进行了理论分析。我们推导出明确的偏差、方差和成本表达式，确定控制层级修正项方差的量，并据此在掩码复用的耦合成本模型下，制定跨层级的样本分配策略。所得分配恢复了经典多层蒙特卡洛结构 $M_\ell \propto \sqrt{V_\ell / C_\ell}$，同时给出了由随机失活耦合 V_ℓ 所引起的、针对具体问题的层级方差表达式。我们还讨论了实用的层级设计，以及基于经验层级方差估计的停止准则。

我们在取自科学机器学习领域的基准问题上展示了该方法。首先，我们研究一个针对边界层常微分方程的正向物理信息神经网络–Uzawa 问题，该方程具有已知的闭式解；我们考察单保真度和多层随机失活估计量的表现，包括对预测抽样方差收敛率的经验验证，以及一项固定成本分配测试。其次，我们研究一个逆向偏微分方程约束优化基准，其中随机目标由一个标量噪声变量生成，且真实目标矩可用闭式形式表示。在这两种情形下，我们报告估计量方差如何随保真度阶梯和样本分配而变化，并将经验上最优的分配与理论预测的连续分配进行比较。

神经网络代理模型的不确定性研究涉及多个相互交叉的方向，而这些方向也汇聚于本文。随机失活最初是作为正则化手段提出的 [Sri+14]，在测试时可以保持启用状态，并将其解释为近似贝叶斯推断，从而无需改变架构或训练过程，即可为认知不确定性提供一种实用途径 [GG16]。相关主题还包括偶然不确定性（源于数据噪声）与认知不确定性（源于模型不确定性）之间的区别，这如今已成为不确定性感知深度学习中的标准做法 [KG17]。互补策略包括贝叶斯神经网络和深度集成，其中独立预测器能够提供基于方差的估计，并且在实践中通常表现出较强的效果[LPB17]。在科学代理模型流程中，当需要重复进行不确定性评估时，蒙特卡洛随机失活也被直接用作一种实用的不确定性量化工具 [PP25]。

多层蒙特卡洛在成本高昂的随机评估中提供了良好的成本—精度权衡。奠基性研究建立了跨耦合保真度层级的望远镜估计量，并表明，在典型的离散化设定下，在偏差、方差衰减和单样本成本满足标准假设时，可以实现诸如 $O(\varepsilon^{-3})$ 到 $O(\varepsilon^{-2})$ 的复杂度改进 [Hei01；Gil08；Gil15]。该方法已在随机微分方程和偏微分方程的随机模拟中得到广泛验证，包括具有随机系数的椭圆问题 [Cli+11]。本文将相同原理应用于神经网络预测器的推理时阶段，为蒙特卡洛随机失活引入估计量层级，并在不同保真度之间进行耦合，从而使层级差值具有较低的方差。

近期研究开始探索机器学习本身中的多层思想。相关示例包括用于随机变分推断的多层方差缩减 [FS21]，用于网络的分层训练策略[Ger+21] 以及可加速丢弃训练的无偏多层构造 [Bla+23]。此外，当有多个模拟器可用时，人们对用于基于模拟的推断的多层和多保真度方法也日益关注；研究表明，通过适当的耦合，这些方法在固定计算预算下能够提升性能 [Hik+25]。本文的独特之处在于聚焦于推理时不确定性的量化

，针对神经代理模型，多层采样旨在降低估计蒙特卡洛随机失活各阶矩的成本。

在科学机器学习领域，物理信息神经网络为求解正向与逆向问题提供了灵活的框架，但通常需要专门的机制来校准不确定性[RPK19]。为此，人们开发了贝叶斯变体和多保真不确定性感知代理模型 [YMK21；MBK21]。在我们的设定中，将蒙特卡洛随机失活估计量视为分层评估可使 PINN 代理模型受益，从而以显著更少的高成本前向传播计算随机失活引起的认知不确定性。

总体而言，本文将用于认知不确定性的蒙特卡洛随机失活 [GG16], 、科学应用中的不确定性感知代理建模 [RPK19；MBK21] 以及用于提高计算效率的多层蒙特卡洛[Gil15；FS21]结合起来。由此得到的框架旨在降低不确定性估计的计算成本，并表明多层采样能够以较低的预算，为神经代理模型提供准确的基于随机失活的不确定性量化。

本文其余部分安排如下。在§2 中，我们定义由随机失活产生的关注目标，并介绍基于耦合内层保真度、用于预测均值和方差的多层蒙特卡洛估计量。在§3 中，我们给出实用的耦合抽样算法，定义掩码复用条件下的耦合成本模型，并推导固定计算预算下各层级的最优样本分配。在§4中，我们通过正向与逆向 PINNs–Uzawa 基准测试验证理论预测，其中包括经验方差率研究和固定成本分配测试。在§5 中，我们总结与可复现性和计算复杂度相关的实现细节。最后在§6作总结。附录汇集了分析中所用技术引理的证明，包括重叠样本方差估计量的协方差恒等式以及分配公式的证明。

2. Multilevel estimators for MC-dropout moments

本节将蒙特卡洛随机失活推断表述为对随机化网络输出的矩进行数值估计。从这一视角来看，计算瓶颈十分明显：不确定性估计需要进行许多次随机前向传播。随后，我们将构建多层蒙特卡洛估计量，在针对相同随机失活诱导量的同时降低这一成本。

2.1. 随机失活诱导的随机预测器。 我们首先确定一个由训练后的前馈神经网络表示的确定性代理模型。具体而言，对于输入 $x \in \mathbb{R}^{N_{\text{in}}}$，该网络通过仿射映射和非线性函数的复合，在 $\mathbb{R}^{N_{\text{out}}}$ 中产生输出。例如，记作 $z_0 := x$，并且对于$k = 0, \ldots\ldots, K-1$,

$$z_{k+1} := \sigma\big(W_k z_k + b_k\big), \qquad \mathscr{D}(x;\hat{\theta}) := W_K z_K + b_K, \tag{1}$$

其中 σ 是适当的激活函数，$\hat{\theta} := \{(W_k,\ b_k)\}_{k=0}^{K} \in \Theta$ 表示训练后的权重和偏置，并在本节中始终保持不变。图1展示了这一确定性评估的示意图。

在蒙特卡洛随机失活中，评估时仍保持随机失活启用。因此，每次前向传播都会随机抽取一组丢弃掩码，并将其应用于选定的层激活值。对于给定的 $\theta \in \Theta$，我们将所有此类测试时随机性的来源汇总为一个随机元素 $\theta \sim \pi_{\hat{\theta}}$，其中 $\pi_{\hat{\theta}}$ 是由所选丢弃率和参数 $\theta \in \Theta$ 所诱导的相应概率律。

在蒙特卡洛随机失活这一特定情形下，对于固定的已学习参数 θ，随机元素 θ 可视为单次前向传播期间应用于所有 Dropout 层的伯努利掩码集合。具体而言，对于宽度为 N_{width} 的随机失活层，我们抽取一个掩码

$$\boldsymbol{B} = (B_1, \ldots, B_{N_{\text{width}}})^{\top}, \qquad B_j \overset{\text{iid}}{\sim} \text{Bernoulli}(1 - p_{\text{drop}}), \tag{2}$$

wh其中 $p_{\text{drop}} \in$ (0,1) 表示丢弃概率。典型的随机失活层随后按如下方式作用：

$$z_{k+1} := \boldsymbol{B}^{(k)} \odot \sigma\big(W_k z_k + b_k\big), \tag{3}$$

其中 $\odot$表示逐分量乘法。（任何标准的 Dropout 缩放约定，例如反向 Dropout，均已纳入 $\mathscr{D}(x;\theta)$的定义中。）

收集随机失活掩码的各个实例$\{\boldsymbol{B}^{(k)}\}_{k=0}^{L-1}$并将其与训练后的权重$\hat{\theta} \in \Theta$配对，即得到$\theta \sim \pi_{\hat{\theta}}$的一个实现。重复的随机前向传播对应于独立抽取$\theta_1$、$\theta_2$.....$\overset{\text{iid}}{\sim} \pi_{\hat{\theta}}$，其中$\hat{\theta}$保持固定。为简化记号，我们将所得的随机化预测器记为$\hat{\mathscr{D}}(x;\ \theta)$，其中$\theta$保持固定，而$\theta$仅编码测试时随机性。这种随机评估如图2所示。

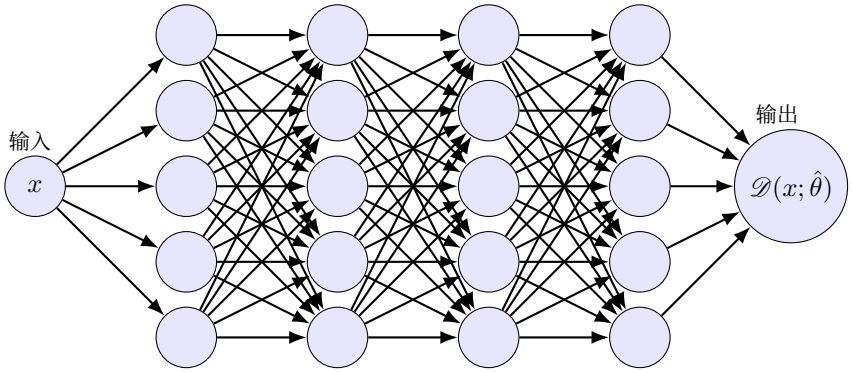


FIGURE 1. 代理网络架构的确定性形式示意图。$\hat{}$训练参数θ固定不变，重复评估会产生相同的输出$\mathscr{D}(x;\theta)$。

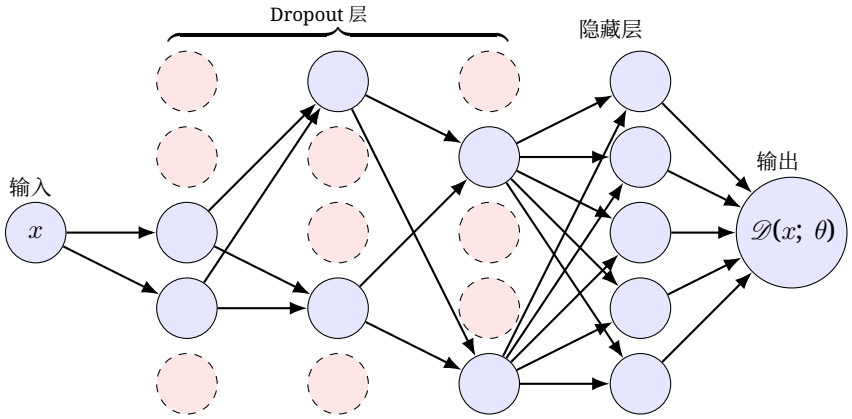


FIGURE 2. 同一个网络在启用随机失活的情况下进行评估。每次前向传播都会抽取丢弃掩码（以$\theta \sim \pi_{\hat{\theta}}$编码），从而得到一个随机化输出$\mathscr{D}(x;\theta)$。重复评估时会重新抽取$\theta$，因此可能产生不同的输出。

在本节中，我们均以标量输出情形$N_{\text{out}} = 1$给出矩恒等式，其中$\mathscr{D}(x;\ \theta) \in \mathrm{R}$，且$\mu_2(x)$表示预测方差。对于向量值输出$\mathscr{D}(x;\theta) \in \mathrm{R}^{N_{\text{out}}}$，将下文的定义应用于$\mathscr{D}(x;\theta)$的每个坐标，即可逐分量得到相同的构造和结果。

我们关注该随机化预测器相对于随机失活所引入的认知随机性的前两个中心矩，记为

$$\mu(x) := \mathbb{E}_\theta\big[\mathscr{D}(x;\theta)\big], \qquad \mu_2(x) := \mathrm{Var}_\theta\big[\mathscr{D}(x;\theta)\big] = \mathbb{E}_\theta\Big[\big(\mathscr{D}(x;\theta) - \mu(x)\big)^2\Big]. \tag{4}$$

我们使用记号μ，μ_2来强调这些是固定x时丢弃诱导的预测分布的矩。这个选择在后文引入样本估计量及其抽样方差时很方便，因为对多个不同的方差都使用σ^2可能会造成歧义。

在.....的分析中　对于方差估计量，我们还需要四阶中心矩，记作

$$\mu_4(x) := \mathbb{E}_\theta\Big[\big(\mathscr{D}(x;\theta) - \mu(x)\big)^4\Big]. \tag{5}$$

这个矩决定了经验方差估计量的抽样变异性，因此也会影响分配和复杂度方面的结果。

$\mathscr{D}(x;\ \theta)$ 中的随机性反映了由数据有限和模型不确定性引起的认知不确定性，应将其与偶然不确定性区分开来；后者表示数据生成过程中不可约的变异或系统性影响。丢弃诱导的预测分布并非精确的贝叶斯预测分布，因为随机失活仅为贝叶斯推断提供了变分近似 [GG16]。一个已知的局限是，由此得到的不确定性估计可能校准不当，例如，这可能是架构选择、丢弃率或未建模的偶然噪声所致 [Guo+17]。尽管如此，蒙特卡洛随机失活的计算开销较低，因此被广泛用作估计认知不确定性的实用方法 [Gal16；KG17]。

2.2. **单保真度蒙特卡洛估计量**。 为了在固定输入 处近似丢弃诱导的矩（4），一种自然的方法是对测试时随机性的独立实现进行蒙特卡罗抽样。设 $\{\theta_t\}_{t=1}^T$ 为来自 $\pi_{\hat\theta}$ 的独立同分布抽样，其中 $\pi_{\hat\theta}$ 是固定权重 $\hat\theta$ 时由随机失活诱导的概率分布，记作$\theta_t \overset{\text{iid}}{\sim} \pi_{\hat\theta}$。为简化记号，我们省略相关概率分布 $\pi_{\hat\theta}$ 对训练后的权重 $\hat\ \theta \in \Theta$ 的显式依赖，并将该分布记为 π。随后，我们定义预测均值和预测方差的单保真度（单 -T ）估计量如下：

$$Y(x;T) := \frac{1}{T}\sum_{t=1}^{T}\mathscr{D}(x;\theta_t), \qquad V(x;T) := \frac{1}{T-1}\sum_{t=1}^{T}\big(\mathscr{D}(x;\theta_t) - Y(x,T)\big)^2. \tag{6}$$

这里 T 是保真度参数，较大的 T 会带来更低的抽样变异性，但需要更多的随机前向传播。

2.2.1. 单保真度估计量的矩。 估计量 $Y(x;\ \ T)$ 和 $V(x;\ \ T)$ 是随机变量，因为它们依赖于随机抽样 $\{\theta_t\}_{t=1}^T$。它们的期望恢复了目标矩（4），其抽样方差则量化了在保真度 T 下进行单次蒙特卡洛估计的准确性。

