arXiv:2301.03415v1 [cs.PL] 9 Jan 2023

通过平滑实现概率编程快速且正确的基于梯度的优化

Basim Khajwal[1],C.-H. Luke Ong[1,2],和 Dominik Wagner[1](✉)

[1] 牛津大学[2] 南洋理工大学

摘要。 我们研究概率编程中变分推断的基础。变分推断将后验推断表述为一个优化问题。实践中占主导地位的优化方法是随机梯度下降。尤其是，采用所谓的重参数化梯度估计器的一种变体，在传统统计学场景中表现出快速收敛。不幸的是，编程语言很容易表达的不连续性，可能会损害这种方法的正确性。我们考虑一种带有条件表达式的简单（高阶、概率）编程语言，并为其赋予可测值语义和平滑的（近似）值语义。我们提出了能够确立技术性前提条件的类型系统。因此，我们可以证明，应用于平滑问题时，采用重参数化梯度估计器的随机梯度下降是正确的。此外，通过适当选择精度系数，我们可以将原始问题求解到任意误差容限以内。实证结果表明，我们的方法收敛表现与一种主要竞争方法相近，但更简单、更快，而且经工作量归一化的方差降低了几个数量级。

关键词： 概率编程 · 变分推断 · 重参数化梯度 · 值语义 · 类型系统。

1 引言

概率编程是一种编程范式，其愿景是让广大受众都能使用统计方法，尤其是贝叶斯推断。这是通过关注点分离实现的：希望获得统计洞见的领域专家专注于建模，而推断则自动执行。（在一些近期系统中 [4,9]，用户可以通过编写自己的推断代码来提高效率。）

从本质上说，概率编程语言通过加入 **Score** 或 **观测** （以及**采样**）等构造，扩展了更传统的编程语言，用于定义先验 $p(\mathbf{z})$ 和似然 $p(\mathbf{x} \mid \mathbf{z})$。推断的任务是推导后验分布 $p(\mathbf{z} \mid \mathbf{x})$；原则上，该后验分布由贝叶斯定律支配，但通常难以求解。

尽管这一范式最初是在统计学和贝叶斯机器学习的背景下提出的，但近年来概率编程

已成为编程语言社区中一个极富成果的研究主题。研究人员在理论方面做出了重要贡献，例如以严谨的（范畴论）语义为语言奠定基础 [37,36,16,39,12,10]，以及研究推断算法的正确性 [17,7,22]。后者大多是在“传统”统计学背景下设计的，而条件语句等在编程中无处不在的特性，对正确性构成了重大挑战。

推断算法大致分为两类：马尔可夫链蒙特卡罗（MCMC）会生成一个样本序列，该序列渐近地趋近于真实后验分布；另一类是变分推断。

变分推断。 在贝叶斯统计的变分推断方法中，近似难以计算的后验概率分布这一问题被转化为一个优化问题。其思路是使用潜变量$\mathbf{z}$上的一族“更简单”的密度 $q_{\boldsymbol{\theta}}(\mathbf{z})$，来近似后验概率 $p(\mathbf{z} \mid \mathbf{x})$，该密度族由 $\boldsymbol{\theta}$ 参数化。接下来，优化问题就是寻找参数 $\boldsymbol{\theta}^*$，使得 $q_{\boldsymbol{\theta}^*}(\mathbf{z})$ 与真实后验分布 $p(\mathbf{z} \mid \mathbf{x})$“尽可能接近”。由于变分族可能不包含真实后验分布，因此一般来说，$q_{\boldsymbol{\theta}^*}$ 是一种近似。在实践中，变分推断已被证明能够比 MCMC 快得多地得到良好的近似。

形式上，这一思想可通过最小化*KL*散度 [30,5] 来表达，即变分近似与真实后验分布之间的 KL 散度。这等价于最大化 ELBO函数，该函数只依赖于联合分布$p(\mathbf{x}, \mathbf{z})$，而不 依赖于我们最终要推断的后验分布：

$$\mathrm{ELBO}_{\boldsymbol{\theta}} \coloneqq \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\theta}}(\mathbf{z})}[\log p(\mathbf{x}, \mathbf{z}) - \log q_{\boldsymbol{\theta}}(\mathbf{z})] \tag{1}$$

基于梯度的优化。 实际中，常采用随机梯度下降（*SGD*）的各种变体来求解以下形式的优化问题：$\operatorname{argmin}_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{s} \sim q(\mathbf{s})}[f(\boldsymbol{\theta}, \mathbf{s})]$。在最简单的版本中，SGD 在每一步都沿着梯度的蒙特卡罗估计进行更新：

$$\boldsymbol{\theta}_{k+1} \coloneqq \boldsymbol{\theta}_k - \gamma_k \cdot \underbrace{\frac{1}{N} \sum_{i=1}^{N} \nabla_\theta f\left(\boldsymbol{\theta}_k, \mathbf{s}_k^{(i)}\right)}_{\text{gradient estimator}}$$

其中$\mathbf{s}_k^{(i)} \sim q\left(\mathbf{s}_k^{(i)}\right)$，而 γ_k 是步长。

为了保证 SGD 的正确性，梯度估计必须是无偏的，即其期望值正确：

$$\mathbb{E}_{\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(N)} \sim q}\left[\frac{1}{N} \sum_{i=1}^{N} \nabla_\theta f\left(\boldsymbol{\theta}, \mathbf{s}^{(i)}\right)\right] = \nabla_\theta \mathbb{E}_{\mathbf{s} \sim q(\mathbf{s})}[f(\boldsymbol{\theta}, \mathbf{s})]$$

这一关于微分与积分可交换的性质，可以由控制收敛定理证明 [21,定理 6.28]。

请注意，我们无法直接用蒙特卡罗估计式 (1) 中证据下界的梯度，因为计算期望所依据的分布也取决于参数。不过，可以使用所谓的 对数导数技巧 推导出无偏估计器，该估计器称为 *Score* 或 *REIN-FORCE* 估计器 [31,40,27]。

重参数化梯度。 尽管 Score 估计量适用范围极广，但不幸的是它存在较高的方差，这可能导致随机梯度下降得出很差的结果[3]。

重参数化梯度估计器是变分推断中的主流方法，它通过微分同胚变换 $\boldsymbol{\phi_\theta}$（例如位置尺度变换或累积分布函数），将潜变量 $\mathbf{z}$ 表示为基随机变量 $\mathbf{s}$（视为熵源）的函数。例如，如果潜变量 z的分布是参数为 $\boldsymbol{\theta} = \{\mu, \sigma\}$ 的高斯分布 $\mathcal{N}(z \mid \mu, \sigma^2)$，那么以标准正态分布为基分布的位置尺度变换可得到如下重参数化形式

$$z \sim \mathcal{N}(z \mid \mu, \sigma^2) \iff z = \phi_{\mu,\sigma}(s), \quad s \sim \mathcal{N}(0, 1). \tag{2}$$

其中 $\phi_{\mu,\sigma}(s) := s \cdot \sigma + \mu$。这种设置（通常称为“重参数化技巧” [20,38,32]）的主要优势在于，我们消除了分布对 $\boldsymbol{\theta}$的依赖，而期望正是相对于该分布计算的。因此，现在我们可以通过从基分布 $\mathbf{s}$中抽样进行蒙特卡罗模拟，并对变分分布的参数 $\boldsymbol{\theta}$ 求导（通过反向传播）。简而言之，我们有

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\theta}}(\mathbf{z})}[f(\boldsymbol{\theta}, \mathbf{z})] = \nabla_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{s} \sim q(\mathbf{s})}[f(\boldsymbol{\theta}, \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s}))] = \mathbb{E}_{\mathbf{s} \sim q(\mathbf{s})}[\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s}))]$$

重参数化梯度估计器的主要优点是，其方差显著低于 Score 估计量，因此收敛速度更快。

重参数化梯度的偏差。 遗憾的是，对于使用条件表达式即可轻松表达的不可微模型 [23],，重参数化梯度估计器存在偏差:

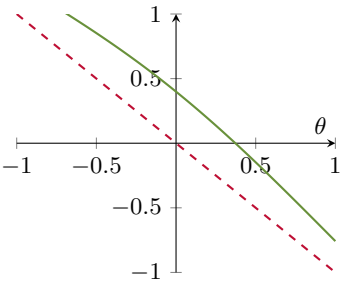
例*1*。 [23,命题 2], 中的反例，其目标函数是一个不可微模型的证据下界，可简化为

$$f(\theta, s) = -0.5 \cdot \theta^2 + \begin{cases} 0 & \text{if } s + \theta < 0 \\ 1 & \text{otherwise} \end{cases}$$

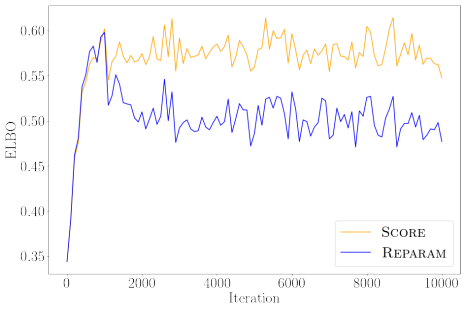
请注意（见图1a）：

$$\nabla_\theta \, \mathbb{E}_{s \sim \mathcal{N}(0,1)} \left[f(\theta, s) \right] = -\theta + \mathcal{N}(-\theta \mid 0, 1) \neq -\theta = \mathbb{E}_{s \sim \mathcal{N}(0,1)} \left[\nabla_\theta f(\theta, s) \right]$$

[3] see e.g. Fig. 5a



(a) 虚线 red: 有偏的 估计量器 $\mathbb{E}_{s\sim\mathcal{N}(0,1)}[\nabla_\theta f(\theta,s)]$, 实线绿色: 真实梯度 $\nabla_\theta \mathbb{E}_{s\sim\mathcal{N}(0,1)}[f(\theta,s)]$。



(b) ELBO轨迹（越高越好——较佳）得自 with 我们的实现（参见第7节）

图1：例1中重参数化梯度估计器的偏差。

关键在于这可能会妨碍收敛到临界点或极大值点：即使我们能找到一个使梯度估计器为零的点，它也未必是原优化问题的临界点（更不用说最优点了）（参见图1b）

非正式方法

我们的出发点是一个简单类型λ演算的变体，其中包含实数、条件表达式和采样构造。我们将证据下界的优化抽象为以下一般优化问题

$$\operatorname{argmin}_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[[\![M]\!](\boldsymbol{\theta},\mathbf{s})] \tag{3}$$

其中，$[\![M]\!]$ 是程序 M 的值函数 [7,26]，而 $\mathcal{D}$ 与参数 $\boldsymbol{\theta}$ 无关，它由 M进行采样所依据的分布决定。由于存在条件表达式，函数 $[\![M]\!]$ 可能不连续，更不用说可微了。

例1可表示为

$$(\lambda z.\, -0.5\cdot\theta^2 + (\mathbf{if}\ z<0\ \mathbf{then}\ 0\ \mathbf{else}\ 1))\,(\mathbf{sample}_{\mathcal{N}} + \theta)$$

我们的方法基于对（新）笛卡尔闭范畴**VectFr**中程序的指称语义$[\![(-)]\!]_\eta$（精度系数为 $\eta>0$）。该范畴推广了光滑流形，并为 Frölicher 空间（参见例如 [13,35]）赋予了向量空间结构。

直观地说，我们将解释条件表达式时通常出现的 Heaviside 阶跃函数替换为平滑近似。具体而言，我们将例1中的条件语句解释为

$$[\![\mathbf{if}\ s+\theta<0\ \mathbf{then}\ \underline{0}\ \mathbf{else}\ \underline{1}]\!]_\eta(\theta,s) := \sigma_\eta(s+\theta)$$

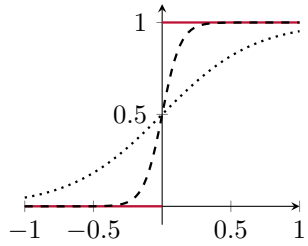


图2：（逻辑）S 形函数 σ_η（点线：$\eta=\frac{1}{3}$，虚线：$\eta=\frac{\Gamma}{15}$）和海维赛德阶跃函数（红色实线）。

其中 σ_η 是一个平滑函数。例如，我们可以选择 $\sigma_\eta(x) := \sigma(\frac{x}{\eta})$ where $\sigma(x) := \frac{1}{1+\exp(-x)}$ 为（逻辑）Sigmoid 函数（参见图 2）。因此，程序 M 可由一个平滑函数 $[\![M]\!]_\eta$ 来解释，其重参数化梯度可以无偏地估计。因此，我们对平滑后的程序应用随机梯度下降。

贡献

本文的主要贡献是为概率编程中的正确且高效的（变分）推断奠定理论基础。我们采用程序的平滑解释来获得无偏的（重参数化）梯度估计量，并通过类型系统确立必要的技术前提。具体而言：

1. 我们提出一种带条件表达式的简单（高阶）编程语言。我们采用轨迹类型来精确刻画在完全急切的按值调用求值策略中抽取的样本。 2. 我们为该语言赋予了可测的指称值语义和一种平滑（因而近似的）值语义。对于后者，我们基于 Frölicher 空间构造了一个范畴模型。 3. 我们开发了用于确保关键技术前提的类型系统：重参数化梯度估计器的无偏性、随机梯度下降的正确性，以及平滑过程一致收敛到原始问题。因此，原则上，我们的平滑方法能够在任意误差容限内给出正确解。 4. 我们进行了实证评估，结果表明，我们的方法具有与 [23] 采用的重参数化梯度估计器无偏校正相似的收敛性——这是我们的主要基线。不过，我们的估计器更简单、更高效：速度更快，且工作归一化方差降低了数个数量级。

概述。 下一节将介绍一种简单的高阶概率编程语言、它的指称值语义和操作语义，并提出优化问题 1。第 3 节将专门讨论平滑后的指称值语义，并提出平滑优化问题 2。在第 4 节和第 5 节中，我们将构建基于注解的类型系统，分别保证随机梯度下降的正确性和平滑过程的收敛性。第 6 节将简要讨论相关工作，随后在第 7 节介绍实证评估结果。最后，我们将在第 8 节总结全文并讨论未来方向。

符号约定。我们采用以下约定：向量和列表使用粗体，$+\!\!+$ 列表连接使用 $\nabla_{\boldsymbol{\theta}}$ 梯度（关于 $\boldsymbol{\theta}$），$[\phi]$ 谓词的 Iverson 括号使用 ϕ 分布使用花体字，特别是 $\mathcal{N}$ 正态分布。此外，我们使用红色标出值得注意的内容。

2 一种简单的编程语言

本章节将介绍我们的编程语言。它是带实数的简单类型 λ 演算，并扩展了条件表达式和从连续分布中采样的功能。

2.1 语法

编程语言的原始项由以下文法定义：

$$\begin{aligned} M ::= {} & x \mid \theta_i \mid \underline{r} \mid \underline{+} \mid \underline{\cdot} \mid \underline{-} \mid \underline{^{-1}} \mid \underline{\exp} \mid \underline{\log} \\ & \mid \mathbf{if}\, M < 0\, \mathbf{then}\, M\, \mathbf{else}\, M \mid \mathbf{sample}_{\mathcal{D}} \mid \lambda x.\, M \mid M\, M \end{aligned}$$

其中， x 和 θ_i 分别取值于（可数集合中的）变量和参数， $r \in \mathbb{R}$，而 $\mathcal{D}$ 是定义在 $\mathbb{R}$ 上的概率分布（其支撑集可能是 $\mathbb{R}$的真子集）。按照惯例，我们使用中缀、后缀和前缀表示法：$M + N$（加法）、 $M \cdot N$（乘法）、 M^{-1}（逆）以及 $\underline{-}M$（数值取负）。为使表达更简洁，我们经常省略下划线。

例*2*（为变分推断编码证据下界）。我们考虑 [23] 在其命题 2 中用于证明重参数化梯度存在偏差的例子。（例1中我们讨论了该例的一个简化版本。）联合密度为

$$p(z) \coloneqq \mathcal{N}(z \mid 0, 1) \cdot \begin{cases} \mathcal{N}(0 \mid -2, 1) & \text{if } z < 0 \\ \mathcal{N}(0 \mid 5, 1) & \text{otherwise} \end{cases}$$

并且他们使用密度为 $q_\theta(z) \coloneqq \mathcal{N}(z \mid \theta, 1)$ 的变分族，该变分族使用标准正态噪声分布和变换$s \mapsto s + \theta$进行重参数化。

首先，我们为均值为 m 、标准差为 s的正态分布的概率密度函数定义一个辅助项：

$$N \equiv \lambda x, m, s.\, \left(\underline{\sqrt{2\pi}} \cdot s\right)^{-1} \cdot \underline{\exp}\left(\underline{-0.5} \cdot \left((x + (-m)) \cdot s^{-1}\right)^2\right)$$

然后，我们可以定义

$$M \equiv \big(\lambda z.\, \underbrace{\underline{\log}\,(N\, z\, \underline{0}\, \underline{1}) + (\mathbf{if}\, z < 0\, \mathbf{then}\, \underline{\log}\,(N\, \underline{0}\, \underline{(-2)}\, \underline{1})\, \mathbf{else}\, \underline{\log}\,(N\, \underline{0}\, \underline{5}\, \underline{1}))}_{\log p} - \underbrace{\underline{\log}\,(N\, z\, \theta\, \underline{1})}_{\log q_\theta}\big)\, \big(\mathbf{sample}_{\mathcal{N}} + \theta\big)$$

2.2 基于轨迹的类型系统

类型由基本类型（R 和 $R_{>0}$，实数和正实数）以及轨迹类型（通常为 Σ，即有限个概率分布构成的列表）生成

以及形式为 $\tau \bullet \Sigma \to \tau'$的基于轨迹的函数空间构造子。形式上，类型由以下文法定义：

轨迹类型	$\Sigma ::= [\mathcal{D}_1, \ldots, \mathcal{D}_n]$	$n \geq 0$
基本类型	$\iota ::= R \mid R_{>0}$	
安全类型	$\sigma ::= \iota \mid \sigma \bullet [] \to \sigma$	
类型	$\tau ::= \iota \mid \tau \bullet \Sigma \to \tau$	

其中， $\mathcal{D}_i$ 是概率分布。直观地说，轨迹类型描述了概率程序的执行轨迹空间。使用轨迹类型后，我们的类型系统有一个鲜明特点：程序的类型能够精确刻画其可能执行轨迹的空间 [24]。对于轨迹类型，我们使用列表拼接记法 $+\!\!+$ ，并将形式为 $\tau_1 \bullet [] \to \tau_2$的函数类型简写为 $\tau_1 \to \tau_2$ 。直观地说，如果一个项在给定类型为 τ的值后，使用 Σ中的所有样本归约为类型为 τ'的值，那么它的类型就是 $\tau \bullet \Sigma \to \tau'$ 。

形式为 $\Gamma \mid \Sigma \vdash M : \tau$的双上下文类型判断定义于图3b中，其中 $\Gamma = x_1 : \tau_1, \cdots, x_n : \tau_n, \theta_1 : \tau'_1, \cdots, \theta_m : \tau'_m$ 是一个有限映射，用于描述一组变量类型绑定和参数类型绑定；轨迹类型Σ 精确刻画了在对项 M进行（完全急切的）按值调用求值时，抽取样本所依据的分布。

类型的子类型关系（如图3a所定义）基本上是标准的；对于上下文，我们定义：$\Gamma \sqsubseteq \Gamma'$若对每个 $x : \tau$，在 Γ中都存在一个 $x : \tau'$，在 Γ'中使得 $\tau' \sqsubseteq \tau$。

轨迹类型是唯一的（参见附录 A.1）：

引理 1. 若 $\Gamma \mid \Sigma \vdash M : \tau$ 且 $\Gamma \mid \Sigma' \vdash M : \tau'$，则 $\Sigma = \Sigma'$。

若一个项不包含**采样**$_{\mathcal{D}}$，或者 σ 是基类型，则该项具有安全类型 σ 。因此，虽然这可能稍显令人困惑，我们有$\mid [\mathcal{D}] \vdash$ **采样**$_{\mathcal{D}} : R$，而 R被视为安全类型。请注意，我们用元变量 σ 表示安全类型。