引理 2.3（单保真度估计量的矩）。 如果 $\theta_t \overset{\text{iid}}{\sim} \pi$ 且 $T \geq 2$，则

$$\begin{aligned}
\mathbb{E}\big[Y(x;T)\big] &= \mu(x), \quad \operatorname{Var}\big[Y(x;T)\big] = \frac{1}{T}\mu_2(x), \\
\mathbb{E}\big[V(x;T)\big] &= \mu_2(x), \quad \operatorname{Var}\big[V(x;T)\big] = \frac{1}{T}\left(\mu_4(x) - \frac{T-3}{T-1}\mu_2^2(x)\right),
\end{aligned} \tag{7}$$

其中 $\mu_4(x)$ 定义见 (5)。

2.3.2. 外层抽样与经验估计量。在应用中，我们通常不仅希望量化估计量 $Y(x;\ T)$ 和 $V(x;\ T)$ 本身，还希望量化它们的抽样变异性 $\operatorname{Var}[Y(x;\ T)]$ 和 $\operatorname{Var}[V(x;\ T)]$。一种直接的经验方法是抽取 M 个估计量（6）的独立重复样本。也就是说，对于 m =1, . . . , M，令 $\{\theta_t^{(m)}\}_{t=1}^T$ 为来自 π 的独立同分布抽样，并定义

$$Y^{(m)}(x;T) := \frac{1}{T}\sum_{t=1}^{T}\mathscr{D}(x;\theta_t^{(m)}), \qquad V^{(m)}(x;T) := \frac{1}{T-1}\sum_{t=1}^{T}\big(\mathscr{D}(x;\theta_t^{(m)}) - Y^{(m)}(x;T)\big)^2.$$

然后，我们构造经验估计量

$$\begin{aligned}
\overline{Y}(x;M,T) &:= \frac{1}{M}\sum_{m=1}^{M}Y^{(m)}(x,T), \quad \mathcal{S}_Y^2(x;M,T) := \frac{1}{M-1}\sum_{m=1}^{M}\big(Y^{(m)}(x,T) - \overline{Y}(x;M,T)\big)^2, \\
\overline{V}(x;M,T) &:= \frac{1}{M}\sum_{m=1}^{M}V^{(m)}(x,T), \quad \mathcal{S}_V^2(x;M,T) := \frac{1}{M-1}\sum_{m=1}^{M}\big(V^{(m)}(x,T) - \overline{V}(x;M,T)\big)^2.
\end{aligned} \tag{8}$$

单保真度构造需要对 $\mathscr{D}(x;\ \theta)$ 进行 MT 次评估，因此其成本随 M 和 T 线性增长。因此，要使经验方差 $\mathcal{S}_Y^2(x;$ $M,T)$ 和 $\mathcal{S}_V^2(x;\ M,T)$ 较小，可能会变得

计算成本高昂。这促使我们在下一小节介绍多层构造，它们旨在通过结合多种保真度，在达到预定精度的同时降低成本。

2.4. 预测均值的多层估计量。 现在，我们为随机失活引起的预测均值 $\mu(x)$ 构造一个多层蒙特卡洛（MLMC）估计量。其核心思想是结合保真度逐步提高的估计量 $T_0 < \cdots < T_L$：使用大量低成本的低保真度样本和少量高成本的高保真度样本，同时耦合相邻保真度，以降低层级差值的方差。

备注 2.5 (估计量层级，而非离散化层级)。 不同于针对有偏离散化的经典 *MLMC*，估计量 $Y(x, T_\ell)$ 在 (6) 中定义，对 $\mu(x)$ 而言，任意 ℓ下均为无偏估计。保真度参数 T_ℓ 控制抽样方差和成本，而非建模偏差。因此，多层构造是一种结构化的控制变量方案，由针对同一目标矩的耦合估计量构成。

2.5.1. 跨保真度耦合。 固定一个严格递增的保真度序列

$$1 \le T_0 < T_1 < \cdots < T_L, \tag{9}$$

并令 $\boldsymbol{T} := (T_0, \dots, T_L)$ 和 $\boldsymbol{M} := (M_0, \dots, M_L)$ 分别表示保真度阶梯和对应的样本数（每层）。MLMC 构造在每个层级差值内耦合相邻保真度：对于固定的层级 $\ell \ge 1$ 和外层索引 $m \in [1, M_\ell]$，我们通过在 T_ℓ 个样本的均值中重复使用前 $T_{\ell-1}$ 次随机失活实现，并仅生成 $T_\ell - T_{\ell-1}$ 次额外抽样，来构造一个耦合对 $\big(Y^{(m)}(x, T_{\ell-1})\ Y^{(m)}(x, T_\ell)\big)$。不同外层重复 m 之间的随机失活实现相互独立，因此对于任意固定的 ℓ，只要 $m_1 \ne m_2$，估计量 $Y^{(m_1)}(x; T_\ell)$ 和 $Y^{(m_2)}(x; T_\ell)$ 就相互独立；而对于给定的 m，在保真度 $T_{\ell-1}$ 下使用的样本会作为前缀，在 T_ℓ 下重复使用。等价地，若 $\theta_t^{(m)}$ 独立同分布 $\sim \pi$，则

$$T_\ell Y^{(m)}(x; T_\ell) = T_{\ell-1} Y^{(m)}(x; T_{\ell-1}) + \sum_{t=T_{\ell-1}+1}^{T_\ell} \mathscr{D}\big(x; \theta_t^{(m)}\big), \qquad m = 1, \dots, M_\ell, \quad \ell = 1, \dots, L. \tag{10}$$

我们将耦合层增量定义为

$$\Delta Y^{(m)}(x; T_\ell) := Y^{(m)}(x; T_\ell) - Y^{(m)}(x; T_{\ell-1}), \qquad m = 1, \dots, M_\ell, \quad \ell = 1, \dots, L. \tag{11}$$

耦合确保 $Y^{(m)}(x;\ T_\ell)$ 和 $Y^{(m)}(x;\ T_{\ell-1})$ 之间具有很强的相关性，因此 $\Delta Y^{(m)}(x;\ T_\ell)$ 的方差通常远小于任一单独项的方差。特别地，望远镜恒等式成立[Gil08]，

$$\mathbb{E}\big[Y^{(m)}(x; T_L)\big] = \mathbb{E}\big[Y^{(m)}(x; T_0)\big] + \sum_{\ell=1}^{L} \mathbb{E}\big[\Delta Y^{(m)}(x; T_\ell)\big] \qquad m = 1, \dots, M_L. \tag{12}$$

2.5.2. 多层蒙特卡洛均值估计量及其方差估计量。 受（12）启发，我们将 $\mu(x)$ 的多层蒙特卡洛估计量定义为

$$\mathcal{Y}(x; \boldsymbol{M}, \boldsymbol{T}) := \overline{Y}(x; M_0, T_0) + \sum_{\ell=1}^{L} \overline{\Delta Y}(x; M_\ell, T_\ell), \qquad \overline{\Delta Y}(x; M_\ell, T_\ell) := \frac{1}{M_\ell} \sum_{m=1}^{M_\ell} \Delta Y^{(m)}(x; T_\ell). \tag{13}$$

其中，对于每个 ℓ，样本 $\{\Delta Y^{(m)}(x;\ T_\ell)\}_{m=1}^{M_\ell}$ 在 m 上是独立同分布的。我们使用经验方差估计量来量化（13）的抽样方差

$$S_Y^2(x; \boldsymbol{M}, \boldsymbol{T}) := \frac{1}{M_0} \mathcal{S}_Y^2(x; M_0, T_0) + \sum_{\ell=1}^{L} \frac{1}{M_\ell} \mathcal{S}_{\Delta Y}^2(x; M_\ell, T_\ell), \tag{14}$$

其中 $\mathcal{S}_Y^2(x;\ M_0,\ T_0)$ 定义见 (8)，并且

$$\mathcal{S}_{\Delta Y}^2(x; M_\ell, T_\ell) := \frac{1}{M_\ell - 1} \sum_{m=1}^{M_\ell} \big(\Delta Y^{(m)}(x; T_\ell) - \overline{\Delta Y}(x; T_\ell)\big)^2, \tag{15}$$

2.5.3. 矩恒等式。 MLMC估计量 (13) 对 $\mu(x)$ 是无偏的，而经验方差估计量 (14) 的期望值等于方差 μ_2 乘以一个分别依赖于 $\boldsymbol{T}$ 和 $\boldsymbol{M}$ 的函数。

引理 2.6 （MLMC均值估计量的矩）。 设 $\boldsymbol{T} = (T_0, \dots, T_L)$ 满足 (9)，并假设对每个层级 ℓ，外层样本 $\{\Delta Y_{(m)}(x; T_\ell)\}_{m=1}^{M_\ell}$ 均为独立同分布。则

$$\mathbb{E}\big[\mathcal{Y}(x; \boldsymbol{M}, \boldsymbol{T})\big] = \mu(x), \tag{16}$$

and

$$\mathbb{E}\big[S_Y^2(x; \boldsymbol{M}, \boldsymbol{T})\big] = \mu_2(x) \left(\frac{1}{M_0 T_0} + \sum_{\ell=1}^{L} \frac{1}{M_\ell} \left(\frac{1}{T_{\ell-1}} - \frac{1}{T_\ell} \right) \right). \tag{17}$$

2.7. **预测方差的多层估计量**。 接下来，我们针对丢弃法诱导的预测方差 $\mu_2(x)$ 构造一个MLMC估计量。与单保真度设置相同，我们使用基于 T 次随机前向传播的无偏样本方差 $V(x, T)$ 来估计 $\mu_2(x)$。多层构造与均值的构造类似，但需要在不同保真度间谨慎地耦合方差估计量。

2.7.1. 通过在线方差更新实现耦合。 固定 $2 \le T_0 < T_1 < \cdots < T_L$，并设 $\{\theta_t\}_{t=1}^{T_L}$ 为来自 π 的独立同分布抽样。此外，假设对 $\ell = 1, \dots, L$， $T_\ell - T_{\ell-1} \ge 2$ 成立，从而每个新数据块上的无偏样本方差都有定义。为方便记号，写作

$$D_t(x) := \mathscr{D}(x; \theta_t), \qquad t = 1, \dots, T_L,$$

并对整数 $1 \le a < b$ 定义，

$$\overline{D}_{a:b}(x) := \frac{1}{b-a+1} \sum_{t=a}^{b} D_t(x), \qquad V_{a:b}(x) := \frac{1}{b-a} \sum_{t=a}^{b} \big(D_t(x) - \overline{D}_{a:b}(x)\big)^2.$$

采用此记号，式（6）中的单保真度方差估计量为 $V(x, T) = V_{1:T}(x)$.

为耦合 $V(x, T_{\ell-1})$ 和 $V(x, T_\ell)$，我们复用 $T_{\ell-1}$-样本估计中的前 T_ℓ 个随机失活抽样，并仅额外生成 $T_\ell - T_{\ell-1}$ 个新抽样。由此得到的耦合方差满足合并方差恒等式（等价地，也可通过在线方差更新计算） [CGL79]:

$$\begin{aligned}(T_\ell - 1)V(x, T_\ell) = (T_{\ell-1} - 1)V(x, T_{\ell-1}) + (T_\ell - T_{\ell-1} - 1)V_{T_{\ell-1}+1:T_\ell}(x) \\ + \frac{T_{\ell-1}(T_\ell - T_{\ell-1})}{T_\ell} \Big(\overline{D}_{1:T_{\ell-1}}(x) - \overline{D}_{T_{\ell-1}+1:T_\ell}(x)\Big)^2.\end{aligned} \tag{18}$$

我们将耦合层增量定义为

$$\Delta V(x; T_\ell) := V(x; T_\ell) - V(x; T_{\ell-1}), \qquad \ell = 1, \dots, L. \tag{19}$$

由于 $V(x; T_{\ell-1})$ 和 $V(x; T_\ell)$ 均为 $\mu_2(x)$ 的无偏估计量（引理 2.3），因此有 $\mathbb{E}[\Delta V(x; T_\ell)] = 0$，而 MLMC 构造同样依赖于耦合差值所带来的方差降低。

2.7.2. *MLMC* 方差估计量及其方差估计量。 设 $\boldsymbol{T} = (T_0, \dots, T_L)$ 和 $\boldsymbol{M} = (M_0, \dots, M_L)$。对于每个层级 ℓ，我们生成 M_ℓ 个独立同分布的耦合增量外层重复样本 $\Delta V(x; T_\ell)$，并将 $\mu_2(x)$ 的 MLMC 估计量定义为

$$\mathcal{V}(x; \boldsymbol{M}, \boldsymbol{T}) := \overline{V}(x; M_0, T_0) + \sum_{\ell=1}^{L} \overline{\Delta V}(x; M_\ell, T_\ell), \quad \overline{\Delta V}(x; M_\ell, T_\ell) := \frac{1}{M_\ell} \sum_{m=1}^{M_\ell} \Delta V^{(m)}(x; T_\ell). \tag{20}$$

我们使用经验方差估计量来量化（20）的抽样方差

$$S_V^2(x; \boldsymbol{M}, \boldsymbol{T}) := \frac{1}{M_0} \mathcal{S}_V^2(x; M_0, T_0) + \sum_{\ell=1}^{L} \frac{1}{M_\ell} \mathcal{S}_{\Delta V}^2(x; M_\ell, T_\ell), \tag{21}$$

其中 $\mathcal{S}_V^2(x;\ M_0,\ T_0)$ 在 (8) 中定义，且

$$\mathcal{S}_{\Delta V}^2(x; M_\ell, T_\ell) := \frac{1}{M_\ell - 1} \sum_{m=1}^{M_\ell} \big(\Delta V^{(m)}(x; T_\ell) - \overline{\Delta V}(x; M_\ell, T_\ell)\big)^2. \tag{22}$$

2.7.3. 矩恒等式。 MLMC估计量 (20) 对 $\mu_2(x)$ 是无偏的，而经验方差估计量 (21) 的期望值仅取决于 2nd 阶和 4th 阶矩， $\boldsymbol{M}$，以及 $\boldsymbol{T}$。

引理 *2.8*（重叠样本无偏样本方差估计量的协方差）。固定 x ，令$\{\theta_i\}_{i=1}^{T_\ell}$ 为来自 π的独立同分布抽样。假设 $\mathscr{D}(x;\theta)$具有有限四阶中心矩，并将式（*4*）和（*5*）中定义的二阶和四阶中心矩分别记为 $\mu_2(x)$和 $\mu_4(x)$。对于 $\ell \geq 1$，用前$T_{\ell-1}$ 个样本构造 $V(x;T_{\ell-1})$，并用前 T_ℓ 个样本构造 $V(x;T_\ell)$。则

$$\text{(23)}\qquad \operatorname{Cov}\big[V(x;T_\ell),V(x;T_{\ell-1})\big]=\frac{T_{\ell-1}-1}{T_\ell-1}\operatorname{Var}[V(x;T_{\ell-1})]+\frac{T_\ell-T_{\ell-1}}{T_{\ell-1}T_\ell(T_\ell-1)}\big(\mu_4(x)-3\mu_2^2(x)\big).$$