条件表达式。 条件表达式的分支必须具有安全类型。否则，就不清楚如何为诸如以下的项确定类型：

$$M \equiv \mathbf{if}\, x < 0\, \mathbf{then}\, (\lambda x.\, \mathbf{sample}_{\mathcal{N}})\, \mathbf{else}\, (\lambda x.\, \mathbf{sample}_{\mathcal{E}} + \mathbf{sample}_{\mathcal{E}})$$
$$N \equiv (\lambda f.\, f\, (f\, \mathbf{sample}_{\mathcal{N}}))\, M$$

因为分支会从不同的分布中抽取不同数量的样本，并且它们的类型分别为 $R \bullet [\mathcal{N}] \to R$ 和 $R \bullet [\mathcal{E}, \mathcal{E}] \to R$。然而，对于$M' \equiv \mathbf{if}\ x < 0\, \mathbf{then}\, \mathbf{sample}_{\mathcal{N}}\, \mathbf{else}\, \mathbf{sample}_{\mathcal{E}} + \mathbf{sample}_{\mathcal{E}}$，我们可以安全地进行类型判定

$$\frac{\dfrac{x : R \mid [\mathcal{N}, \mathcal{E}, \mathcal{E}] \vdash M' : R}{\mid [] \vdash \lambda x.\, M' : R \bullet [\mathcal{N}, \mathcal{E}, \mathcal{E}] \to R}}{\mid [\mathcal{N}, \mathcal{N}, \mathcal{E}, \mathcal{E}, \mathcal{N}, \mathcal{E}, \mathcal{E}] \vdash (\lambda f.\, f\, (f\, \mathbf{sample}_{\mathcal{N}}))\, (\lambda x.\, M') : R}$$

$$\frac{}{\iota \sqsubseteq \iota} \qquad \frac{}{R_{>0} \sqsubseteq R} \qquad \frac{\tau_1' \sqsubseteq \tau_1 \quad \tau_2 \sqsubseteq \tau_2'}{(\tau_1 \bullet \Sigma \to \tau_2) \sqsubseteq (\tau_1' \bullet \Sigma \to \tau_2')}$$

(a) Subtyping

$$\frac{\Gamma \mid \Sigma \vdash M : \tau}{\Gamma' \mid \Sigma \vdash M : \tau'}\ \Gamma \sqsubseteq \Gamma', \tau \sqsubseteq \tau' \qquad \frac{}{x : \tau \mid [] \vdash x : \tau}$$

$$\frac{}{\mid [] \vdash \underline{r} : R}\ r \in \mathbb{R} \qquad \frac{}{\mid [] \vdash \underline{r} : R_{>0}}\ r \in \mathbb{R}_{>0}$$

$$\frac{}{\mid [] \vdash \underline{\circ} : R \to R \to R}\ \circ \in \{+, \cdot\} \qquad \frac{}{\mid [] \vdash \underline{\circ} : R_{>0} \to R_{>0} \to R_{>0}}\ \circ \in \{+, \cdot\}$$

$$\frac{}{\mid [] \vdash \underline{-} : R \to R} \qquad \frac{}{\mid [] \vdash \underline{^{-1}} : R_{>0} \to R_{>0}}$$

$$\frac{}{\mid [] \vdash \underline{\exp} : R \to R_{>0}} \qquad \frac{}{\mid [] \vdash \underline{\log} : R_{>0} \to R}$$

$$\frac{\Gamma \mid \Sigma \vdash L : R \quad \Gamma \mid \Sigma' \vdash M : \sigma \quad \Gamma \mid \Sigma'' \vdash N : \sigma}{\Gamma \mid \Sigma +\!\!+ \Sigma' +\!\!+ \Sigma'' \vdash \mathbf{if}\, L < 0\, \mathbf{then}\, M\, \mathbf{else}\, N : \sigma} \qquad \frac{}{\mid [\mathcal{D}] \vdash \mathbf{sample}_{\mathcal{D}} : R}$$

$$\frac{\Gamma, y : \tau_1 \mid \Sigma \vdash M : \tau_2}{\Gamma \mid [] \vdash \lambda y.\, M : \tau_1 \bullet \Sigma \to \tau_2} \qquad \frac{\Gamma \mid \Sigma_1 \vdash M : \tau_1 \bullet \Sigma_3 \to \tau_2 \quad \Gamma \mid \Sigma_2 \vdash N : \tau_1}{\Gamma \mid \Sigma_1 +\!\!+ \Sigma_2 +\!\!+ \Sigma_3 \vdash M\, N : \tau_2}$$

(b) Typing judgments

图 3：基本的基于轨迹的类型系统

示例 *3.* 考虑以下项：

$$L \equiv \lambda x.\, \mathbf{sample}_{\mathcal{N}} + \mathbf{sample}_{\mathcal{N}}$$
$$M \equiv \mathbf{if}\, x < 0\, \mathbf{then}\, (\lambda y.\, y + y)\, \mathbf{sample}_{\mathcal{N}}\, \mathbf{else}\, (\mathbf{sample}_{\mathcal{N}} + \mathbf{sample}_{\mathcal{N}})$$

我们可以推导出以下类型判断：

$$\begin{aligned} &\mid [] \vdash L : R_{>0} \bullet [\mathcal{N}, \mathcal{N}] \to R \\ x : R_{>0} &\mid [\mathcal{N}, \mathcal{N}, \mathcal{N}] \vdash M : R \\ &\mid [] \vdash \lambda x.\, M : R_{>0} \bullet [\mathcal{N}, \mathcal{N}, \mathcal{N}] \to R \\ &\mid [\mathcal{N}, \mathcal{N}, \mathcal{N}, \mathcal{N}] \vdash (\lambda x.\, M)\, \mathbf{sample}_{\mathcal{N}} : R \\ &\mid [\mathcal{N}, \mathcal{N}] \vdash (\lambda f.\, f\, (f\, 0))\, (\lambda x.\, \mathbf{sample}_{\mathcal{N}}) : R \end{aligned}$$

注意，$\mathbf{if}\ x < 0\, \mathbf{then}\, (\lambda x.\, \mathbf{采样}_{\mathcal{N}})\, \mathbf{else}\, (\lambda x.\, x)$ *not*是不可类型判定的。

2.3 指称值语义

接下来，我们为语言赋予（可测的）值语义。众所周知，可测空间与可测函数构成的范畴不是笛卡尔闭的 [1], ，这意味着 λ 演算无法解释为可测函数。

这些困难促使 [15] 人们发展了拟博雷尔空间的范畴 **QBS** 。我们在附录 A.2 中回顾其定义。值得注意的是，态射可以分段组合，而这正是我们处理条件表达式所需要的。

我们在拟博雷尔空间范畴 **QBS** 中解释我们的编程语言。类型的解释如下：

$$[\![R]\!] \coloneqq (\mathbb{R}, M_{\mathbb{R}}) \qquad [\![R_{>0}]\!] \coloneqq (\mathbb{R}_{>0}, M_{\mathbb{R}_{>0}}) \qquad [\![[\mathcal{D}_1, \ldots, \mathcal{D}_n]]\!] \coloneqq (\mathbb{R}, M_{\mathbb{R}})^n$$
$$[\![\tau_1 \bullet \Sigma \to \tau_2]\!] \coloneqq [\![\tau_1]\!] \times [\![\Sigma]\!] \Rightarrow [\![\tau_2]\!]$$

其中 $M_{\mathbb{R}}$ 是可测函数 $\mathbb{R} \to \mathbb{R}$ 的集合； $M_{\mathbb{R}_{>0}}$ 的情况也类似。（对于轨迹类型，我们使用列表记法（以及列表拼接）表示轨迹。）

我们首先定义一个便捷的辅助函数，用于解释应用。对于 f：$[\![\Gamma]\!] \times \mathbb{R}^{n_1} \Rightarrow [\![\tau_1 \bullet \Sigma_3 \to \tau_2]\!]$ 和 g：$[\![\Gamma]\!] \times \mathbb{R}^{n_2} \Rightarrow [\![\tau_1]\!]$，定义

$$f \mathbin{@} g : [\![\Gamma]\!] \times \mathbb{R}^{n_1+n_2+|\Sigma_3|} \Rightarrow [\![\tau_2]\!]$$
$$(\gamma, \mathbf{s}_1 +\!\!+ \mathbf{s}_2 +\!\!+ \mathbf{s}_3) \mapsto f(\gamma, \mathbf{s}_1)(g(\gamma, \mathbf{s}_2), \mathbf{s}_3) \quad \mathbf{s}_1 \in \mathbb{R}^{n_1}, \mathbf{s}_2 \in \mathbb{R}^{n_2}, \mathbf{s}_3 \in \mathbb{R}^{|\Sigma_3|}$$

我们对上下文中的项 $[\![\Gamma \mid \Sigma \vdash M\colon \tau]\!]\colon [\![\Gamma]\!] \times [\![\Sigma]\!] \to [\![\tau]\!]$ 作如下解释：

$$[\![\Gamma \mid [\mathcal{D}] \vdash \textbf{采样}\ _{\mathcal{D}}\colon \quad R]\!](\gamma, [s])\colon = s$$
$$[\![\Gamma \mid [] \vdash \lambda y.\, M\colon \tau_1 \bullet \Sigma \to \tau_2]\!](\gamma,\ [])\coloneqq$$
$$(v,\ \mathbf{s}) \in [\![\tau_1]\!] \times [\![\Sigma]\!] \mapsto [\![\Gamma,\ x\colon\ \tau_1 \mid \Sigma \vdash M\colon\ \tau_2]\!]((\gamma, v),\ \mathbf{s})$$
$$[\![\Gamma \mid \Sigma_1 +\!\!+ \Sigma_2 +\!\!+ \Sigma_3 \vdash M\, N\colon\ \tau]\!]\colon =$$
$$[\![\Gamma \mid \Sigma_1 \vdash M : \tau_1 \bullet \Sigma_3 \to \tau_2]\!] \mathbin{@} [\![\Gamma \mid \Sigma_2 \vdash N : \tau_1]\!]$$
$$[\![\Gamma \mid \Sigma_1 +\!\!+ \Sigma_2 +\!\!+ \Sigma_3 \vdash \textbf{如果}\ L < 0\, \textbf{那么}\ M\, \textbf{否则}\ N\colon\ \tau]\!](\gamma, \mathbf{s}_1 +\!\!+ \mathbf{s}_2 +\!\!+ \mathbf{s}_3))\colon =$$
$$\begin{cases} [\![\Gamma \mid \Sigma_2 \vdash M : \tau]\!](\gamma, \mathbf{s}_2) & \text{if } [\![\Gamma \mid \Sigma_1 \vdash L : R]\!](\gamma, \mathbf{s}_1) < 0 \\ [\![\Gamma \mid \Sigma_3 \vdash N : \tau]\!](\gamma, \mathbf{s}_3) & \text{otherwise} \end{cases}$$

不难看出，这种对上下文中的项的解释是定义良好且全定义的。对于条件子句，我们可以假设轨迹类型和轨迹分别表示为分割 $\Sigma_1 +\!\!+ \Sigma_2 +\!\!+ \Sigma_3$ 和$\mathbf{s}_1 +\!\!+ \mathbf{s}_2 +\!\!+\mathbf{s}_3$。这是合理的，因为由判断式 $\Gamma \mid \Sigma_1 +\!\!+\Sigma_2 +\!\!+ \Sigma_3 \vdash$ **如果** $L < 0$ **则** M **否则** $N\colon \tau$可证得 $\Gamma \mid \Sigma_1 \vdash L\colon R$, $\Gamma \mid \Sigma_2 \vdash M\colon \sigma$和 $\Gamma \mid \Sigma_3 \vdash N\colon \sigma$ ；并且根据引理 1，我们知道 Σ_1, Σ_2 和 Σ_3 各自都是唯一的；它们各自的长度进而决定了分割$\mathbf{s}_1 +\!\!+\mathbf{s}_2 +\!\!+\mathbf{s}_3$。应用子句的情况也类似：根据引理 1，分量 Σ_1 和 Σ_2是确定的，而 Σ_3 则由 M的类型确定。

2.4 与操作语义的关系

我们还可以为语言赋予一种基于大步 CBV 采样的语义，类似于附录 A 的图 6 中定义的 [7,26], 。我们用 $M \Downarrow_w^{\mathbf{s}} V$ 表示 M 通过执行轨迹 $\mathbf{s}$ 规约为值 V（它是一个实常数或抽象），并累积权重 w。

基于此，我们可以定义值函数和权重函数：

$$\mathrm{value}_M(\mathbf{s}) \coloneqq \begin{cases} V & \text{if } M \Downarrow_w^{\mathbf{s}} V \\ \mathrm{undef} & \text{otherwise} \end{cases} \qquad \mathrm{weight}_M(\mathbf{s}) \coloneqq \begin{cases} w & \text{if } M \Downarrow_w^{\mathbf{s}} V \\ 0 & \text{otherwise} \end{cases}$$

我们的语义有些不同寻常：对于条件表达式，我们会急切地计算两个分支。这样做的技术优势在于，对于每个（闭）上下文中的项，$\mid [\mathcal{D}_1, \cdots, \mathcal{D}_n] \vdash M \colon \iota$，$M$ 都能恰好使用类型判定中编码的长度所对应的轨迹归约为一个（唯一的）值，即 n。

因此，从这个意义上说，操作语义是“全定义的”：不存在发散。请注意，由于类型判定的作用，不会出现由偏函数原语（如 $1/x$）引起的非全定义情况。

此外，它与我们的指称值语义之间存在简单的联系：

命题 1。令 $\mid [\mathcal{D}_1, \ldots, \mathcal{D}_n] \vdash M \colon \iota$。则

1. $\mathrm{dom}(\mathrm{value}_M) = \mathbb{R}^n$ *2.* $[\![M]\!] = \mathrm{value}_M$ *3.* $\mathrm{权重}_M(\mathbf{s}) = \prod_{j=1}^{n} \mathrm{概率密度函数}_{\mathcal{D}_j}(s_j)$

2.5 问题陈述

我们终于可以正式陈述优化问题了：

问题 *1.* 优化

Given: term-in-context, $\theta_1 : \iota_1, \cdots, \theta_m : \iota_m \mid [\mathcal{D}_1, \ldots, \mathcal{D}_n] \vdash M : R$

求： $\mathrm{argmin}_{\boldsymbol{\theta}}\ \mathbb{E}_{s_1 \sim \mathcal{D}_1, \ldots, s_n \sim \mathcal{D}_n} [[\![M]\!](\boldsymbol{\theta}, \mathbf{s})]$

3 平滑指称值语义

现在我们转向平滑指称值语义，并用它来避免重参数化梯度估计器中的偏差。它由一族光滑函数 $\sigma_\eta : \mathbb{R} \to [0,1]$ 参数化。直观地说，我们用光滑近似（参见图 2）替代条件表达式解释中出现的 Heaviside 阶跃函数。具体而言，条件表达式 **如果** $z < 0$ **那么0 否则 1** 被解释为 $z \mapsto \sigma_\eta(z)$，而非 $[z \geq 0]$（使用 Iverson 括号）。

我们的主要示例是 $\sigma_\eta(x) \coloneqq \sigma(\frac{x}{\eta})$，其中 σ 是（逻辑）S形函数 $\sigma(x) \coloneqq \frac{1}{1+\exp(-x)}$，~~见图 2。~~虽然在此阶段，除平滑性外不需要其他性质，但稍后我们将需要限制 σ_η，使其具有良好的性质，尤其是能够收敛到赫维赛德阶跃函数。

我们提出向量 Frölicher 空间 *VectFr*$\mathbf{VectFr}$作为一种范畴模型。据我们所知，这是一种新的构造，可对平滑后的条件表达式作出简单直接的解释。

3.1 Frölicher 空间

我们回顾 Frölicher 空间的定义，它是对光滑空间的推广[4]：*Frölicher* 空间 是一个三元组 $(X, \mathcal{C}_X, \mathcal{F}_X)$，其中 X 是一个集合，$\mathcal{C}_X \subseteq \mathbf{Set}(\mathbb{R}, X)$ 是一个由曲线 组成的集合，而 $\mathcal{F}_X \subseteq \mathbf{Set}(X, \mathbb{R})$ 是一个由泛函组成的集合，满足

1. 如果 $c \in \mathcal{C}_X$ 和 $f \in \mathcal{F}_X$，则 $f \circ c \in C^\infty(\mathbb{R}, \mathbb{R})$ 2. 如果 $c\colon \mathbb{R} \to X$，使得对于所有 $f \in \mathcal{F}_X$, $f \circ c \in C^\infty(\mathbb{R}, \mathbb{R})$，则 $c \in \mathcal{C}_X$3. 如果 $f\colon X \to \mathbb{R}$，使得对于所有 $c \in \mathcal{C}_X$, $f \circ c \in C^\infty(\mathbb{R}, \mathbb{R})$，则 $f \in \mathcal{F}_X$。

Frölicher 空间 $(X, \mathcal{C}_X, \mathcal{F}_X)$ 与 $(Y, \mathcal{C}_Y, \mathcal{F}_Y)$ 之间的一个态射 是一个映射$\phi\colon X \to Y$，满足对于所有 $f \in \mathcal{F}_Y$和 $c \in \mathcal{C}_X$，$f \circ \phi \circ c \in C^\infty(\mathbb{R}, \mathbb{R})$。

Frölicher 空间及其态射构成一个范畴$\mathbf{Fr}$，众所周知，该范畴是笛卡尔闭的 [13,35]。

3.2 向量 Frölicher 空间

为了平滑地解释我们的编程语言，我们希望将条件表达式解释为其分支的 σ_η-加权凸组合：

$$[\![\textbf{如果 } L < 0 \textbf{ 则 } M \textbf{ 否则 } N]\!]_\eta(\gamma, \mathbf{s}_1 +\!\!+ \mathbf{s}_2 +\!\!+ \mathbf{s}_3) := \sigma_\eta(-[\![L]\!]_\eta(\gamma, \mathbf{s}_1)) \cdot [\![M]\!]_\eta(\gamma, \mathbf{s}_2) + \sigma_\eta([\![L]\!]_\eta(\gamma, \mathbf{s}_1)) \cdot [\![N]\!]_\eta(\gamma, \mathbf{s}_3) \quad (4)$$

根据目前的讨论，只有当分支具有基类型时，这才有意义，因为 Frölicher 空间没有配备向量空间结构，而我们要对态射取加权组合。特别是，如果ϕ_1，ϕ_2：$X \to Y$ 和 $\alpha\colon X \to \mathbb{R}$ 是态射，那么 $\alpha\,\phi_1 + \phi_2$ 也应该是一个态射。因此，我们为 Frölicher 空间增添额外的向量空间结构：

定义 1. 一个 $\mathbb{R}$-向量 Frölicher 空间 是一个 *Frölicher* 空间 $(X, \mathcal{C}_X, \mathcal{F}_X)$，满足 X 是一个 $\mathbb{R}$-向量空间，并且当 c、$c' \in \mathcal{C}_X$ 和 $\alpha \in C^\infty(\mathbb{R}, \mathbb{R})$ 时，$\alpha\, c + c' \in \mathcal{C}_X$（逐点定义）。

两个 $\mathbb{R}$-向量 Frölicher 空间 之间的 态射 是 *Frölicher* 空间之间的态射，即如果对于所有 $c \in \mathcal{C}_X$ 和 $f \in \mathcal{F}_Y$，$f \circ \phi \circ c \in C^\infty(\mathbb{R}, \mathbb{R})$，则 $\phi\colon (X, \mathcal{C}_X, \mathcal{F}_X) \to (Y, \mathcal{C}_Y, \mathcal{F}_Y)$ 是一个态射。

$\mathbb{R}$-向量 *Frölicher* 空间 及其态射构成一个范畴 $\mathbf{VectFr}$。显然存在一个遗忘函子，将 $\mathbf{VectFr}$ 完全忠实地嵌入 $\mathbf{Fr}$ 中。请注意，上述限制比要求 $\mathcal{C}_X$ 同样是一个向量空间要更强一些。（α 不一定是常数。）其主要益处如下，这对于按照公式（4）解释条件表达式至关重要：

引理 2。 如果 $\phi_1, \phi_2 \in \mathbf{VectFr}(X, Y)$ 且 $\alpha \in \mathbf{VectFr}(X, \mathbb{R})$，那么 $\alpha\,\phi_1 + \phi_2 \in \mathbf{VectFr}(X, Y)$（逐点定义）。

证明。 假设 $c \in \mathcal{C}_X$ 和 $f \in \mathcal{F}_Y$。那么 $(\alpha_1\,\phi_1 + \phi_2) \circ c = (\alpha \circ c) \cdot (\phi_1 \circ c) + (\phi_2 \circ c) \in \mathcal{C}_Y$（逐点定义），结论由此得出。