定理 2.9 （MLMC方差估计量的矩）。 设 $\boldsymbol{T}=(T_0,\dots,T_L)$ 满足 $T_0\geq 2$，并假设对于每个层级 ℓ，外层样本 $\{\Delta V_{(m)}(x;T_\ell)\}_{m=1}^{M_\ell}$ 均为独立同分布。则

$$\text{(24)}\qquad \mathbb{E}\big[\mathcal{V}(x;\boldsymbol{M},\boldsymbol{T})\big]=\mu_2(x),$$

and

$$\begin{aligned}\text{(25)}\qquad \mathbb{E}\big[S_V^2(x;\boldsymbol{M},\boldsymbol{T})\big]&=\frac{1}{M_0T_0}\left(\mu_4(x)-\frac{T_0-3}{T_0-1}\mu_2^2(x)\right)+\sum_{\ell=1}^{L}\frac{1}{M_\ell}\left(\frac{1}{T_{\ell-1}}-\frac{1}{T_\ell}\right)\big(\mu_4(x)-3\mu_2^2(x)\big)\\&\quad+\sum_{\ell=1}^{L}\frac{2}{M_\ell}\left(\frac{1}{T_{\ell-1}-1}-\frac{1}{T_\ell-1}\right)\mu_2^2(x),\end{aligned}$$

其中 $\mu_4(x)$ 定义见(5)。

证明。 无偏性（24）由期望的线性性和引理 2.3 得出： $\mathbb{E}[V(x;T_0)]=\mu_2(x)$ 且$\mathbb{E}[\Delta V(x;T_\ell)]=0$。对于抽样方差，根据期望的线性性，我们有

$$\text{(26)}\qquad \mathbb{E}[S_V^2(x;\boldsymbol{M},\boldsymbol{T})]=\frac{1}{M_0}\mathbb{E}[\mathcal{S}_V^2(x;M_0,T_0)]+\sum_{\ell=1}^{L}\frac{1}{M_\ell}\mathbb{E}[\mathcal{S}_{\Delta V}^2(x;M_\ell,T_\ell)],$$

$\mathcal{S}_V^2$和 $\mathcal{S}_{\Delta V}^2$的定义是，它们分别是 V和 ΔV的无偏方差估计量。假设外层样本独立同分布，则无偏方差估计量的期望等于方差，即 $\mathrm{E}[\mathcal{S}_Z^2]=\mathsf{Var}[Z]$，因此

$$\text{(27)}\qquad \mathbb{E}[S_V^2(x;\boldsymbol{M},\boldsymbol{T})]=\frac{1}{M_0}\operatorname{Var}\big[V(x;T_0)\big]+\sum_{\ell=1}^{L}\frac{1}{M_\ell}\operatorname{Var}\big[\Delta V(x;T_\ell)\big].$$

展开求和项，可得以下表达式

$$\operatorname{Var}\big[\Delta V(x;T_\ell)\big]=\operatorname{Var}\big[V(x;T_{\ell-1})\big]-2\operatorname{Cov}\big[V(x;T_{\ell-1}),V(x;T_\ell)\big]+\operatorname{Var}\big[V(x;T_\ell)\big].$$

第一项和最后一项由引理2.3给出。对于 $\ell\geq 1$，耦合会重用 $T_{\ell-1}$ 个随机前向传播，这些传播用于 T_ℓ样本方差。因此，将引理2.8 应用于嵌套对$\big(V(x;T_{\ell-1})\ V(x;T_\ell)\big)$，可得闭式表达式:

$$\operatorname{Var}\big[\Delta V(x;T_\ell)\big]=\left(\frac{1}{T_{\ell-1}}-\frac{1}{T_\ell}\right)\big(\mu_4(x)-3\mu_2^2(x)\big)+2\left(\frac{1}{T_{\ell-1}-1}-\frac{1}{T_\ell-1}\right)\mu_2^2(x),$$

将其代入上述方差分解后，可得（25）。 □

注记2.10。 ((25) 右侧非负，因为根据构造，它等于 方差$\big[\mathcal{V}(x;\boldsymbol{M},\boldsymbol{T})\big]$。此外，在超额峰度为零的假设下，$\mu_4(x)=3\mu_2^2(x)$，表达式(25) 可简化为

$$\text{(28)}\qquad \mathbb{E}\big[S_V^2(x;\boldsymbol{M},\boldsymbol{T})\big]=2\mu_2^2(x)\left(\frac{1}{M_0}\frac{1}{T_0-1}+\sum_{\ell=1}^{L}\frac{1}{M_\ell}\left(\frac{1}{T_{\ell-1}-1}-\frac{1}{T_\ell-1}\right)\right).$$

具体而言，对于几何层级（例如 $T_\ell=2^\ell T_0$），单层校正方差按 $O(T_{\ell-1}^{-2})$ 衰减，而最优分配 $\boldsymbol{M}$ 仅取决于 $\boldsymbol{T}$，因为对底层分布的依赖通过公共因子 $2\mu_2^2(x)$ 体现。

3. Computation and allocation

在本节中，我们将第 2 节中的估计量转化为实用的耦合抽样流程，然后讨论如何在各保真度层级之间分配样本，以便在给定计算预算的情况下最小化抽样方差。

3.1. 耦合抽样算法。 预测均值和方差的多层估计量依赖于每个层级增量内的耦合评估。对于层级 $\ell \geq 1$，我们从相同的底层随机失活实现中计算 $Y(x;\ T_{\ell-1})$ 和 $Y(x;\ T_\ell)$（以及类似的 $V(x;\ T_{\ell-1})$ 和 $V(x;\ T_\ell)$），并重用 T_ℓ 个样本估计量中的前$T_{\ell-1}$ 次随机前向传播。算法 1 实现了这种耦合，并返回 MLMC估计量 $\mathcal{Y}(x;\ \boldsymbol{M},\boldsymbol{T})$ 和 $\mathcal{V}(x;\ \boldsymbol{M},\boldsymbol{T})$，以及式（13）–（21）中定义的经验方差估计量$S_Y^2(x;\ \boldsymbol{M},\boldsymbol{T})$ 和 $S_V^2(x;\ \boldsymbol{M},\boldsymbol{T})$。

算法 1 用于多层蒙特卡洛均值和方差估计量的耦合抽样

要求： 输入 x；保真度阶梯 $\boldsymbol{T} = (T_0, \ldots, T_L)$，其中 $2 \leq T_0 < \cdots < T_L$；样本数 $\boldsymbol{M} = (M_0, \ldots, M_L)$，其中 $M_\ell \geq 1$，且在计算样本方差时 $M_\ell \geq 2$

确保： 样本 $\{Y^{(m)}(x;\ T_0),\ V^{(m)}(x;\ T_0)\}$ 以及增量 $\{\Delta Y^{(m)}(x;\ T_\ell),\ \Delta V^{(m)}(x;\ T_\ell)\}$；估计量 $\mathcal{Y},\ \mathcal{V}$ 以及经验方差 $S_Y^2,\ S_V^2$

初始化数组 $\mathrm{Y}_0 \leftarrow []$、$\mathrm{V}_0 \leftarrow []$
对于 $m = 1$ 到 M_0，**执行**
    抽取 T_0 个独立同分布的 $\theta^{(m)} \sim \pi$，并存储评估结果 $\mathrm{D}^{(m)} = [D_1^{(m)}, \ldots, D_{T_0}^{(m)}]$，其中 $D_t^{(m)} = \mathscr{D}(x;\ \theta_t^{(m)})$
    计算 $Y^{(m)} \leftarrow$ 均值$(\mathrm{D}^{(m)})$
    计算 $V^{(m)} \leftarrow$ 样本方差$(\mathrm{D}^{(m)})$
    将 $Y^{(m)}$ 追加到 Y_0，并将 $V^{(m)}$ 追加到 V_0
结束循环
初始化数组 $\mathrm{dY}_\ell \leftarrow []$、$\mathrm{dV}_\ell \leftarrow []$，用于 $\ell = 1, \ldots, L$
对于 $\ell = 1$ 到 L，**执行**
    对于 $m = 1$ 到 M_ℓ **执行**
        抽取 $T_\ell - T_{\ell-1}$ 独立同分布 $\theta^{(m)} \sim \pi$ 并附加评估结果 $[D_{T_{\ell-1}+1}^{(m)}, \ldots, D_{T_\ell}^{(m)}]$ 到 $\mathrm{D}^{(m)}$
        计算 $Y_{(m)\mathrm{c}} \leftarrow$ 均值$([D_1^{(m)}, \ldots, D_{T_{\ell-1}}^{(m)}])$ 以及 $Y_{(m)\,\mathrm{f}} \leftarrow$ 均值$([D_1^{(m)}, \ldots, D_{T_\ell}^{(m)}])$
        计算 $V_{(m)\mathrm{c}} \leftarrow$ 样本方差$([D_1^{(m)}, \ldots, D_{T_{\ell-1}}^{(m)}])$ 以及 $V_{(m)\,\mathrm{f}} \leftarrow$ 样本方差$([D_1^{(m)}, \ldots, D_{T_\ell}^{(m)}])$
        将 $Y_{(m)\,\mathrm{f}} - Y_{(m)\mathrm{c}}$ 附加到 dY_ℓ
        将 $V^{(m)\,\mathrm{f}} - V^{(m)\,\mathrm{c}}$ 添加到 dV_ℓ
    结束循环
结束循环
计算 $\mathcal{Y} \leftarrow$ 均值$(\mathrm{Y}_0) + \sum_{\ell=1}^{L}$均值$(\mathrm{dY}_\ell)$
计算 $\mathcal{V} \leftarrow$ 均值$(\mathrm{V}_0) + \sum_{\ell=1}^{L}$均值$(\mathrm{dV}_\ell)$
计算 $S_Y^2 \leftarrow$ 样本方差$(\mathrm{Y}_0)/M_0 + \sum_{\ell=1}^{L}$样本方差$(\mathrm{dY}_\ell)/M_\ell$
计算 $S_V^2 \leftarrow$ 样本方差$(\mathrm{V}_0)/M_0 + \sum_{\ell=1}^{L}$样本方差$(\mathrm{dV}_\ell)/M_\ell$

3.2. 成本模型。 为了比较单保真度和多层策略，我们需要一个成本模型。全文中，我们以训练好的网络的随机前向传播次数来衡量成本，也就是说，以独立抽取 $\theta \sim \pi$时对 $\mathscr{D}(x;\theta)$的总调用次数来衡量。当主要开销来自网络评估本身，且与之相比，计算均值和方差所需的归约运算等额外开销可以忽略不计时，这种替代指标是恰当的。

3.2.1. 非耦合成本基准。 一个有用的基准是：在构造每个增量时，若粗层与细层估计量独立生成，所需的成本。在这种非耦合设置下，0 级需要 M_0T_0次评估；对于每个 $\ell \geq 1$，还需为精细估计量生成 T_ℓ 次评估，因此得到 $M_\ell T_\ell$

ℓ级的评估次数。因此，非耦合总成本为

$$c_{\mathrm{unc}} := M_0 T_0 + \sum_{\ell=1}^{L} M_\ell T_\ell, \tag{29}$$

3.2.2. 掩码复用下的有效成本。 在常见情形 $M_0 \geq \cdots \geq M_L$ 中，一个耦合的单个 ℓ 层估计量会抽取 T_ℓ 个随机失活实现，并复用其中前 $T_{\ell-1}$ 个来计算粗估计量。因此，一个耦合的第 ℓ 层重复实验需要进行 $T_\ell - T_{\ell-1}$ 次评估，而非 T_ℓ 次。有效耦合成本为

$$c_{\mathrm{cpl}} = T_0 M_0 + \sum_{\ell=1}^{L} (T_\ell - T_{\ell-1}) M_\ell = \sum_{\ell=0}^{L-1} T_\ell (M_\ell - M_{\ell+1}) + T_L M_L. \tag{30}$$

因此 $c_{\mathrm{cpl}} \leq c_{\mathrm{unc}}$。在进行分配时，我们采用耦合预算 $c := c_{\mathrm{cpl}}$。

3.3. **各层级的最优分配。** 第 2 节中的多层蒙特卡洛估计量取决于样本数 $\boldsymbol{M} = (M_0, \ldots, M_L)$ 的选择。本小节将提出并求解分配问题：在计算预算固定的情况下，选择 $\boldsymbol{M}$ 以使目标估计量的抽样方差最小。我们采用耦合成本 $c := c_{\mathrm{cpl}}$，其形式见式（30）。优化先针对连续的 M_ℓ 进行，然后在实现时将其取整为整数。

3.3.1. 均值估计量的分配。 多层蒙特卡洛均值估计量的经验抽样方差的期望由引理 2.6明确给出。记

$$v_0(x) := \mathrm{Var}\big[Y(x; T_0)\big] = \frac{\mu_2(x)}{T_0}, \qquad v_\ell(x) := \mathrm{Var}\big[\Delta Y(x; T_\ell)\big] = \mu_2(x) \left(\frac{1}{T_{\ell-1}} - \frac{1}{T_\ell} \right), \ \ell \geq 1, \tag{31}$$

我们有，期望估计量的经验抽样方差的期望为

$$\mathbb{E}[S_Y^2(x; \boldsymbol{M}, \boldsymbol{T})] = \frac{v_0(x)}{M_0} + \sum_{\ell=1}^{L} \frac{v_\ell(x)}{M_\ell}. \tag{32}$$

在满足预算约束的条件下最小化（32）$c = T_0 M_0 + \sum_{\ell=1}^{L} (T_\ell - T_{\ell-1}) M_\ell$，可得引理 3.4中的分配。

引理 3.4（均值估计量的最优分配）。 固定 x 以及一个严格递增的保真度阶梯$2 \leq T_0 < \cdots < T_L$。令 $a_0 := T_0$和 $a_\ell := T_\ell - T_{\ell-1}$，其中 $\ell \geq 1$，并令 $c > 0$ 为耦合成本预算$c = \sum_{\ell=0}^{L} a_\ell M_\ell$。在满足此预算约束的连续选择 $M_\ell > 0$ 中，抽样方差$\mathbb{E}[S_Y^2(x; \boldsymbol{M}, \boldsymbol{T})]$ 在以下条件下达到最小

$$M_\ell^\star = c \frac{\sqrt{v_\ell(x)/a_\ell}}{\sum_{k=0}^{L} \sqrt{v_k(x) a_k}}, \qquad \ell = 0, \ldots, L. \tag{33}$$

等价地，代入显式的 v_ℓ 和 a_ℓ 可得

$$M_0^\star = \frac{c}{T_0} \left(1 + \sum_{\ell=1}^{L} \frac{T_\ell - T_{\ell-1}}{\sqrt{T_\ell T_{\ell-1}}} \right)^{-1}, \qquad M_i^\star = \frac{c}{\sqrt{T_i T_{i-1}}} \left(1 + \sum_{\ell=1}^{L} \frac{T_\ell - T_{\ell-1}}{\sqrt{T_\ell T_{\ell-1}}} \right)^{-1}, \quad i = 1, \ldots, L. \tag{34}$$