[4] $C^\infty(\mathbb{R}, \mathbb{R})$ 是平滑函数的集合 $\mathbb{R} \to \mathbb{R}$

与 Frölicher 空间类似，如果 X 是一个 $\mathbb{R}$-向量空间，那么任意 $\mathcal{C} \subseteq$集合$(X, \mathbb{R})$ 生成一个 $\mathbb{R}$-向量 Frölicher 空间（X，$\mathcal{C}_X$，$\mathcal{F}_X$），其中

$$\mathcal{F}_X \coloneqq \{f : X \to \mathbb{R} \mid \forall c \in \mathcal{C}.\, f \circ c \in C^\infty(\mathbb{R}, \mathbb{R})\}$$
$$\widetilde{\mathcal{C}}_X \coloneqq \{c : \mathbb{R} \to X \mid \forall f \in \mathcal{F}_X.\, f \circ c \in C^\infty(\mathbb{R}, \mathbb{R})\}$$
$$\mathcal{C}_X \coloneqq \left\{ \sum_{i=1}^{n} \alpha_i\, c_i \mid n \in \mathbb{N}, \forall i \leq n.\, \alpha_i \in C^\infty(\mathbb{R}, \mathbb{R}), c_i \in \widetilde{\mathcal{C}}_X \right\}$$

修改了由一组曲线生成的 Frölicher 空间的概念后，笛卡尔闭性证明仍然成立（更多细节见附录B），因此我们得出：

命题 2。 $\mathbf{VectFr}$ 是笛卡尔闭的。

3.3 平滑解释

至此，我们已经讨论了在笛卡尔闭范畴 $\mathbf{VectFr}$ 中（以平滑方式）解释我们语言所需的全部要素。我们将 $[\![M]\!]$ 的 η-平滑 称为 $[\![M]\!]_\eta$（或者宽泛地说，称为 M 的 η-平滑 ）。这种解释大体上是标准的，遵循第2.3节，条件表达式的情形除外。后者由公式 (4) 给出，而这需要额外的向量空间结构。

最后，我们可以表述优化问题 1 的平滑版本：

问题 *2.* η-平滑优化

给定： 上下文中的项，$\theta_1 : \iota_1, \cdots, \theta_m : \iota_m \mid [\mathcal{D}_1, \ldots, \mathcal{D}_n] \vdash M : R$，以及精度 系数 $\eta > 0$

求： $\operatorname{argmin}_{\boldsymbol{\theta}} \mathbb{E}_{s_1 \sim \mathcal{D}_1, \ldots, s_n \sim \mathcal{D}_n} [[\![M]\!]_\eta(\boldsymbol{\theta}, \mathbf{s})]$

4 平滑问题中随机梯度下降的正确性及重参数化梯度的无偏性

接下来，我们将随机梯度下降（随机梯度下降）与重参数化梯度估计器应用于平滑问题（批次大小为 $N = 1$）：

$$\boldsymbol{\theta}_{k+1} \coloneqq \boldsymbol{\theta}_k - \gamma_k \cdot \nabla_\theta [\![M]\!]_\eta (\boldsymbol{\theta}_k, \mathbf{s}_k) \qquad\qquad \mathbf{s}_k \sim \mathcal{D} \tag{5}$$

其中 $\boldsymbol{\theta} \mid [\mathbf{s} \sim \mathcal{D}] \vdash M : R$（在轨迹类型中稍微滥用了符号）。

步长序列的一种经典选择是 $\gamma_k \in \Theta(1/k)$，它满足所谓的 *Robbins-Monro* 准则：

$$\sum_{k \in \mathbb{N}} \gamma_k = \infty \qquad\qquad \sum_{k \in \mathbb{N}} \gamma_k^2 < \infty \tag{6}$$

在本章节中，我们希望确立将随机梯度下降过程应用于平滑公式 (5) 的正确性。

4.1 期望条件

首先，我们应当退一步，注意到由于缺乏可积性，我们试图解决的优化问题可能没有明确定义：取$M \equiv (\lambda x.\exp(x \cdot x))$ 采样 $_{\mathcal{N}}$：我们有 $\mathbb{E}_{z\sim\mathcal{N}}[[\![M]\!](z)] = \infty$，且与参数无关。因此，我们旨在保证：

(SGD0) 优化问题（平滑和未平滑的）都有良好定义。

由于 $\mathbb{E}[[\![M]\!]_\eta(\boldsymbol{\theta}, \mathbf{s})]$（以及 $\mathbb{E}[[\![M]\!](\boldsymbol{\theta}, \mathbf{s})]$）关于参数 $\boldsymbol{\theta}$可能不是凸函数，因此我们无法期望总能找到全局最优解。相反，我们寻求驻点，即关于参数 $\boldsymbol{\theta}$ 的梯度为零的点。以下结果（其证明是标准的）给出了随机梯度下降收敛到驻点的充分条件（例如参见 [3]或 [2,第 2]章）：

命题 3（收敛）。假设$(\gamma_k)_{k\in\mathbb{N}}$ 满足 *Robbins-Monro* 准则，见公式(6)，且 $g(\boldsymbol{\theta})$:$= \mathbb{E}_{\mathbf{s}}[f(\boldsymbol{\theta}, \mathbf{s})]$ 有良好定义。若 $\boldsymbol{\Theta} \subseteq \mathbb{R}^m$ 满足

(SGD1) 无偏性：$\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{s}}[\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, \mathbf{s})]$ 对所有 $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

(SGD2) g 在 $\boldsymbol{\Theta}$ 上是 *L-Lipschitz* 平滑的，其中 $L > 0$:

$$\|\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}')\| \leq L \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \qquad \textit{for all } \boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$$

(SGD3) 有界方差：$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{E}_{\mathbf{s}}[\|\nabla_{\boldsymbol{\theta}} f_k(\boldsymbol{\theta}, \mathbf{s})\|^2] < \infty$

那么，对于某个 $i \in \mathbb{N}$，$\inf_{i\in\mathbb{N}} \mathbb{E}[\|\nabla g(\boldsymbol{\theta}_i)\|^2] = 0$ 或 $\boldsymbol{\theta}_i \notin \boldsymbol{\Theta}$。

无偏性 (SGD1) 要求微分与积分可以交换。该操作的有效性可通过控制收敛定理 [21, 定理 6.28]，（见附录 C.1）来确立。要使其适用，f 关于参数的偏导数需要由一个可积函数一致地控制。形式化地说：

定义 2。设 $f\colon \boldsymbol{\Theta} \times \mathbb{R}^n \to \mathbb{R}$ 和 $g\colon \mathbb{R}^n \to \mathbb{R}$。如果对所有 $(\boldsymbol{\theta}、\mathbf{s}) \in \boldsymbol{\Theta} \times \mathbb{R}^n$，$|f(\boldsymbol{\theta}、\mathbf{s})| \leq g(\mathbf{s})$，我们称 g 一致支配 f。

还要注意，对于 Lipschitz 平滑性（SGD2），只需对二阶偏导数进行一致有界控制。

本章节接下来将介绍两种类型系统，它们会限制语言，以保证性质（SGD0）至（SGD3）。

4.2 分段多项式与有限矩分布

作为第一步说明，我们考虑一个类型系统 $\vdash_{\mathrm{poly}}$，它将项限制为（分段）多项式以及具有有限矩的分布。回顾一下，分布 $\mathcal{D}$ 具有有限矩，是指对所有 $p \in \mathbb{N}$，$\mathbb{E}_{s\sim\mathcal{D}}[|s|^p] < \infty$均成立。具有有限矩的分布包括以下常用分布：正态分布、指数分布、逻辑斯谛分布和伽马分布。柯西分布则是一个反例，众所周知，它甚至不存在期望。

定义 3。 对于具有有限矩的分布 $\mathcal{D}$， $f\colon \mathbb{R}^n \to \mathbb{R}$ 具有有限矩，是指对所有 $p \in \mathbb{N}$， $\mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[|f(\mathbf{s})|^p] < \infty$ 均成立。

具有有限矩的函数具有良好的闭包性质：

引理 3。 如果 $f, g\colon \mathbb{R}^n \to \mathbb{R}$ 具有（全部）有限矩，那么 $-f, f+g, f\cdot g$也具有有限矩。

特别地，如果一个分布具有有限矩，那么多项式也具有有限矩。因此，直观而言，只需（详细内容明确列于附录 C.2）：

1. 要求采样规则中的分布 $\mathcal{D}$ 具有有限矩：

$$\frac{}{\mid [\mathcal{D}] \vdash_{\mathrm{poly}} \mathbf{sample}_{\mathcal{D}} : R}\ \mathcal{D} \text{ has finite moments}$$

2. 从类型系统 $\vdash_{\mathrm{poly}}$中移除 $^{-1}$、exp 和 log 的规则。

类型可靠性 I：定义良好性。 今后，我们固定参数 $\theta_1 :\iota_1, \ldots, \theta_m\colon \iota_m$。直观来看，只要 $\boldsymbol{\theta} \mid \Sigma \vdash_{\mathrm{poly}} M\colon \iota$， $[\![M]\!]$ 显然就是分段多项式。尽管如此，我们仍对这一性质给出形式化证明，以说明本文其余部分所采用的证明技巧——逻辑关系的一种变体。

我们在 $\boldsymbol{\Theta} \times \mathbb{R}^n \to [\![\tau]\!]$上定义一个稍强的逻辑谓词 $\mathcal{P}_\tau^{(n)}$，从而得到一个一致上界：

1. $f \in \mathcal{P}_\iota^{(n)}$若 f 被一个具有有限二阶矩的函数一致受控 2. $f \in \mathcal{P}_{\tau_1 \bullet \Sigma_3 \to \tau_2}^{(n)}$ 若对所有 $n_2 \in \mathbb{N}$ 以及 $g \in \mathcal{P}_{\tau_1}^{(n+n_2)}$， $f \odot g \in \mathcal{P}_{\tau_2}^{(n+n_2+|\Sigma_3|)}$