3.4.2. 方差估计量的分配。 多层蒙特卡洛方差估计量的抽样方差由定理2.9给出。记

$$w_0(x) := \mathrm{Var}\big[V(x; T_0)\big], \qquad w_\ell(x) := \mathrm{Var}\big[\Delta V(x; T_\ell)\big], \ \ell \geq 1, \tag{35}$$

我们可以将目标函数表示为

$$\mathbb{E}[S_V^2(x; \boldsymbol{M}, \boldsymbol{T})] = \frac{w_0(x)}{M_0} + \sum_{\ell=1}^{L} \frac{w_\ell(x)}{M_\ell}, \tag{36}$$

在相同的预算约束下。一般而言，$w_\ell(x)$ 取决于 $\mu_2(x)$ 和 $\mu_4(x)$，因此在实践中，通常要么通过先导运行估计 $\{w_\ell(x)\}$，要么采用诸如 $\mu_4(x) = 3\mu_2^2(x)$ 这样的矩闭合，见公式（28）。相应的最优分配由引理 3.5给出。

引理 3.5 (方差估计量的最优分配)。 固定 x 和一个保真度阶梯 $2 \le T_0 < \cdots < T_L$。令 $a_0 := T_0$，并令 $a_\ell := T_\ell - T_{\ell-1}$，其中 $\ell \ge 1$；设 $c > 0$ 为耦合成本预算 $c = \sum_{\ell=0}^{L} a_\ell M_\ell$。在满足预算约束的连续选择 $M_\ell > 0$ 中，方差估计量的经验抽样方差的期望为，

$$\mathbb{E}\big[S_V^2(x;\boldsymbol{M},\boldsymbol{T})\big] = \frac{w_0(x)}{M_0} + \sum_{\ell=1}^{L} \frac{w_\ell(x)}{M_\ell} \tag{37}$$

由.....最小化

$$M_\ell^\star = c\frac{\sqrt{w_\ell(x)/a_\ell}}{\sum_{k=0}^{L}\sqrt{w_k(x)a_k}}, \qquad \ell = 0,\dots,L. \tag{38}$$

3.6. **选择保真度层级。** 多层构造要求保真度序列严格递增 $\boldsymbol{T} = (T_0,\dots,T_L)$。在我们的设定中 T_ℓ 是单次重复估计量内部使用的随机前向传播次数 $Y(x,T_\ell)$ 和 $V(x,T_\ell)$，因此层级通过引理中的耦合增量方差控制方差–成本权衡 2.6 和定理 2.9。我们将讨论 $\boldsymbol{T}$ 的实用取值，这些取值与第 2.5.1 节和第 2.7.1节中的耦合方式相一致。

一种方便的默认选择是采用几何层级

$$T_\ell := \lceil T_0 r^\ell \rceil, \qquad \ell = 0,\dots,L, \tag{39}$$

其固定比例为 $r > 1$，最小保真度为 T_0。这种选择既能使相邻保真度保持较大间隔，又能限制层级数量。在本文中，我们同时估计预测均值和预测方差，因此要求 $T_0 \ge 2$，以确保无偏样本方差 $V(x,\ T_0)$ 定义良好。此外，方差耦合（18）涉及$V_{T_{\ell-1}+1}$新 块的样本方差$V_{T_{\ell-1}+1:\ T_\ell}(x)$，因此我们还要求

$$T_\ell - T_{\ell-1} \ge 2, \qquad \ell = 1,\dots,L, \tag{40}$$

使得每一步细化至少引入两个新的随机失活抽样。一个简单且稳健的特例是二进制阶梯

$$T_0 = 2, \qquad T_\ell = 2^{\ell+1} \quad (\ell = 0,\dots,L), \tag{41}$$

它恰好满足（40）。

给定 T_0 和 r，最终层级 L 的选取应达到所需的最大内部保真度 $T_{\max}$，例如，采用传统蒙特卡洛 Dropout 基线时所使用的随机失活前向传播次数。具体而言，我们取

$$L := \max\{\ell \in \mathbb{N}_0 : T_\ell \le T_{\max}\}, \tag{42}$$

使得 $T_L \le T_{\max} < T_{L+1}$。在实际应用中，$T_{\max}$ 由每个输入可用的评估预算或内部估计量的目标误差水平 $Y(x,\ T)$和 $V(x,\ T)$决定。

比率 r 的选择主要是一种实用的建模选择。若 r 过于接近 1，则 L会变大，从而增加记账工作量，并且在分配后可能导致某些层级的 M_ℓ 非常小（尤其是若要通过 S_Y^2 和 S_V^2估计抽样方差，通常需要 $M_\ell \ge 2$ 才能计算出非退化样本方差）。若 r 过大，则每次细化都会一次性增加许多新样本，使层级更像是少数几次由粗到细的跃迁。因此，在下文的数值结果中，我们采用二进制阶梯（41）或具有适中比率的几何层级（39）（例如，$r = 2$），它满足（40），并使 L 保持较小。

最后，也可以利用先导样本自适应地选择层级。一种标准的停止准则是不断增大 L，直到最细修正项的经验贡献相对于当前估计量方差可以忽略不计，例如，直到以下两项均满足：

$$\frac{\mathcal{S}^2_{\Delta Y}(x;M_L,T_L)}{M_L} \ll S_Y^2(x;\boldsymbol{M},\boldsymbol{T}), \qquad \frac{\mathcal{S}^2_{\Delta V}(x;M_L,T_L)}{M_L} \ll S_V^2(x;\boldsymbol{M},\boldsymbol{T}), \tag{43}$$

使得在当前预算下进一步细化不会显著降低抽样变异性。该准则与引理 2.6 和定理 2.9中的方差分解相一致，并自然地将 L 的选择与第 3.3节中的分配程序联系起来。

4. Numerical experiments

本节在两个基准物理信息神经网络–Uzawa 问题上评估所提出的多层估计量。目标是：(i) 验证第 2–3 节中的抽样理论；(ii) 量化多层蒙特卡洛相对于单保真度 MC-dropout 的方差–成本权衡；以及 (iii) 在固定成本下，将第 3.3 节中的分配公式与穷举结果进行比较。

4.1. **实验设置与常用定义。** 本节中的所有数值结果均通过对一个固定的、经过训练的代理网络进行重复测试时评估得到，并保持随机失活处于激活状态。一次网络评估 即为一次随机前向传播 $x \mapsto \mathscr{D}(x;\ \theta,$ 其中实现 $\theta \sim \pi$ （第 2.1节）是独立重新采样得到的。所有报告的估计量均在 x上逐点计算，随后对这些逐点量应用范数或回归拟合。除非另有说明，单保真度蒙特卡洛和多层蒙特卡洛均采用相同的实现方式，并按照算法 1，在每个层级增量内实现耦合。

4.1.1. 所报告的估计量与范数。 固定一个评估网格 $\{x_i\}_{i=1}^{N_x} \subset (0,1)$。对于选定的内层保真度$T \geq 2$ 和外层样本数 $M \geq 1$，我们计算单保真度估计量

$$Y^{(m)}(x_i,T) = \frac{1}{T}\sum_{t=1}^{T} \mathscr{D}(x_i;\theta_t^{(m)}), \qquad V^{(m)}(x_i,T) = \frac{1}{T-1}\sum_{t=1}^{T} \big(\mathscr{D}(x_i;\theta_t^{(m)}) - Y^{(m)}(x_i,T)\big)^2,$$

其中 $\theta_t^{(m)}$ 独立同分布$\sim \pi$，然后计算外层平均值

$$\overline{Y}(x_i;M,T) = \frac{1}{M}\sum_{m=1}^{M} Y^{(m)}(x_i,T), \qquad \overline{V}(x_i;M,T) = \frac{1}{M}\sum_{m=1}^{M} V^{(m)}(x_i,T),$$

以及相应的经验外层方差估计量

$$\begin{aligned} \mathcal{S}_Y^2(x_i;M,T) &= \frac{1}{M-1}\sum_{m=1}^{M} \big(Y^{(m)}(x_i,T) - \overline{Y}(x_i;M,T)\big)^2, \\ \mathcal{S}_V^2(x_i;M,T) &= \frac{1}{M-1}\sum_{m=1}^{M} \big(V^{(m)}(x_i,T) - \overline{V}(x_i;M,T)\big)^2. \end{aligned} \tag{44}$$

对于多层实验，我们使用保真度阶梯 $\boldsymbol{T} = (T_0,\dots,T_L)$ 和样本数 $\boldsymbol{M} = (M_0,\dots,M_L)$，并计算耦合多层蒙特卡洛估计量 $\mathcal{Y}(x_i;\ \boldsymbol{M},\boldsymbol{T})$ 和 $\mathcal{V}(x_i;\ \boldsymbol{M},\boldsymbol{T})$，以及相应的经验方差估计量 $S_Y^2(x_i;\ \boldsymbol{M},\boldsymbol{T})$ 和 $S_V^2(x_i;\ \boldsymbol{M},\boldsymbol{T})$，这些量的定义见第 2 节，具体实现见算法 1。

当需要对 x 汇总为一个标量时（例如在方差–保真度图或固定成本曲面中），我们报告离散的 L^1 代理量

$$\|g\|_{L^1} \approx \sum_{i=1}^{N_x} |g(x_i)|\Delta x, \qquad \Delta x := x_{i+1} - x_i, \tag{45}$$

对于均匀网格， Δx 为常数。对于双对数收敛率图，我们通过最小二乘法对 $\log \|g\|_{L^1}$ 关于 $\log\ T$ 拟合线性模型，并报告估计斜率及其置信区间，具体如相应图注所示。

4.1.2. 成本核算。 我们以训练好的网络的随机前向传播总次数来衡量计算成本，也就是说，以调用 $\mathscr{D}(x;\theta)$ 的总次数衡量，其中 $\theta \sim \pi$ （第 3.2 节）独立抽取。对于参数为 $(M,\ T)$ 的单保真度实验，每个输入点的成本为 MT 次评估。

对于多层实验，我们区分非耦合基准成本和有效耦合成本：在非耦合基准成本中，每个增量里的粗略估计量和精细估计量彼此独立生成，记为 c_{unc}；而在有效耦合成本中，前 $T_{\ell-1}$ 个样本会在 T_ℓ样本估计量中重复使用，记为 c_{cpl}。

在固定成本分配研究中，我们枚举满足所选成本约束的整数元组 $\boldsymbol{M}$ （以及任何额外的可行性条件，例如当需要在层级 ℓ 处计算样本方差时的 $M_\ell \geq 2$ ），然后将所得估计量方差曲面的经验最小值与理论分配进行比较

第3.3节。除非另有明确说明，所有“固定成本”多层实验均将 c_{cpl}（公式（30））视为固定值。

4.2. **前向PINNs基准测试：边界层问题**。 我们首先考虑一个具有显式参考解的前向模型问题。这使我们能够：**(i)** 可视化蒙特卡洛随机失活置信带；**(ii)** 验证第 2 节中的单保真度采样率；以及 **(iii)** 测试第3 节中的固定成本分配预测。

4.2.1. 问题定义与参考解。 对于参数 $\epsilon > 0$，我们考虑边界层问题

$$u(x) - \epsilon^2 u''(x) = 1 \quad \text{for } x \in (0,1), \qquad u(0) = u(1) = 0. \tag{46}$$

精确解为

$$u(x) = 1 - \frac{e^{x/\epsilon} + e^{(1-x)/\epsilon}}{1 + e^{1/\epsilon}}. \tag{47}$$

4.2.2.训练与架构。我们使用 $\mathscr{D}$，通过针对边界层问题的 PINNs–Uzawa 设置进行训练（46），并在 L_d 层中启用随机失活。在测试时，训练好的确定性权重保持固定，唯一的随机性来源于 $\theta \sim \pi$，因此每次评估 $\mathscr{D}(x;\ \theta)$ 都对应于第2.1节所述的一次随机前向传播。本实验采用的训练超参数汇总于表1。

N_{in}	N_{out}	N_{width}	L_d	L_h	p_{drop}	N_{SGD}	N_{Uz}	N_{repeats}	γ	ρ	η	ϵ
1	1	64	3	0	0.1	50,000	50	5	100	0.01	0.5	1.0

TABLE 1. 用于训练前向（边界层）替代模型的超参数。一次随机前向传播对应于使用新抽取的随机变量 $\mathscr{D}(x;\theta)$ 进行求值。 $\theta \sim \pi$参数 $N_{\mathrm{in}}, N_{\mathrm{out}}, N_{\mathrm{width}}, L_d, L_h$和 p_{drop} 在第 2.1 节所述的网络架构中定义。参数 $N_{\mathrm{SGD}}, N_{\mathrm{Uz}}$ 和N_{repeats} 分别表示训练轮次总数、每次 Uzawa 更新对应的轮次数，以及每轮计算 PINNs 项所用的重复次数。参数 γ 用于对损失泛函中的边界条件项加权，ρ 是 Uzawa 更新参数，而 η 是优化器学习率。

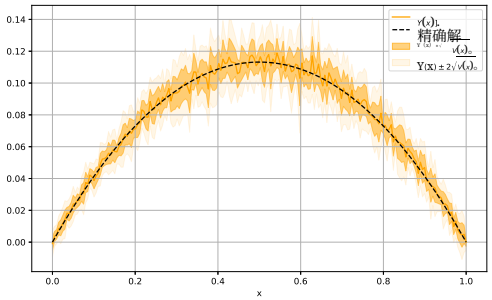
4.2.3. 随着内层保真度增加的单保真度行为。我们首先说明内层保真度 T 如何影响单保真度蒙特卡洛估计量。对于每个 x，我们使用 $M = 1$个外层重复，根据式（6）计算 $Y(x,T)$ 和 $V(x,T)$，并绘制均值估计量及其逐点带 $Y(x,T) \pm \sqrt{V(x,T)}$ 和$Y(x,T) \pm 2\sqrt{V(x,T)}$。图3中的结果表明，增大 T可以减少明显的采样噪声，并使不确定性带更加平滑。

4.2.4. 抽样方差收敛率的经验验证。 接下来，我们验证预测的 T^{-1} 抽样方差缩放规律。对于每个 T，我们生成 $M_0 = 10$ 个单保真度估计量的独立外层重复样本，并根据式（$\mathcal{S}_Y^2(\cdot;\ M_0,\ T)$ 和 $\mathcal{S}_V^2(\cdot;\ M_0,\ T)$ 计算经验估计量方差8。我们按照第 L^1 **4.1** 节所述，报告其在评估网格上的离散 4.1 范数。图 4 显示，双对数斜率近似接近 -1，与引理 2.3一致。

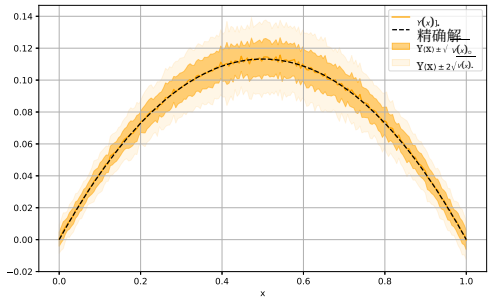
4.2.5. 对丢弃概率的敏感性。 固定 $T_0 = 10^4$ 和 $M_0 = 10$，我们研究经验估计量方差如何随丢弃概率 p_{drop} 变化。图 5 显示，随着 $p_{\mathrm{drop}} \to 0$ （确定性评估），方差趋于零；随着 $p_{\mathrm{drop}} \to 1$，方差则增大。为量化这一趋势，我们拟合广义逻辑模型

$$\log \|S_\cdot^2\|_{L^1} \approx \alpha_1 + \alpha_2 \log p_{\mathrm{drop}}^{\alpha_4} - \alpha_3 \log\left(1 - p_{\mathrm{drop}}^{\alpha_4}\right), \tag{48}$$

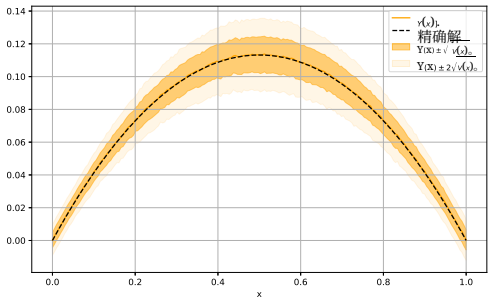
并在表中报告拟合参数和置信区间 2。该实验作为辅助诊断纳入，但不用于多层蒙特卡洛构造，其中保真度由 T 控制。



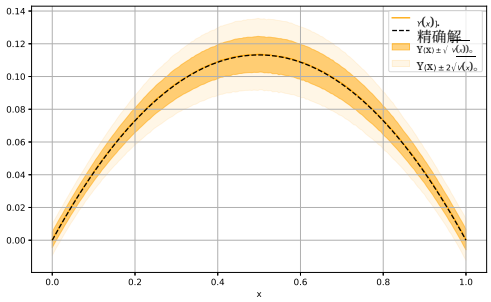
(A) $T = 10$。



(B) $T = 100$。

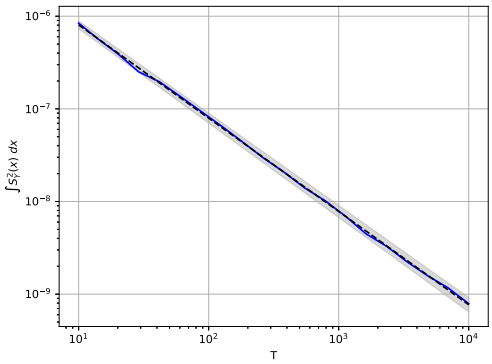


(C) $T = 1000$。

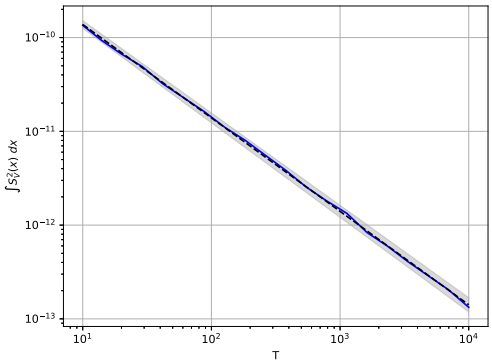


(D) $T = 10000$。

FIGURE 3. 内层保真度逐渐增加时的单保真度单层蒙特卡洛随机失活估计量T。每个子图展示估计均值 $Y(x,\ T)$ 和置信带 $Y(x,\ T) \pm\sqrt{V(x,\ T)}$ 和 $Y(x,\ T) \pm 2\sqrt{V(x,\ T)}$。精确解（47）以虚线曲线表示。



(A) $\|\mathcal{S}_Y^2\|_{L^1}$ 与 T。

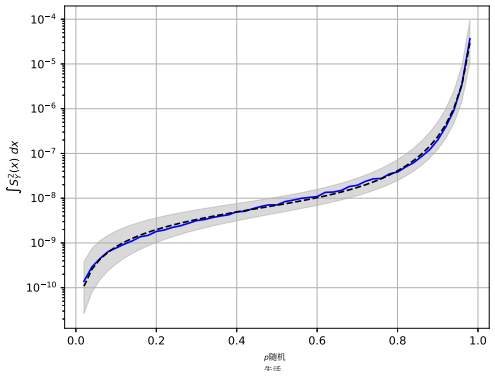


(B) $\|\mathcal{S}_V^2\|_{L^1}$ 与 T。

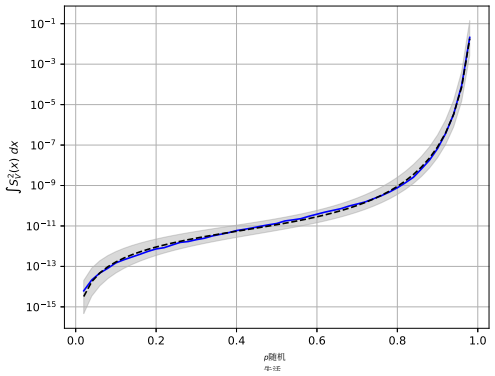
FIGURE 4. 正向问题中经验估计量方差与内层保真度 T 的关系。每个子图均报告对应估计量的经验外样本方差的离散 L^1 范数，并给出双对数线性拟合及拟合斜率的 99% 置信区间。

	α_1	α_2	α_3	α_4
Y	-18.002（-18.143，-17.862）	0.250（0.223，0.277）	3.259（3.037，3.481）	5.059（4.373，5.744）
V	-23.797（-24.064，-23.531）	0.513（0.476，0.550）	8.307（7.917，8.697）	4.766（4.306，5.226）

TABLE 2. 正向问题中模型的拟合参数（48）。括号表示99%置信区间。



(A) $\|\mathcal{S}_Y^2\|_{L^1}$ 与 p_{drop}相比。



(B) $\|\mathcal{S}_V^2\|_{L^1}$ 与 p_{drop}相比。

FIGURE 5. 经验估计量方差与丢弃概率的关系 p_{drop} 对于固定 $T_0 = 10^4$ 的正向问题以及 $M_0 = 10$。虚线曲线表示拟合模型（48）及其99%置信带。

4.2.6. 固定成本多层分配测试。现在，我们在固定耦合成本预算下测试多层分配理论。我们考虑三级层级

$$L = 2, \qquad \boldsymbol{T} = (T_0, T_1, T_2) = (4, 8, 16), \qquad c_{\mathrm{cpl}} = 1000. \tag{49}$$

对于每个满足耦合成本约束（小节 $\boldsymbol{M} = (M_0, M_1, M_2)$ 的整数分配 $M_\ell \geq 2$（第 4.1.2 节）

$$T_0 M_0 + (T_1 - T_0) M_1 + (T_2 - T_1) M_2 = c_{\mathrm{cpl}}, \tag{50}$$

我们使用算法 1，根据式（14）和（21）计算经验多层蒙特卡洛方差估计量 $S_Y^2(\cdot;\, \boldsymbol{M}, \boldsymbol{T})$ 和 $S_V^2(\cdot;\, \boldsymbol{M}, \boldsymbol{T})$。由于（50）留下了两个自由度，我们用 (M_1, M_2) 参数化可行分配，其中M_0 由预算确定，并绘制倒数曲面 $1/\|S_Y^2\|_{L^1}$ 和 $1/\|S_V^2\|_{L^1}$ 以突出显示极小值。图 6 将经验上最优的分配与第 3.3 节推导的连续最优分配进行比较，即用于均值估计量的引理 3.4，以及用于方差估计量的引理 3.5（后者在推导时采用了零超额峰度闭合，即方程（28））。

对于层级 $\boldsymbol{T} = (4、8、16)$ 和成本 $c_{\mathrm{cpl}} = 1000$，理论预测的连续最优分配对于均值估计量而言与 x 无关；而对于方差估计量，在零超额峰度闭合条件下，除一个总体因子外，与底层分布无关。具体而言，均值分配公式给出 $(M_0^\star、 M_1^\star、 M_2^\star) = (1000/12、 1000/24、 1000/48)$，而闭合条件下的方差分配公式给出 $(M_0^\star、 M_1^\star、 M_2^\star) \approx (82。64、44。17、19。76)$。

4.3. 逆问题 PINN 基准测试：随机目标源推断。在本小节中，我们在一个训练目标具有随机性的逆问题 PINN–Uzawa 问题上测试多层估计量。该设置为目标矩提供了闭式参考值，并使我们能够将以下两类误差区分开来：(i) 估计丢弃诱导的矩时产生的蒙特卡洛采样误差；(ii) 由训练不完善导致的代理模型失配。

4.3.1. 问题定义与真实矩。 我们考虑以下约束优化问题：给定 $\alpha > 0$ 和一个（可能是随机的）目标 $u_{\mathrm{target}} \in L^2(0,1)$，最小化

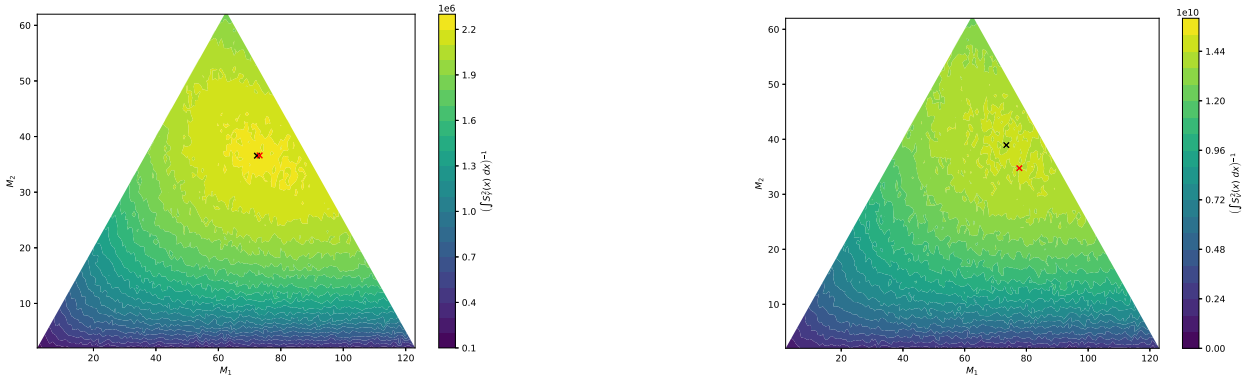
$$\frac{1}{2}\|u - u_{\mathrm{target}}\|_{L^2(0,1)}^2 + \frac{\alpha}{2}\|f\|_{L^2(0,1)}^2 \tag{51}$$

满足偏微分方程约束

$$-u''(x) = f(x) \ \text{ on } (0,1), \qquad u(0) = u(1) = f(0) = f(1) = 0. \tag{52}$$

对于一个合成基准测试 对于具有已知矩的标记，我们通过一个标量噪声变量生成随机性

$$u_{\mathrm{target}}(x) = (1 + \omega)(1 + \alpha\pi^4)\sin(\pi x), \quad \omega \sim \mathrm{Unif}(-\delta/2, \delta/2), \tag{53}$$



(A) $1/\|S_Y^2\|_{L^1}$ 作为 $(M_1,\ M_2)$ 的函数。

(B) $1/\|S_V^2\|_{L^1}$ 作为 $(M_1,\ M_2)$ 的函数。

FIGURE 6. 正向问题的固定成本分配测试，其中 $\boldsymbol{T} = (4、8、16)$ 和 $c_{\mathrm{cpl}} = 1000$。黑色标记表示可行整数分配中的经验最小值点，红色标记表示引理 3.4 （左图）或引理 3.5（右图）给出的连续最优解。请注意，方差分配公式是在零超额峰度闭合假设下推导的。

并且我们使用闭式族

$$u(x;\omega) = (1+\omega)\sin(\pi x), \qquad f(x;\omega) = (1+\omega)\pi^2 \sin(\pi x). \tag{54}$$

相应的真实矩为

$$\mathbb{E}[u(x)] = \sin(\pi x), \qquad \mathbb{E}[f(x)] = \pi^2 \sin(\pi x), \tag{55}$$

并且，利用 $\mathrm{Var}(\omega) = \delta^2/12$,

$$\mathrm{Var}[u(x)] = \frac{\delta^2}{12}\sin^2(\pi x), \qquad \mathrm{Var}[f(x)] = \frac{\delta^2}{12}\pi^4 \sin^2(\pi x). \tag{56}$$

这些表达式为评估经验估计量提供了参考基准。经验估计量是按照第 4.1 节所述方法计算的丢弃诱导的矩。

4.3.2. 代理模型、架构与超参数。 我们训练一个双输出丢弃网络 $\mathscr{D} = (\mathscr{U},\ \mathscr{F})$，以近似与随机目标（53）相关联的解对（$u,\ f$），该解对对应于（51）–（52）。该架构采用共享主干，并在其后为 $\mathscr{U}$ 和 $\mathscr{F}$ 设置分支输出头；我们通过输出因子 $x(1-x)$ 强制施加齐次狄利克雷边界条件。示意图见图 7，训练超参数列于表 3。

在评估时，我们固定训练后的权重并保持随机失活处于激活状态，因此每次前向传播都会抽取一个新的实现 $\theta \sim \pi$ 并生成一个随机预测 $\mathscr{D}(x;\ \theta)$。随后，按照第 $\mathcal{Y}$、 $\mathcal{V}$、 S_Y^2 和 S_V^2 节中的说明，并依照算法2中的实现，通过重复进行随机前向传播来计算估计量1。

N_{in}	N_{out}	N_{width}	L_d	$L_{q,u}$	$L_{q,f}$	p_{drop}	N_{SGD}
1	2	128	4	0	1	0.2	200000
N_{Uz}	N_{repeats}	N_{lag}	α	β	η	ρ	δ
50	5	20	10^{-4}	10^{-4}	2.5×10^{-5}	10^{-3}	2.5×10^{-2}

TABLE 3. 用于训练逆向基准替代模型的超参数 $\mathscr{D} = (\mathscr{U},\ \mathscr{F})$。该架构采用共享主干/分支头丢弃设计，见图 7。许多超参数的定义与表 1中相同。超参数 β对应损失泛函中物理信息神经网络的权重， N_{lag} 则表示每次 Uzawa 更新的评估次数。

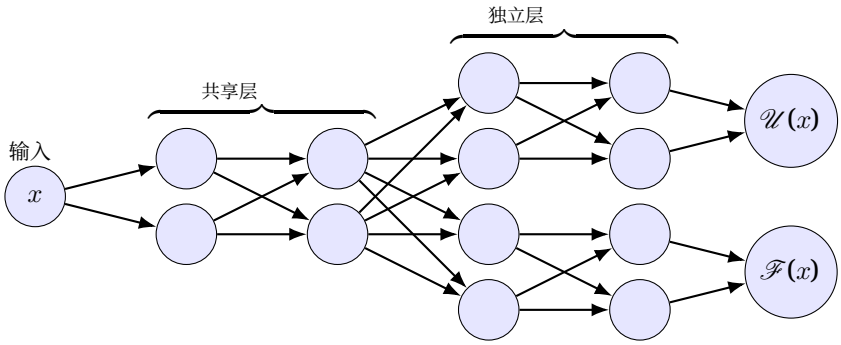


FIGURE 7. 逆向基准的代理网络架构示意图。该网络包含一个由随机失活层组成的共享主干，后接分别用于 $\mathscr{U}$ 和 $\mathscr{F}$的分支输出头。每个输出头均应用边界约束输出因子 $x(1-x)$。