其中，对于 $f\colon \boldsymbol{\Theta} \times \mathbb{R}^{n_1} \to [\![\tau_1 \bullet \Sigma_3 \to \tau_2]\!]$ 和 $g\colon \boldsymbol{\Theta} \times \mathbb{R}^{n_1+n_2} \to [\![\tau_1]\!]$，我们定义

$$\begin{aligned} f \odot g : \boldsymbol{\Theta} \times \mathbb{R}^{n_1+n_2+|\Sigma_3|} &\to \tau_2 \\ (\boldsymbol{\theta}, \mathbf{s}_1 +\!\!+ \mathbf{s}_2 +\!\!+ \mathbf{s}_3) &\mapsto f(\boldsymbol{\theta}, \mathbf{s}_1)(g(\boldsymbol{\theta}, \mathbf{s}_1 +\!\!+ \mathbf{s}_2), \mathbf{s}_3) \end{aligned}$$

直观来说， g 可能依赖于 $\mathbf{s}_2$ 中的样本（以及 $\mathbf{s}_1$），而函数应用可能会消耗更多样本 $\mathbf{s}_3$ （由轨迹类型Σ_3决定）。通过对安全类型进行归纳法，我们证明以下结果，这对条件表达式很重要：

引理 4. 如果 $f \in \mathcal{P}_\iota^{(n)}$ 和 $g,\ h \in \mathcal{P}_\sigma^{(n)}$，那么 $[f(-) < 0]\cdot g + [f(-) \geq 0]\cdot h \in \mathcal{P}_\sigma^{(n)}$。

证明。 对于基本类型，这由引理 3 得出。因此，假设 σ 的形式为$\sigma_1 \bullet [] \to \sigma_2$。令 $n_2 \in \mathbb{N}$ 和 $x \in \mathcal{P}_{\sigma_1}^{n+n_2}$。根据定义，$(g \odot x)$，$(h \odot x) \in \mathcal{P}^{(n+n_2)\sigma_2}$。令 $\widehat{f}$ 为将 f 扩展到 $\boldsymbol{\Theta} \times \mathbb{R}^{n+n_2} \to \mathbb{R}$（忽略额外样本）所得的结果。容易看出，同样有 $\widehat{f} \in \mathcal{P}_\iota^{(n+n_2)}$ 根据归纳假设，

$$[\widehat{f}(-) < 0] \cdot (g \odot x) + [\widehat{f}(-) \geq 0] \cdot (h \odot x) \in \mathcal{P}_{\sigma_2}^{(n+n_2)}$$

最后，根据定义，

$$([f(-) < 0] \cdot g + [f(-) \geq 0] \cdot h) \odot x = [\widehat{f}(-) < 0] \cdot (g \odot x) + [\widehat{f}(-) \geq 0] \cdot (h \odot x)$$

假设 1 我们假设 $\boldsymbol{\Theta} \subseteq [\![\iota_1]\!] \times \cdots \times [\![\iota_m]\!]$ 是紧致的。

引理 5（基本）。如果 $\boldsymbol{\theta},\ x_1 :\ \tau_1,\ \ldots\ldots,\ x_\ell :\ \tau_\ell \mid \Sigma \vdash_{\mathrm{poly}} M :\ \tau,$ $n \in \mathbb{N},\ \xi_1 \in \mathcal{P}^{(n)}_{\tau_1},\ \ldots\ldots,\ \xi_\ell \in \mathcal{P}^{(n)}_{\tau_\ell}$那么 $[\![M]\!] * \langle \xi_1, \ldots, \xi_\ell \rangle \in \mathcal{P}^{(n+|\Sigma|)}_{\tau},$ 其中

$$[\![M]\!] * \langle \xi_1, \ldots, \xi_\ell \rangle : \boldsymbol{\Theta} \times \mathbb{R}^{n+|\Sigma|} \to [\![\tau]\!]$$
$$(\boldsymbol{\theta}, \mathbf{s} +\!\!+ \mathbf{s}') \mapsto [\![M]\!]((\boldsymbol{\theta}, \xi_1(\boldsymbol{\theta}, \mathbf{s}), \ldots, \xi_\ell(\boldsymbol{\theta}, \mathbf{s})), \mathbf{s}')$$

值得注意的是，与更标准的基本引理相比，这里我们需要捕捉自由变量对若干个额外样本 n的依赖。例如，在 $(\lambda x.\, x)$**采样** $_{\mathcal{N}}$的上下文中，子项 x 依赖于一个样本，尽管单独考察 x 时，这一点并不明显。

引理 5 通过结构归纳法证明（详情参见附录 C.2）。最有趣的情形包括：参数、基本操作和条件表达式。对于参数情形，我们利用 $\boldsymbol{\Theta}$ 的紧致性（假设 1）。对于基本操作，注意到根据引理 3，每个 $\mathcal{P}^{(n)}_{\iota}$在取负、加法和乘法下都封闭。最后，对于条件表达式，我们利用引理 3。

类型可靠性 II：随机梯度下降的正确性。接下来，我们讨论平滑问题的可积性，以及 (SGD1) 至 (SGD3)。我们将证明，不仅 $[\![M]\!]_\eta$，其直至二阶的偏导数也都一致受控于具有有限矩的函数。为使这一结论有可能成立，我们要求：

假设 2 对于每个 $\eta > 0$,

$$\sup_{x \in \mathbb{R}} |\sigma_\eta(x)| < \infty \qquad \sup_{x \in \mathbb{R}} |\sigma'_\eta(x)| < \infty \qquad \sup_{x \in \mathbb{R}} |\sigma''_\eta(x)| < \infty$$

例如，逻辑斯蒂 sigmoid 函数满足假设 2。

随后，我们可以利用$\mathbf{VectFr}$中的逻辑谓词，证明一个类似于引理 5 的基本引理，作相应的修改。如果 $f \in \mathcal{Q}^{(n)}_{\iota}$的直至二阶的偏导数一致受控于一个具有有限矩的函数，我们就称其满足相应条件。除引理 3 外，我们还利用微分的标准规则（例如求和法则、乘积法则和链式法则），以及假设 2。我们得出：

命题 4. 如果 $\boldsymbol{\theta} \mid \Sigma \vdash_{\mathrm{poly}} M :\ R$，那么 $[\![M]\!]_\eta$ 的二阶及以下偏导数一致受控于一个具有所有有限矩的函数。

因此，平滑优化问题 2 不仅定义良好，而且根据控制收敛定理 [21, 定理 6。28], 重参数化梯度估计器是无偏的。此外，(SGD1) 至 (SGD3) 均得到满足，且随机梯度下降正确。

[5] 对于 $\iota = R$

讨论。 类型系统 $\vdash_{\text{poly}}$ 简单明了，却能保证随机梯度下降的正确性。不过，它在某种程度上有所限制；尤其是，许多变分推断中出现的证据下界无法直接用它表示，因为这些证据下界通常具有指数项的对数形式（参见例 2）。

4.3 带注释的通用类型系统

接下来，我们介绍一种带注释的通用类型系统。在第 4.4 节中，我们将给出一种实例化，使 $\vdash_{\text{poly}}$ 更易于许可；在第 5 节中，我们将转而研究另一种性质：平滑化的一致收敛。

类型判断的形式为 $\Gamma \mid \Sigma \vdash_? M : \tau$，其中“?”表示我们旨在确立的性质，并且我们为基类型添加注释。因此，类型由以下形式生成：

$$
\begin{array}{ll}
\textbf{trace types} & \Sigma ::= [s_1 \sim \mathcal{D}_1, \ldots, s_n \sim \mathcal{D}_n] \\
\textbf{base types} & \iota ::= R \mid R_{>0} \\
\textbf{safe types} & \sigma ::= \iota^{\beta} \mid \sigma \bullet [] \to \sigma \\
\textbf{types} & \tau ::= \iota^{\alpha} \mid \tau \bullet \Sigma \to \tau
\end{array}
$$

注释取自一个集合，并且对于安全类型，可能会受到限制。其次，轨迹类型现在带有变量注释，通常为 $\Sigma = [s_1 \sim \mathcal{D}_1, \ldots, s_n \sim \mathcal{D}_n]$，其中变量 s_j 两两不同。

对于子类型关系，我们可以在基类型层面约束注释（见图 8a）；向更高类型的扩展则和之前一样完成。

类型规则的形式相同，但扩展了基类型上的注释，以及可能对这些注释加以约束的附加条件。例如，加法、求幂和采样的规则修改如下：

$$
\frac{}{\mid [] \vdash_? \underline{+} : \iota^{\alpha_1} \to \iota^{\alpha_2} \to \iota^{\alpha}} \text{ (cond. Add)} \qquad \frac{}{\mid [] \vdash_? \underline{\exp} : R^{\alpha} \to R_{>0}^{\alpha'}} \text{ (cond. Exp)}
$$

$$
\frac{}{\mid [s_j \sim \mathcal{D}] \vdash_? \mathbf{sample}_{\mathcal{D}} : R^{\alpha}} \text{ (cond. Sample)}
$$

子类型关系、变量、抽象和应用的规则完全无需更改，但它们使用的是带注释的类型，而非章节 2.2 中的类型。

$$
\frac{\Gamma \mid \Sigma \vdash_? M : \tau}{\Gamma' \mid \Sigma \vdash_? M : \tau'} \; \Gamma \sqsubseteq_? \Gamma', \tau \sqsubseteq_? \tau' \qquad \frac{}{x : \tau \mid [] \vdash_? x : \tau}
$$

$$
\frac{\Gamma, y : \tau_1 \mid \Sigma \vdash_? M : \tau_2}{\Gamma \mid [] \vdash_? \lambda y.\, M : \tau_1 \bullet \Sigma \to \tau_2} \qquad \frac{\Gamma \mid \Sigma_2 \vdash_? M : \tau_1 \bullet \Sigma_3 \to \tau_2 \quad \Gamma \mid \Sigma_1 \vdash_? N : \tau_1}{\Gamma \mid \Sigma_1 +\!\!+ \Sigma_2 +\!\!+ \Sigma_3 \vdash_? M\, N : \tau_2}
$$

完整的类型系统见附录 C.3。

$\vdash_{\text{poly}}$ 可视为 $\vdash_?$ 的一个特例：我们将单元素集合 $*$用作注释，为不需要的原语 $^{-1}$、exp 和 log 指定矛盾的侧条件（例如 false），并像上文一样，为 sample 指定侧条件“$\mathcal{D}$ 具有有限矩”。