4.3.3. 单保真度行为：预测均值。 我们首先评估对丢弃引起的预测均值进行单保真度蒙特卡洛估计的效果。对于每个输出，我们在较低的内层保真度 $Y(x;\ M,\ T)$和较高的内层保真度 $T=10$下计算 $T=10^4$ ，并与真实期望（55）进行比较。图8 展示了估计期望及相应的逐点误差。

4.3.4. 单保真度行为：预测方差。 接下来，我们通过 $V(x)$ 评估对随机失活所致预测方差进行单保真度估计的效果。我们报告标准差 $\sqrt{V(x)}$，并与由 (56) 得到的解析标准差进行比较。图 9展示了 $T=10$ 和 $T=10^4$ 的结果。

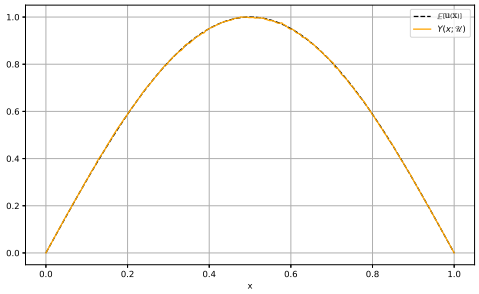
4.3.5. 抽样方差收敛率的经验验证。 现在，我们验证单保真度估计量的抽样方差是否按预测的 $\mathcal{O}(T^{-1})$ 衰减（引理 2.3）。对于每个保真度 T，我们计算 M_0 个独立的外层重复样本，并分别针对两个输出构造经验外层方差估计量 $\mathcal{S}_Y^2(x;\ M_0,\ T)$ 和 $\mathcal{S}_V^2(x;\ M_0,\ T)$，然后按照第 4.1.1 小节所述，使用评估网格上的离散 L^1 范数汇总其量级。图 10 给出了相应的双对数拟合结果，表 4 列出了估计斜率及其 99%置信区间。

	99%置信区间下限	梯度估计值	99%置信区间上限
$\mathcal{S}_Y^2(\mathscr{U})$	-1.0159	-1.0033	-0.9907
$\mathcal{S}_Y^2(\mathscr{F})$	-1.0094	-0.9976	-0.9857
$\mathcal{S}_V^2(\mathscr{U})$	-1.0000	-0.9852	-0.9704
$\mathcal{S}_V^2(\mathscr{F})$	-0.9963	-0.9842	-0.9722

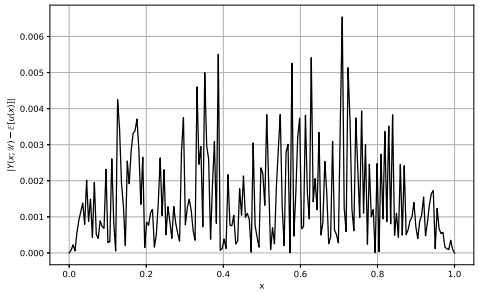
TABLE 4. 图 10 中单保真度采样方差的对数-对数梯度估计值相对于 T的变化，并附有 99%置信区间。我们观察到，估计量方差的 L^1-范数呈现近似 $\mathcal{O}(T^{-1}$ 衰减。

4.3.6. 固定成本多层分配测试。最后，我们针对多层估计量进行固定成本分配实验。我们采用三级层级 $\boldsymbol{T}=(T_0,T_1,T_2)=(4,8,16)$（因此 $L=2$），并设定固定的耦合评估预算 $c_{\mathrm{cpl}}=1000$（第 4.1.2 小节）。我们枚举满足 $T_0M_0+(T_1-T_0)M_1+(T_2-T_1)M_2=c_{\mathrm{cpl}}$ 及约束 $M_\ell\geq 2$的可行整数分配 $\boldsymbol{M}=(M_0,M_1,M_2)$。我们枚举满足预算和约束 $M_\ell\geq 2$ （从而使样本方差有定义）的可行整数分配 $\boldsymbol{M}=(M_0,M_1,M_2)$，并针对两种输出估计经验多层蒙特卡洛方差估计量 S_Y^2 和 S_V^2。为便于可视化，我们绘制倒数曲面 $1/\|S_\cdot^2\|_{L^1}$ 关于 (M_1,M_2) 的函数图，并注意到 M_0 随后由成本约束确定。

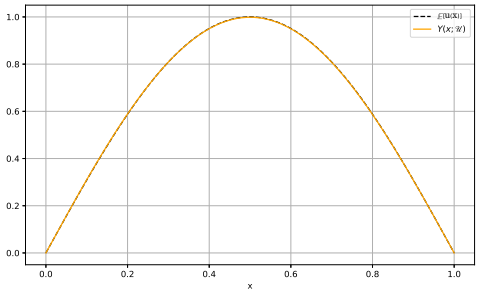
黑色标记表示可行整数分配中的经验最小值点。红色标记表示根据§3.3.1 中均值估计量的分配公式得到的连续最优解



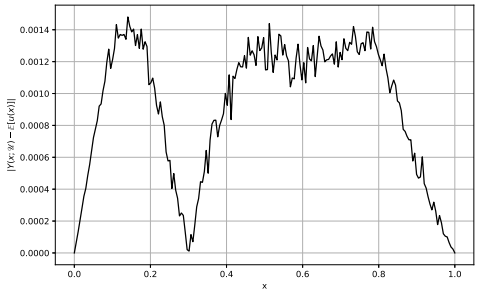
(A) $Y(\mathscr{U})$ 与 $\mathbf{E}[u]$ 的比较， $T = 10$。



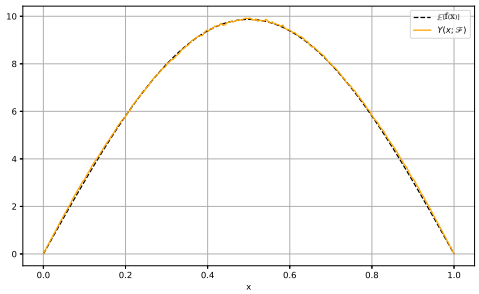
(B) $Y(\mathscr{U})$ 的逐点误差， $T = 10$。



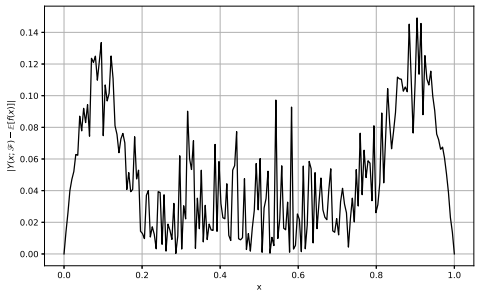
(C) $Y(\mathscr{U})$ 与 $\mathbf{E}[u]$, $T = 10^4$ 的比较。



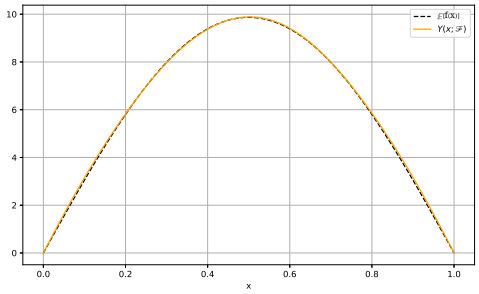
(D) $Y(\mathscr{U})$ 的逐点误差， $T = 10^4$。



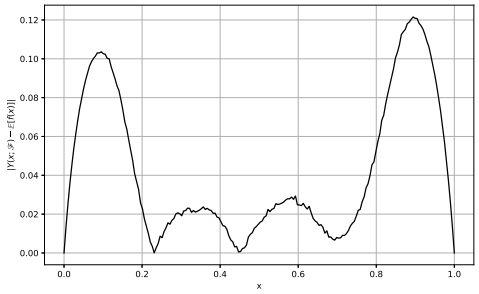
(E) $Y(\mathscr{F})$ 与 $\mathbf{E}[f]$, $T = 10$ 的比较。



(F) $Y(\mathscr{F})$ 的逐点误差， $T = 10$。

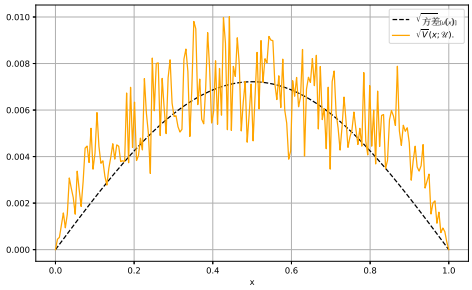


(G) $Y(\mathscr{F})$ 与 $\mathbf{E}[f]$, $T = 10^4$ 的比较。

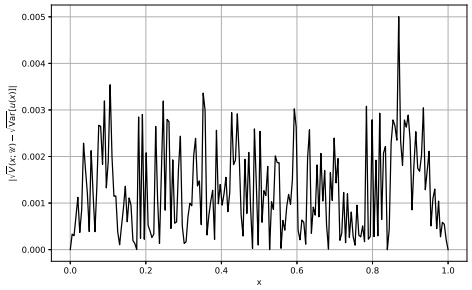


(H) $Y(\mathscr{F})$ 的逐点误差， $T = 10^4$。

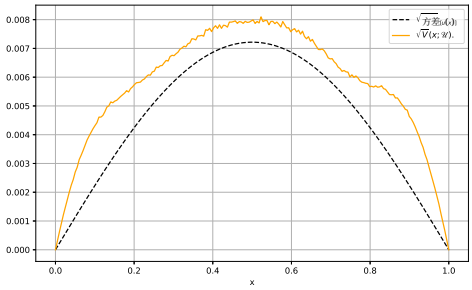
FIGURE 8. 逆向基准预测均值的单保真度单层蒙特卡罗估计。前两行报告 $Y(x)$，对应于 $\mathscr{U}$，并将其与 $\mathbb{E}[u(x)]$ （见（55））进行比较；后两行报告 $Y(x)$，对应于 $\mathscr{F}$，并将其与 $\mathbb{E}[f(x)]$ 进行比较。对于每个输出，我们展示低保真度估计（$T = 10$）和高保真度估计（$T = 10^4$），以及相应的逐点误差。与图9中的方差估计不同，预测均值随着 T 的增大而收敛至解析期望，这表明与蒙特卡洛噪声相比，认知误差可以忽略不计。



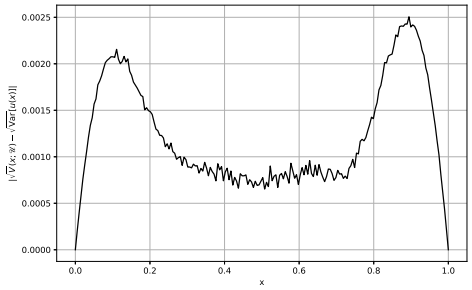
(A) $\sqrt{V(\mathscr{U})}$ 与 $\sqrt{\text{方差}[u]}$, $T = 10$。



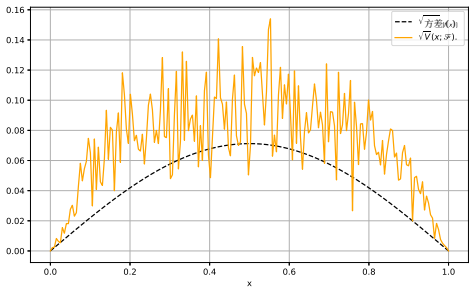
(B) $\sqrt{V(\mathscr{U})}$ 的逐点误差, $T = 10$。



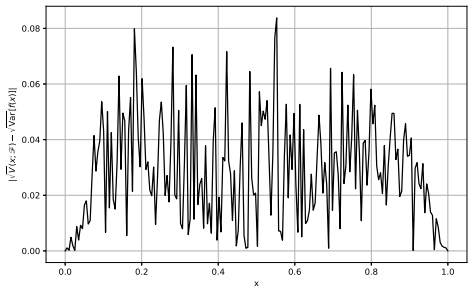
(C) $\sqrt{V(\mathscr{U})}$与 $\sqrt{\text{方差}[u]}$, $T = 10^4$。



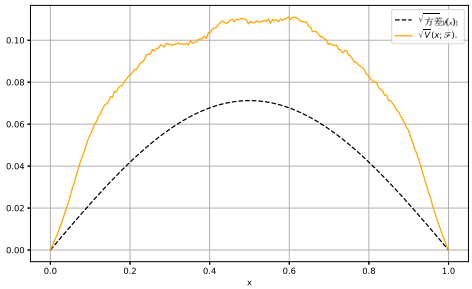
(D) $\sqrt{V(\mathscr{U})}$的逐点误差, $T = 10^4$。



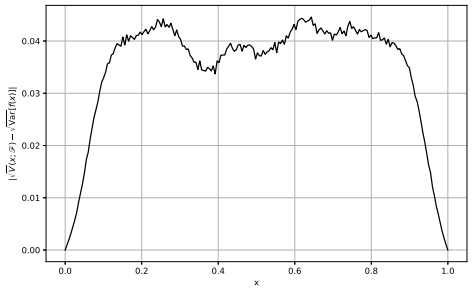
(E) $\sqrt{V(\mathscr{F})}$与 $\sqrt{\text{方差}[f]}$, $T = 10$。



(F) $\sqrt{V(\mathscr{F})}$的逐点误差, $T = 10$。

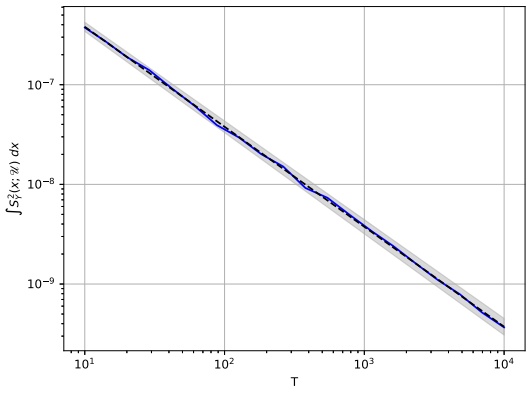


(G) $\sqrt{V(\mathscr{F})}$与 $\sqrt{\text{方差}[f]}$, $T = 10^4$。

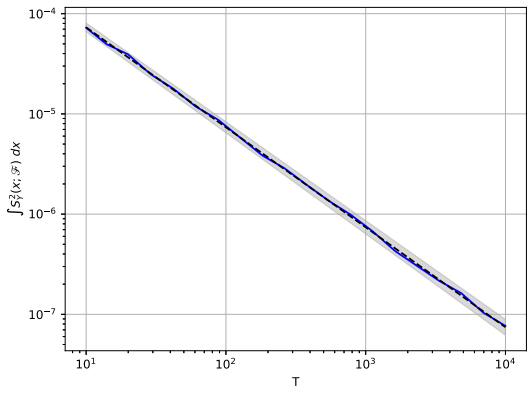


(H) $\sqrt{V(\mathscr{F})}$的逐点误差, $T = 10^4$。

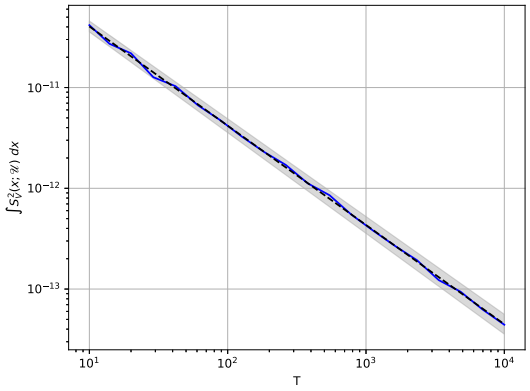
FIGURE 9. 逆向基准中预测标准差的单保真度单层蒙特卡洛估计。顶行报告 $\sqrt{V(x)}$ 对于 $\mathscr{U}$，并将其与公式 (56) 中的 $\sqrt{\text{方差}[u(x)]}$ 进行比较；底行报告 $\sqrt{V(x)}$ 对于 $\mathscr{F}$，并将其与 $\sqrt{\text{方差}[f(x)]}$进行比较。低保真度和高保真度结果（$T = 10$ 和 $T = 10^4$）及其逐点误差一并展示。我们观察到，增大 T 会降低 V估计量的采样噪声；在较高 T 下仍然存在的偏差反映的是代理模型引起的认知不确定性失配，而非蒙特卡罗误差，这表明丢弃法后验未能充分捕捉目标映射的偶然变异。



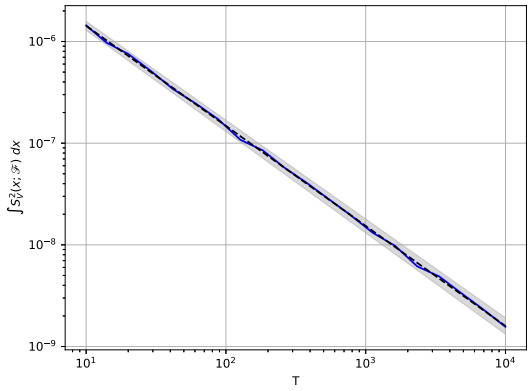
(A) $\|\mathcal{S}_Y^2(\mathscr{U})\|_{L^1}$ 与 T相比。



(B) $\|\mathcal{S}_Y^2(\mathscr{F})\|_{L^1}$ 与 T相比。



(C) $\|\mathcal{S}_V^2(\mathscr{U})\|_{L^1}$ 与 T相比。



(D) $\|\mathcal{S}_V^2(\mathscr{F})\|_{L^1}$ 与 T相比。

FIGURE 10. 逆向基准的经验采样方差衰减。每个面板均报告了单保真度估计量的经验抽样方差在评估网格上的离散 L^1 范数，该范数是内层保真度的函数 T。虚线表示在双对数坐标轴上的线性拟合，阴影区域表示拟合斜率的99%置信区间。对应的梯度列于表4。

以及§ 3.4.2 中针对方差估计量的内容（后一种情况是在简化分配所用的零超额峰度闭合条件下）。

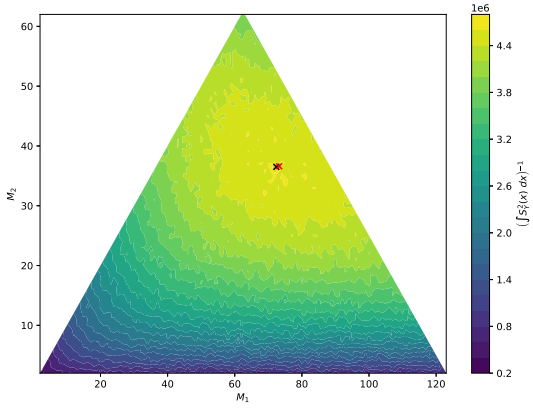
$\boldsymbol{T} = (4, 8, 16)$和 $c_{\mathrm{cpl}} = 1000$ 的理论连续最优解与前一个问题相同，因为最优的 $\boldsymbol{M}$ 仅取决于保真度阶梯 $\boldsymbol{T}$,

$$\boldsymbol{M}_Y^\star = \left(\frac{1000}{12}, \frac{1000}{24}, \frac{1000}{48}\right), \qquad \boldsymbol{M}_V^\star \approx (82.64, 44.17, 19.76), \tag{57}$$

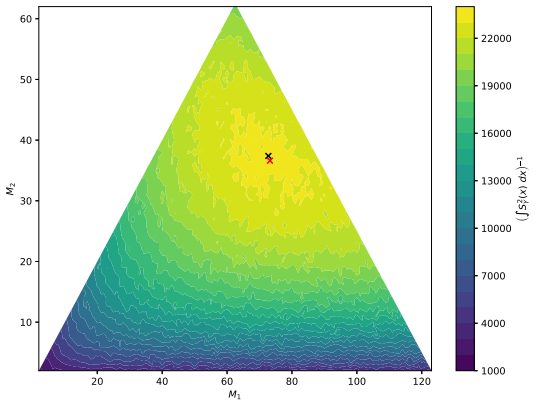
其中， $\boldsymbol{M}_Y^\star$ 是均值估计量分配， $\boldsymbol{M}_V^\star$ 是零超额峰度闭合条件下的方差估计量分配。

5. CODE DISCUSSION

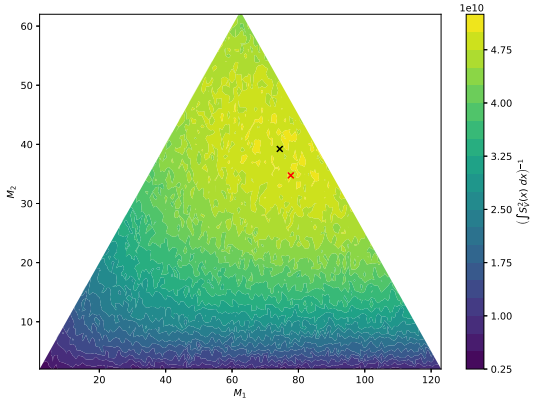
5.1. 可复现性。 所有实验均可复现，但以下情况除外：(i) 基于随机失活的推断所固有的随机性；(ii) GPU 算术运算和库实现所产生的不确定性。在我们的代码中，伪随机数生成器会在训练开始时使用 CONFIG.py 中指定的值进行设定，并且评估时也可使用相同的设定机制，以便在需要时使图表具有可重复性。每次训练运行都会创建一个带时间戳的输出目录，其中会复制完整的配置文件并保存训练后的模型权重，从而可以无歧义地重新加载运行结果。尽管



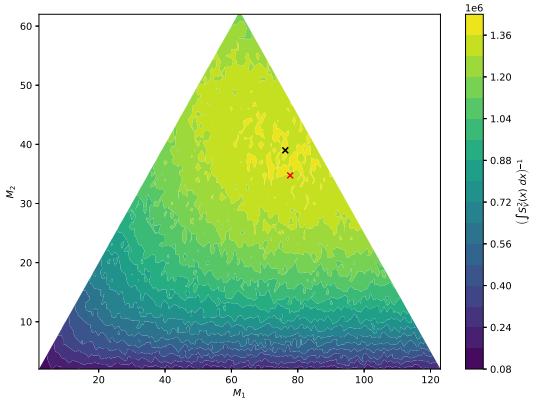
(A) $1/\|S_Y^2(\mathscr{U})\|_{L^1}$ 作为（M_1， M_2）的函数。



(B) $1/\|S_Y^2(\mathscr{F})\|_{L^1}$ 作为（M_1， M_2）的函数。



(C) $1/\|S_V^2(\mathscr{U})\|_{L^1}$ 作为（M_1， M_2）的函数。



(D) $1/\|S_V^2(\mathscr{F})\|_{L^1}$ 作为（M_1， M_2）的函数。

FIGURE 11. 逆向基准的固定成本分配测试，其中 $\boldsymbol{T} = (4,\ 8,\ 16)$ 和$c_{\mathrm{cpl}} = 1000$。每个面板展示了 MLMC估计量的倒方差曲面，该曲面以整数分配（M_1， M_2）为自变量，其中 M_0 由耦合成本约束确定。黑色标记表示在可行整数分配中找到的经验最小值点，红色标记表示根据第§ 3.3.1 节（上排）和第§ 3.4.2节（下排，简化分配采用零超额峰度闭合）中的分配公式得到的连续最优解。

这些步骤能够确保在相同的软件栈和硬件上重复获得结果，但无法保证在不同的 GPU架构或不同版本的 PyTorch/CUDA 之间实现逐位确定性。

本文的代码仓库_可在此处获取：蒙特_卡洛_随机失活_物理信息神经网络.git。

5.2. **计算成本与实现说明**。 在全文中，计算成本的实际度量标准是训练好的网络的随机前向传播次数，这与小节 3.2中的成本模型一致。具体而言，耦合采样策略会在不同保真度间重复使用丢弃掩码，因此与非耦合基线相比，形成多层增量所需的额外网络评估次数更少。

在实践中，将样本聚合为 $Y(x,\ T)$、 $V(x,\ T)$ 以及多层估计量所需的额外算术运算，与本文所考察架构的网络评估成本相比可以忽略不计。因此，我们不尝试进行针对特定硬件的浮点运算计数，而是以网络前向传播次数报告预算和分配。

6. Conclusions

我们提出了一种用于蒙特卡洛随机失活的多层蒙特卡洛框架：通过抽样丢弃掩码引入认知不确定性，并通过随机前向传播次数控制保真度。通过复用掩码耦合粗、细内部估计量，所得的预测均值和方差多层估计量对于相应的随机失活诱导量仍保持无偏，同时在给定评估预算下大幅降低抽样方差。我们推导了明确的方差-成本关系、有效耦合成本模型和有理论依据的样本分配规则，并在正向与逆向 PINNs–Uzawa 基准测试中验证了这些预测；实测方差收敛率和固定成本分配实验与理论相符，差异仅来自整数可行性和矩闭合假设的预期影响。

Appendix A. Supplementary derivations

A.1. 期望与方差恒等式的证明。

引理的证明 *2.3*。 关于 $\mathbb{E}[Y]$ 和方差 $\mathrm{Var}[Y]$ 的恒等式是标准结果。样本方差的无偏性意味着 $\mathbb{E}[V] = \mu_2$，而 $\mathrm{Var}[V]$ 的表达式是无偏方差估计量抽样方差的经典公式，例如见 [Ben18，公式 10]。 □

引理的证明 *2.6*。 根据期望的线性性和（12），

$$\mathbb{E}\big[\mathcal{Y}(x;\boldsymbol{M},\boldsymbol{T})\big] = \mathbb{E}\big[Y(x;T_0)\big] + \sum_{\ell=1}^{L} \mathbb{E}\big[\Delta Y(x;T_\ell)\big] = \mathbb{E}\big[Y(x;T_L)\big] = \mu(x),$$

这里用到了引理 2.3。考虑期望估计量的经验抽样方差的期望，由期望的线性性可得

$$\mathbb{E}[S_Y^2(x;\boldsymbol{M},\boldsymbol{T})] = \frac{1}{M_0}\mathbb{E}[\mathcal{S}_Y^2(x;M_0,T_0)] + \sum_{\ell=1}^{L} \frac{1}{M_\ell}\mathbb{E}[\mathcal{S}_{\Delta Y}^2(x;M_\ell,T_\ell)], \tag{58}$$

$\mathcal{S}_Y^2$ 和 $\mathcal{S}_{\Delta Y}^2$ 的定义分别是：它们是 Y 和 ΔY 的无偏方差估计量。在外层样本独立同分布的假设下，我们有无偏方差估计量的期望等于方差，即 $\mathrm{E}[\mathcal{S}_Z^2] = \mathrm{Var}[Z]$，因此

$$\mathbb{E}[S_Y^2(x;\boldsymbol{M},\boldsymbol{T})] = \frac{1}{M_0}\mathrm{Var}\big[Y(x;T_0)\big] + \sum_{\ell=1}^{L} \frac{1}{M_\ell}\mathrm{Var}\big[\Delta Y(x;T_\ell)\big]. \tag{59}$$

在耦合 (10) 下， $\Delta Y(x;\, T_\ell) = \frac{1}{T_\ell}\sum_{t=T_{\ell-1}+1}^{T_\ell} \mathscr{Q}(x;\,\theta_t) - \frac{T_\ell - T_{\ell-1}}{T_\ell} Y(x;\, T_{\ell-1})$，利用独立性直接计算可得 $\mathrm{Var}[\Delta Y(x;\, T_\ell)] = \mu_2(x)\left(\frac{1}{T_{\ell-1}} - \frac{1}{T_\ell}\right)$，而引理 2.3 给出 $\mathrm{Var}[Y(x;\, T_0)] = \mu_2(x)/T_0$。 □

引理 *2.8*的证明。 设 D_1、……、D_{T_ℓ} 独立同分布，并令 $\mu_k := \mathbb{E}\big[(X-\mu)^k\big]$ 表示 k 阶中心矩。引入大小为 新 的区块的均值 $T_\ell - T_{\ell-1}$，

$$\overline{D}_{T_{\ell-1}:T_\ell} := \frac{1}{T_\ell - T_{\ell-1}} \sum_{j=T_{\ell-1}+1}^{T_\ell} D_j. \tag{60}$$

由以下内容给出的标准更新恒等式 [CGL79] 指出

$$\begin{aligned}(T_\ell - 1)V(T_\ell) =& (T_{\ell-1} - 1)V(T_{\ell-1}) + \sum_{i=T_{\ell-1}+1}^{T_\ell} \big(D_i - \overline{D}_{T_{\ell-1}:T_\ell}\big)^2 \\ &+ \frac{(T_\ell - T_{\ell-1})T_{\ell-1}}{T_\ell}\big(Y(T_{\ell-1}) - \overline{D}_{T_{\ell-1}:T_\ell}\big)^2.\end{aligned} \tag{61}$$

右侧的中间项仅取决于新的样本 $D_{T_{\ell-1}+1},\dots,D_{T_\ell}$，且独立于 $V(T_{\ell-1})$，因此它与 $V(T_{\ell-1})$ 的协方差为零。对式（61）中的 $(T_{\ell-1})$ 取协方差，并利用双线性性，可得

$$
\begin{aligned}
(T_\ell-1)\mathrm{Cov}\big[V(T_\ell),V(T_{\ell-1})\big] =&(T_{\ell-1}-1)\mathrm{Var}\big[V(T_{\ell-1})\big]\\
&+\frac{(T_\ell-T_{\ell-1})T_{\ell-1}}{T_\ell}\mathrm{Cov}\Big(\big(Y(T_{\ell-1})-\overline{D}_{T_{\ell-1}:T_\ell}\big)^2,\ V(T_{\ell-1})\Big).
\end{aligned}
\tag{62}
$$