表1：本文类型系统概述。

性质	章节	判断标注	
全性	第2.2节	$\vdash$	–
正确性 随机梯度下降	第4.2节	$\vdash_{\mathrm{poly}}$	无/$*$
	第 4.4 节	$\vdash_{\mathrm{SGD}}$	$0/1$
一致收敛 第5.1节		$\vdash_{\mathrm{unif}}$	(f Δ)/(t Δ)

表1概述了本文中的类型系统及其用途。$\vdash_?$ 及其具体实例细化了第2.2节中的基本类型系统；也就是说，如果某个上下文中的项可在带注释类型系统中证明，那么擦除它（即擦除基本类型和分布的注释）后，该项可在基本类型系统中证明。这一点很容易验证。

4.4 一个更宽松的类型系统

本节讨论通用类型系统的另一种实例化形式 $\vdash_{\mathrm{SGD}}$，以保证 (SGD0) 至 (SGD3)；它比 $\vdash_{\mathrm{poly}}$ 更宽松。具体而言，我们希望支持例 2，其中使用了对数以及涉及指数函数的密度。直观地说，我们需要确保涉及 exp 的子项会被相应的 log“中和”。为此，我们用 0 或 1 标注基本类型，并按离散顺序排列。0 是安全基本类型唯一的标注，可将其视为“可积”；1 表示“需要经过 $\overline{\log}$ 处理”。更准确地说，我们对类型规则施加限制，使得如果 $\boldsymbol{\theta} \mid \Sigma \vdash_{\mathrm{SGD}} M\colon \iota^{(e)}$，那么 [6] $\log^e \circ [\![M]\!]$，且 $\log^e \circ [\![M]\!]_\eta$ 的二阶以内偏导数均被某个具有有限矩的函数一致受控。

我们按如下方式定义基本类型的子类型关系：$\iota_1^{(e_1)} \sqsubseteq_{\mathrm{SGD}} \iota_2^{(e_2)}$ 若 $\iota_1 \sqsubseteq \iota_2$（定义见图3a）且 $e_1 = e_2$，或者 $\iota_1 = R_{>0} \overset{1}{=} \iota_2$ 且 $e_1 \leq e_2$。第二个析取条件可能出人意料，但我们确保类型为 $R_{>0}^{(0)}$ 的项完全不依赖采样。

图4列出了最重要的规则；完整的类型系统见附录 C.4。exp 和 log 分别使注释增大和减小。基本操作和条件表达式的规则源于引理 3 的闭包性质，以及以下初等事实：对于 $f, g\colon \boldsymbol{\Theta} \times \mathbb{R}^n \to \mathbb{R}$，$\log \circ (f \cdot g) = (\log \circ f) + (\log \circ g)$ 且 $\log \circ (f^{-1}) = -\log \circ f$。

Example 4. $\theta : R_{>0}^{(0)} \mid [\mathcal{N}, \mathcal{N}] \vdash_{\mathrm{SGD}} \underline{\log}\,(\theta^{-1} \cdot \underline{\exp}\,(\mathbf{sample}_{\mathcal{N}})) + \mathbf{sample}_{\mathcal{N}} : R^{(0)}$

请注意，条件表达式的分支必须具有安全类型，因此类型为 $R^{(1)}$ 的分支不符合要求。这是因为在条件表达式的平滑解释中，对数与加法组合时表现并不理想。

[6] 采用 $\log^0$ 为恒等映射的约定

$$\frac{}{\mid [] \vdash_{\mathrm{SGD}} \underline{\exp} : R^{(0)} \to R^{(1)}_{>0}} \qquad \frac{}{\mid [] \vdash_{\mathrm{SGD}} \underline{\log} : R^{(e)}_{>0} \to R^{(0)}}$$

$$\frac{}{\mid [] \vdash_{\mathrm{SGD}} \underline{+} : \iota^{(0)} \to \iota^{(0)} \to \iota^{(0)}} \qquad \frac{}{\mid [] \vdash_{\mathrm{SGD}} \underline{\cdot} : \iota^{(e)} \to \iota^{(e)} \to \iota^{(e)}}$$

$$\frac{}{\mid [] \vdash_{\mathrm{SGD}} \underline{-} : R^{(0)} \to R^{(0)}} \qquad \frac{}{\mid [] \vdash_{\mathrm{SGD}} \underline{^{-1}} : R^{(e)}_{>0} \to R^{(e)}_{>0}}$$

$$\frac{\Gamma \mid \Sigma \vdash_{\mathrm{SGD}} L : \iota^{(0)} \quad \Gamma \mid \Sigma' \vdash_{\mathrm{SGD}} M : \sigma \quad \Gamma \mid \Sigma'' \vdash_{\mathrm{SGD}} N : \sigma}{\Gamma \mid \Sigma +\!\!+ \Sigma' +\!\!+ \Sigma'' \vdash_{\mathrm{SGD}} \mathbf{if}\, L < 0\, \mathbf{then}\, M\, \mathbf{else}\, N : \sigma}$$

$$\frac{\mathcal{D} \text{ has finite moments}}{\mid [s_j \sim \mathcal{D}] \vdash_{\mathrm{SGD}} \mathbf{sample}_{\mathcal{D}} : R^{(0)}}$$

图4：随机梯度下降正确性相关类型规则的摘录（参见附录 C.4）。

此外请注意，在对数和逆函数的规则中允许使用 $e = 0$，这可能令人意外[7]。例如，这对于变分推断示例 2 的类型判定是必要的：

示例 *5*（变分推断的类型判定）。成立 $\mid [] \vdash N : \quad R^{(0)} \to R^{(0)} \to R^{(0)}_{>0} \to R^{(1)}_{>0}$ 和 $\theta : \quad R^{(0)} \mid [s_1 \sim \mathcal{N}] \vdash M : \quad R^{(0)}$。

类型可靠性。为了正式确立类型可靠性，我们可以使用一个逻辑谓词，它与第 4.2节中的谓词非常相似（请注意新增的第 2 项）：具体而言，$f \in \mathcal{Q}^{(n)}_{\iota^{(e)}}$ 如果

1. $\log^e \circ f$ 的二阶以内偏导数均一致受控于一个具有有限矩的函数 2. 如果 $\iota^{(e)}$ 是 $R^{(0)}_{>0}$，那么 f 受一个正的常值函数控制

利用这一点以及针对 $[\![(\text{–})]\!]$ 的类似逻辑谓词，我们可以证明：

命题 5。如果 $\theta_1 : \quad \iota^{(0)}, \ldots, \theta_m : \quad \iota^{(0)}_m \mid \Sigma \vdash_{\mathrm{SGD}} M : \quad \iota^{(0)}$ 那么

1. Σ 中的所有分布都具有有限矩；*2.* $[\![M]\!]$，并且对于每个 $\eta > 0$，$[\![M]\!]_\eta$ 的二阶及以下偏导数均被一个具有有限矩的函数一致受控。

因此，平滑优化问题 2 再次不仅定义良好，而且根据控制收敛定理，重参数化梯度估计器是无偏的。此外，(SGD1) 至 (SGD3) 均成立，且随机梯度下降是正确的。

5 一致收敛

在上一章节中，我们已经证明，带有重参数化梯度的随机梯度下降可以正确地（就命题 3 的意义而言）求解

[7] 请记住，类型为 $R^{(0)}_{>0}$ 的项不能依赖于样本。

对于任意固定的精度系数，平滑优化问题 2。不过，先验地，尚不清楚平滑问题 2 的解如何有助于求解原始问题 1。

下面说明可能出现的显著差异：

例 *6.* 考虑 $M \equiv$ **如果** $0 < 0$ **那么** $\theta \cdot \theta + 1$ **否则** $(\theta - 1) \cdot (\theta - 1)$。注意，无论 $\eta > 0$ 如何，$[\![M]\!]_\eta$ 的全局最小值和唯一驻点都在 $\theta = \frac{1}{2}$，其中 $[\![M]\!]_\eta(\frac{1}{2}) = \frac{3}{4}$。另一方面，$[\![M]\!](\frac{1}{2}) = \frac{1}{4}$，且 $[\![M]\!]$ 的全局最小值在 $\theta = 1$。

在本章节中，我们将探讨在什么条件下，平滑目标函数会在 $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ 上一致收敛到原始目标函数：

$$\text{(Unif)}\ \ \mathbb{E}_{\mathbf{s}\sim\mathcal{D}}\left[[\![M]\!]_\eta(\boldsymbol{\theta}, \mathbf{s})\right] \xrightarrow{\text{unif.}} \mathbb{E}_{\mathbf{s}\sim\mathcal{D}}\left[[\![M]\!](\boldsymbol{\theta}, \mathbf{s})\right] \text{ as } \eta \searrow 0 \text{ for } \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

我们设计了一种能够保证这一点的类型系统。

一致收敛的实际意义在于，在运行随机梯度下降之前，对于每个误差容限 $\epsilon > 0$，我们都能找到一个精度系数 $\eta > 0$，使平滑后的目标函数与原始目标函数之间的差异不超过 ϵ，尤其是对于随机梯度下降在 η-平滑问题上得到的 $\boldsymbol{\theta}^*$。

限制条件讨论。为排除例 6 中的病态情况，我们要求守卫几乎处处不为 0。

此外，根据一致极限定理 [29]，，只有当期望 $\mathbb{E}_{\mathbf{s}\sim\mathcal{D}}\left[[\![M]\!](\boldsymbol{\theta}, \mathbf{s})\right]$（作为参数 $\boldsymbol{\theta}$ 的函数）连续时，一致收敛（Unif）才有可能成立。一个简单的反例是取 $M \equiv$ **若** $\theta < 0$ **则** 0 **否则** 1；我们有 $\mathbb{E}_{\mathbf{s}}[[\![M]\!](\theta)] = [\theta \geq 0]$，它在 $\theta = 0$ 处不连续，更不用说可微了。我们的方法要求守卫不直接依赖参数，但允许它们通过微分同胚[8] 重参数化变换间接依赖参数；参见例8。我们称这类守卫为安全的。

总之，我们的目标是直观地确保守卫由随机样本（可能取决于参数）的微分同胚变换与一个几乎处处不为零的函数复合而成。

5.1 守卫安全性的类型系统

为了强制满足这一要求并使变换更加明确，我们引入语法糖 **变换样本** $_{\mathcal{D}}$ **通过** T，用于表示形式为 T **样本** $_{\mathcal{D}}$ 的应用。