由于 $D_{\overline{T_{\ell-1}:\ T_\ell}}$ 与 $\{D_i\}_{i=1}^{T_{\ell-1}}$ 相互独立，且 $\mathrm{E}[D_{T_{\ell-1}:\ \overline{T_\ell}}]=\mu$，经过简单展开并利用独立性，再将协方差表示为它们偏差的乘积，可得

$$
\mathrm{Cov}\Big(\big(Y(T_{\ell-1})-\overline{D}_{T_{\ell-1}:T_\ell}\big)^2,\ V(T_{\ell-1})\Big) =\mathbb{E}\Big[\big(Y(T_{\ell-1})-\mu\big)^2\big(V(T_{\ell-1})-\mu_2\big)\Big].
\tag{63}
$$

还需计算右侧的期望。令 $\xi_i:=D_i-\mu$，则 $Y(T)-\mu=\frac{1}{T}\sum_{i=1}^{T}\xi_i$且 –

$$
(T_{\ell-1}-1)V(T_{\ell-1})=\sum_{i=1}^{T_{\ell-1}}\xi_i^2-T_{\ell-1}\big(Y(T_{\ell-1})-\mu\big)^2.
\tag{64}
$$

利用（64），

$$
\mathbb{E}\big[(Y(T_{\ell-1})-\mu)^2V(T_{\ell-1})\big]=\frac{1}{T_{\ell-1}-1}\mathbb{E}\Big[(Y(T_{\ell-1})-\mu)^2\sum_{i=1}^{T_{\ell-1}}\xi_i^2\Big]-\frac{T_{\ell-1}}{T_{\ell-1}-1}\mathbb{E}\big[(Y(T_{\ell-1})-\mu)^4\big].
\tag{65}
$$

直接进行指标计数论证可得

$$
\mathbb{E}\Big[(Y(T_{\ell-1})-\mu)^2\sum_{i=1}^{T_{\ell-1}}\xi_i^2\Big]=\frac{1}{T_{\ell-1}^2}\mathbb{E}\Big[\Big(\sum_{i=1}^{T_{\ell-1}}\xi_i\Big)^2\Big(\sum_{k=1}^{T_{\ell-1}}\xi_k^2\Big)\Big]=\frac{\mu_4}{T_{\ell-1}}+\frac{T_{\ell-1}-1}{T_{\ell-1}}\mu_2^2,
\tag{66}
$$

类似地

$$
\mathbb{E}\big[(Y(T_{\ell-1})-\mu)^4\big]=\frac{\mu_4}{T_{\ell-1}^3}+\frac{3(T_{\ell-1}-1)}{T_{\ell-1}^3}\mu_2^2.
\tag{67}
$$

将（66）–（67）代入（65）可得

$$
\mathbb{E}\big[(Y(T_{\ell-1})-\mu)^2V(T_{\ell-1})\big]=\frac{\mu_4}{T_{\ell-1}^2}+\frac{T_{\ell-1}-3}{T_{\ell-1}^2}\mu_2^2.
\tag{68}
$$

由于 $\mathrm{E}\big[(Y(T_{\ell-1})-\mu)^2\big]=\mu_2/T_{\ell-1}$，我们得到

$$
\mathbb{E}\Big[\big(Y(T_{\ell-1})-\mu\big)^2\big(V(T_{\ell-1})-\mu_2\big)\Big]=\Big(\frac{\mu_4}{T_{\ell-1}^2}+\frac{T_{\ell-1}-3}{T_{\ell-1}^2}\mu_2^2\Big)-\frac{\mu_2}{T_{\ell-1}}\mu_2=\frac{1}{T_{\ell-1}^2}\big(\mu_4-3\mu_2^2\big).
\tag{69}
$$

结合（62）、（63）和（69）可得

$$
(T_\ell-1)\mathrm{Cov}\big[V(T_\ell),V(T_{\ell-1})\big]=(T_{\ell-1}-1)\mathrm{Var}\big[V(T_{\ell-1})\big]+\frac{T_\ell-T_{\ell-1}}{T_{\ell-1}T_\ell}\big(\mu_4-3\mu_2^2\big).
$$

两边除以 $T_\ell-1$ 可得（23）的第一行。方差$\big[V(T_{\ell-1})\big]$ 项可直接利用方差$[V(T_{\ell-1})]$,的闭式 $\frac{1}{T_{\ell-1}}\big(\mu_4-\frac{T_{\ell-1}-3}{T_{\ell-1}-1}\mu_2^2\big)$进行计算。□

A.2. **最优分配结果的证明**。

引理的证明*3.4*。 固定 x和 $\boldsymbol{T}=(T_0,\dots,T_L)$，其中 $1\le T_0<\cdots<T_L$。回顾加性耦合成本表示

$$
c=T_0M_0+\sum_{\ell=1}^{L}(T_\ell-T_{\ell-1})M_\ell=\sum_{\ell=0}^{L}a_\ell M_\ell,\qquad a_0:=T_0,\quad a_\ell:=T_\ell-T_{\ell-1}\ (\ell\ge1),
\tag{70}
$$

以及方差分解

$$\mathbb{E}[S_Y^2(x;\boldsymbol{M},\boldsymbol{T})] = \frac{v_0(x)}{M_0} + \sum_{\ell=1}^{L} \frac{v_\ell(x)}{M_\ell}, \tag{71}$$

其中

$$v_0(x) = \frac{\mu_2(x)}{T_0}, \qquad v_\ell(x) = \mu_2(x)\left(\frac{1}{T_{\ell-1}} - \frac{1}{T_\ell}\right) = \mu_2(x)\frac{T_\ell - T_{\ell-1}}{T_\ell T_{\ell-1}}, \quad \ell \geq 1. \tag{72}$$

在开集 $\{M_\ell > 0\}_{\ell=0}^{L}$ 上，目标函数 $\sum_{\ell=0}^{L} v_\ell(x)/M_\ell$ 严格凸，约束 $\sum_{\ell=0}^{L} a_\ell M_\ell = c$ 是仿射的，因此任何驻点都是唯一的全局极小值点。

引入拉格朗日函数

$$\mathcal{L}(\boldsymbol{M},\lambda) := \sum_{\ell=0}^{L} \frac{v_\ell(x)}{M_\ell} + \lambda\left(\sum_{\ell=0}^{L} a_\ell M_\ell - c\right). \tag{73}$$

对每个 ℓ，一阶最优性条件为：

$$\frac{\partial \mathcal{L}}{\partial M_\ell} = -\frac{v_\ell(x)}{M_\ell^2} + \lambda a_\ell = 0, \qquad \Longrightarrow \qquad M_\ell = \sqrt{\frac{v_\ell(x)}{\lambda a_\ell}}. \tag{74}$$

利用（72），可得

$$\frac{v_0(x)}{a_0} = \frac{\mu_2(x)}{T_0^2}, \qquad \frac{v_\ell(x)}{a_\ell} = \frac{\mu_2(x)}{T_\ell T_{\ell-1}} \quad (\ell \geq 1), \tag{75}$$

因此（74）意味着

$$M_0 = \frac{\sqrt{\mu_2(x)}}{\sqrt{\lambda}}\frac{1}{T_0}, \qquad M_\ell = \frac{\sqrt{\mu_2(x)}}{\sqrt{\lambda}}\frac{1}{\sqrt{T_\ell T_{\ell-1}}}, \quad \ell = 1,\dots,L. \tag{76}$$

施加预算约束（70）可得

$$\begin{aligned} c &= T_0 M_0 + \sum_{\ell=1}^{L} (T_\ell - T_{\ell-1}) M_\ell \\ &= \frac{\sqrt{\mu_2(x)}}{\sqrt{\lambda}}\left(1 + \sum_{\ell=1}^{L} \frac{T_\ell - T_{\ell-1}}{\sqrt{T_\ell T_{\ell-1}}}\right), \end{aligned} \tag{77}$$

so

$$\frac{\sqrt{\mu_2(x)}}{\sqrt{\lambda}} = c\left(1 + \sum_{\ell=1}^{L} \frac{T_\ell - T_{\ell-1}}{\sqrt{T_\ell T_{\ell-1}}}\right)^{-1}. \tag{78}$$

将（78）代入（76）可得（33）。因子 $\mu_2(x)$ 相消，因此最优分配仅取决于$\boldsymbol{T}$ 和 c。 □

引理 *3.5* 的证明。我们希望最小化方差估计量的经验抽样方差的期望

$$\mathbb{E}[S_V^2(x;\boldsymbol{M},\boldsymbol{T})] = \frac{w_0(x)}{M_0} + \sum_{\ell=1}^{L} \frac{w_\ell(x)}{M_\ell} \tag{79}$$

满足相同的耦合成本约束

$$c = \sum_{\ell=0}^{L} a_\ell M_\ell, \qquad a_0 := T_0, \quad a_\ell := T_\ell - T_{\ell-1} \ (\ell \geq 1). \tag{80}$$

与引理 3.4 的证明类似， $\sum w_\ell/M_\ell$ 在 $M_\ell > 0$ 上的严格凸性意味着 KKT 驻点是唯一的全局极小值点。引入拉格朗日函数

$$\mathcal{L}(\boldsymbol{M},\lambda) := \sum_{\ell=0}^{L} \frac{w_\ell(x)}{M_\ell} + \lambda\left(\sum_{\ell=0}^{L} a_\ell M_\ell - c\right),$$

一阶条件给出

$$-\frac{w_\ell(x)}{(M_\ell^\star)^2} + \lambda a_\ell = 0, \qquad \Longrightarrow \qquad M_\ell^\star = \sqrt{\frac{w_\ell(x)}{\lambda a_\ell}}. \tag{81}$$

施加预算约束（80）得到

$$c = \sum_{k=0}^{L} \sqrt{\frac{w_k(x)a_k}{\lambda}}, \qquad \sqrt{\lambda} = \frac{1}{c}\sum_{k=0}^{L}\sqrt{w_k(x)a_k}, \qquad M_\ell^\star = c\frac{\sqrt{w_\ell(x)/a_\ell}}{\sum_{k=0}^{L}\sqrt{w_k(x)a_k}}. \tag{82}$$

□

References

[Ben18] E. Benhamou. “样本方差的若干性质”。载于：arXiv 预印本 arXiv:1809.03774（2018）。[Bla+23] J. Blanchet 等。“丢弃训练在分布鲁棒意义下是最优的”。载于：《机器学习研究杂志》24.180（2023），第 1–60 页。T. F. Chan、G. H. Golub 和 R. J. LeVeque。《计算样本方差的更新公式与成对算法》。技术报告 STAN-CS-79-773。斯坦福大学计算机科学系，1979 年 11 月。[Cli+11] K. A. Cliffe 等。“多层蒙特卡洛方法及其在随机系数椭圆型偏微分方程中的应用”。载于：《科学计算与可视化》14.1（2011），第 3 页。[Cox+24] A. M. Cox 等。“质子疗法剂量递送验证的贝叶斯逆向方法”。载于：《英国皇家学会学报 A 辑》。第 480 卷，2301 期。英国皇家学会，2024 年，第 20230836 页。[FS21] M. Fujisawa 和 I. Sato。“多层蒙特卡洛变分推断”。载于：《机器学习研究杂志》22.278（2021），第 1–44 页。[Gal16] Y. Gal。《深度学习中的不确定性》。博士论文。剑桥大学，2016 年。[Gaw+23] J. Gawlikowski 等。“深度神经网络不确定性综述”。载于：《人工智能评论》56.增刊 1（2023），第 1513–1589 页。[Ger+21] T. Gerstner 等。“多层蒙特卡洛学习”。载于：arXiv 预印本 arXiv:2102.08734（2021）。[GG16] Y. Gal 和 Z. Ghahramani。“将随机失活作为贝叶斯近似：表示深度学习中的模型不确定性”。载于：国际机器学习会议。PMLR，2016 年，第 1050–1059 页。[Gil08] M. B. Giles。“多层蒙特卡洛路径模拟”。载于：《运筹学》56.3（2008），第 607–617 页。[Gil15] M. B. Giles。“多层蒙特卡洛方法”。载于：《数值学报》24（2015），第 259–328 页。[Guo+17] C. Guo 等。“关于现代神经网络的校准”。载于：国际机器学习会议。PMLR，2017 年，第 1321–1330 页。[Hei01] S. Heinrich。“多层蒙特卡洛方法”。载于：国际大规模科学计算会议。施普林格，2001 年，第 58–67 页。[Hik+25] Y. Hikida 等。“多层神经模拟推断”。载于：arXiv 预印本 arXiv:2506.0608（2025）。[HW21] E. H¨ullermeier 和 W. Waegeman。“机器学习中的偶然不确定性与认知不确定性：概念与方法简介”。载于：《机器学习》110.3（2021），第 457–506 页。[KG17] A. Kendall 和 Y. Gal。“计算机视觉贝叶斯深度学习需要哪些不确定性？”载于：《神经信息处理系统进展》30（2017）。

7

[LPB17] B. Lakshminarayanan、A. Pritzel 和 C. Blundell。“使用深度集成进行简单且可扩展的预测不确定性估计”。载于：《神经信息处理系统进展》30（2017）。[MBK21] X. Meng、H. Babaee 和 G. E. Karniadakis。“多保真贝叶斯神经网络：算法与应用”。载于：《计算物理学杂志》438（2021），第 110361 页。[MPP24] C. G. Makridakis、A. Pim 和 T. Pryer。“一种用于物理信息神经网络和深度 Ritz 方法边界条件的深度 Uzawa-Lagrange 乘子方法”。载于：*arXiv* 预印本 *arXiv:2411.08702*（2024）。[PP25] A. Pim 和 T. Pryer。“使用蒙特卡洛 Dropout 不确定性量化对质子剂量进行代理建模”。载于：*arXiv* 预印本 *arXiv:2509.18155*（2025）。[RPK19] M. Raissi、P. Perdikaris 和 G. E. Karniadakis。“物理信息神经网络：用于求解涉及非线性偏微分方程的正问题和逆问题的深度学习框架”。载于：《计算物理学杂志》378（2019），第 686–707 页。[Sri+14] N. Srivastava 等。“随机失活：一种防止神经网络过拟合的简单方法”。载于：《机器学习研究杂志》15.1（2014），第 1929–1958 页。[YMK21] L. Yang、X. Meng 和 G. E. Karniadakis。“贝叶斯物理信息神经网络：用于含噪数据正问题和逆问题的贝叶斯物理信息神经网络”。载于：《计算物理学杂志》425（2021），第 109913 页。

[1] Institute for Mathematical Innovation University of Bath Bath UK. [2] Department of Mathematical Sci- ,,,,„ ences University of Bath Bath UK.