例*7.* 如公式（2）所示，我们可以通过**变换采样** $_{\mathcal{N}}$ **by** $(\lambda s.\, s \cdot \sigma + \mu)$，从 $\mathcal{N}(\mu, \sigma^2)$ 中获得样本；这是一种项 $(\lambda s.\, s \cdot \sigma + \mu)$ **采样** $_{\mathcal{N}}$ 的语法糖。

[8] 附录D中的例12说明了为什么仅将重新参数化变换限制为不足够的双射（我们要求它是微分同胚）。

我们提出第 4.3 节通用类型系统的另一种实例，$\vdash_{\text{unif}}$，其中我们用 $\alpha = (g, \Delta)$ 标注基本类型，其中 $g \in \{\mathbf{f}, \mathbf{t}\}$ 表示我们是否要证明守卫安全性，而 Δ 是一个有限集合，用于捕获可能存在的对样本的依赖。我们按如下方式对基本类型定义子类型关系：$\iota_1^{(g_1,\Delta_1)} \sqsubseteq_{\text{unif}} \iota_2^{(g_2,\Delta_2)}$ 若 $\iota_1 \sqsubseteq \iota_2$（定义见图3a）、$\Delta_1 \subseteq \Delta_2$ 和 $g_1 \preceq g_2$ 成立，其中 $\mathbf{t} \preceq \mathbf{f}$。其依据的直觉是，我们总可以舍弃[9] 守卫安全性，并添加更多依赖。

条件表达式的规则确保只使用安全守卫。一元运算会保留变量依赖和守卫安全性。参数和常量不具备守卫安全性，也不依赖任何样本（完整类型系统见附录D）：

$$\frac{\Gamma \mid \Sigma \vdash_{\text{unif}} L : \iota^{(\mathbf{t},\Delta)} \quad \Gamma \mid \Sigma' \vdash_{\text{unif}} M : \sigma \quad \Gamma \mid \Sigma'' \vdash_{\text{unif}} N : \sigma}{\Gamma \mid \Sigma +\!\!+ \Sigma' +\!\!+ \Sigma'' \vdash_{\text{unif}} \mathbf{if}\, L < 0\, \mathbf{then}\, M\, \mathbf{else}\, N : \sigma}$$

$$\frac{}{\mid [] \vdash_{\text{unif}} \_ : R^{(g,\Delta)} \to R^{(g,\Delta)}}$$

$$\frac{}{\theta_i : \iota^{(\mathbf{f},\emptyset)} \mid [] \vdash_{\text{unif}} \theta_i : \iota^{(\mathbf{f},\emptyset)}} \qquad \frac{}{\mid [] \vdash_{\text{unif}} \underline{r} : \iota^{(\mathbf{f},\emptyset)}}\ r \in [\![\iota]\!]$$

$$\frac{\boldsymbol{\theta} \mid [] \vdash_{\text{unif}} T : R^{\alpha} \to R^{\alpha}}{\boldsymbol{\theta} \mid [s_j \sim \mathcal{D}] \vdash_{\text{unif}} \mathbf{transform\ sample}_{\mathcal{D}}\ \mathbf{by}\ T : R^{(\mathbf{t},\{s_j\})}}\ T \text{ diffeomorphic}$$

项 $\boldsymbol{\theta} \mid [] \vdash_{\text{unif}} T$：$R^{\alpha} \to R^{\alpha}$ 是微分同胚的，当且仅当对于每个 $\boldsymbol{\theta} \in \boldsymbol{\Theta}$，$[\![T]\!](\boldsymbol{\theta}, []) = [\![T]\!]_{\eta}(\boldsymbol{\theta}, [])$: $\mathbb{R} \to \mathbb{R}$ 都是微分同胚，即可微且为双射，并且其逆也可微。

首先，我们可以表示仿射变换，特别是例7中的位置-尺度变换：

例8（位置-尺度变换）。上下文中的项

$$\sigma : R_{>0}^{(\mathbf{f},\emptyset)}, \mu : R^{(\mathbf{f},\emptyset)} \mid [] \vdash \lambda s.\, \sigma \cdot s + \mu : R^{(\mathbf{f},\{s_1\})} \to R^{(\mathbf{f},\{s_1\})}$$

是微分同胚的。（然而，对于 $\sigma : R^{(\mathbf{f},\emptyset)}$，它不是，因为它包含 $\sigma = 0$。）因此，重参数化变换

$$G \equiv \sigma : R_{>0}^{(\mathbf{f},\emptyset)}, \mu : R^{(\mathbf{f},\emptyset)} \mid [s_1 : \mathcal{D}] \vdash \mathbf{transform\ sample}_{\mathcal{D}}\ \mathbf{by}\ (\lambda s.s{\cdot}\sigma{+}\mu) : R^{(\mathbf{t},\{s_1\})}$$

带有 g 标记 $\mathbf{t}$ 的项是可接受的守卫项。请注意，G 间接通过微分同胚依赖参数 σ 和 μ，这符合类型系统的规定。

若要证明二元运算满足守卫安全性，我们要求操作数之间不存在对样本的共享依赖：

$$\frac{}{\mid [] \vdash_{\text{unif}} \circ : \iota^{(\mathbf{f},\Delta)} \to \iota^{(\mathbf{f},\Delta)} \to \iota^{(\mathbf{f},\Delta)}}\ \circ \in \{+, \cdot\}$$

$$\frac{}{\mid [] \vdash_{\text{unif}} \circ : \iota^{(\mathbf{t},\Delta_1)} \to \iota^{(\mathbf{t},\Delta_2)} \to \iota^{(\mathbf{t},\Delta_1 \cup \Delta_2)}}\ \circ \in \{+, \cdot\}, \Delta_1 \cap \Delta_2 = \emptyset$$

旨在解决以下问题：

[9] 只要不在守卫中使用

例 *9*（非恒定守卫）。我们有 $\mid [] \vdash (\lambda x.x + (-x)) \colon R^{(\mathbf{f},\{s_1\})} \to_{R^{(\mathbf{f},\{s_1\})}}$，注意，对于 $+$ 规则，我们必须使用 $g = \mathbf{f}$；并且由于 $R^{(\mathbf{t},\{s_j\})} \sqsubseteq_{\text{unif}} R^{(\mathbf{f},\{s_j\})}$，我们有

$$\mid [] \vdash (\lambda x.x + (\underline{-}x)) : R^{(\mathbf{t},\{s_1\})} \to R^{(\mathbf{f},\{s_1\})}.$$

现在，**转换采样** $_{\mathcal{D}}$，**使得** $(\lambda y.y)$ 的类型为 $R^{(\mathbf{t},\{s_1\})}$，且 g 标志必然设为$\mathbf{t}$；因此该项

$$M \equiv \big(\lambda x.x + (-x)\big)\, \mathbf{transform\ sample}_{\mathcal{D}}\, \mathbf{by}\, (\lambda y.y)$$

它表示 0，类型为 $R^{(\mathbf{f},\{s_1\})}$，但不是 $R^{(\mathbf{t},\{s_1\})}$。由此可知，M 不能用于守卫中（注意条件语句规则中的附带条件），这正合所需：回想一下例 6。同样地，考虑项

$$N \equiv \big(\lambda x.(\lambda y\, z.\mathbf{if}\ y + (-z) < 0\, \mathbf{then}\, M_1\, \mathbf{else}\, M_2)\, x\, x\big) \\ (\mathbf{transform\ sample}_{\mathcal{D}}\, \mathbf{by}\, (\lambda y.y)) \tag{7}$$

求值后，守卫中的项 $y + (-z)$ 的指称值为 0。基于上述相同的原因，项 N 无法进行细化类型判定。

然而，该类型系统并不完备，也就是说，存在一些上下文中的项满足性质（一致收敛），但无法赋予类型。

例 *10*（不完备性）。以下上下文中的项表示“恒等”：

$$\mid [] \vdash (\lambda x.(\underline{2} \cdot x) + (-x)) : R^{(\mathbf{t},\{s_1\})} \to R^{(\mathbf{f},\{s_1\})}$$

但它没有类型 $R^{(\mathbf{t},\{s_1\})} \to R^{(\mathbf{t},\{s_1\})}$。接下来，采用与例 9 相同的推理，项

$$G \equiv (\lambda x.(\underline{2} \cdot x) + (-x))\, (\mathbf{transform\ sample}_{\mathcal{D}}\, \mathbf{by}\, (\lambda y.y))$$

具有类型 $R^{(\mathbf{f},\{s_1\})}$，但没有 $R^{(\mathbf{t},\{s_1\})}$，因此**如果** $G < 0$ **那么0否则1** 不可定型，尽管 G 可以安全地用于守卫中。

5.2 类型可靠性

今后，我们固定参数 $\theta_1 \colon \iota_1^{(\mathbf{f},\emptyset)}, \ldots, \theta_m \colon \iota_m^{(\mathbf{f},\emptyset)}$。

现在，我们讨论如何证明性质（一致收敛），即对于 $\boldsymbol{\theta} \mid \Sigma \vdash_{\text{unif}} M : \iota^{(g,\Delta)}$，$\eta$-平滑后的 $\mathbb{E}[[\![M]\!]_\eta(\boldsymbol{\theta},\ \mathbf{s})]$ 对 $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ 一致收敛，当 $\eta \searrow 0$ 时。显然，为使这一点成立，我们需要要求 σ_η 具有良好的（一致）收敛性质（就 0 处不可避免的不连续性所允许的范围而言）：

假设 3 对于每个 $\delta > 0$, σ_η 一致收敛。$\longrightarrow [(-) > 0]$ 在 $(-\infty, -\delta) \cup (\delta, \infty)$。

请注意，一般而言，即使 M 可定型，$[\![M]\!]_\eta$ 也不会在 $\boldsymbol{\theta}$ 和$\mathbf{s}$ 上都一致收敛，因为 $[\![M]\!]$ 在$\mathbf{s}$上仍可能不连续:

22 BasimKhajwal、C.-H. Luke Ong 和 Dominik Wagner(✉)

例*11.* 对于 $M \equiv$ **如果（对采样结果进行变换$_{\mathcal{N}}$，变换方式为**（$\lambda s.\, s{+}\theta$）） < 0 **那么**0 _ **否则**1，$[\![M]\!]$（θ， s） $= [s + \theta \geq 0]$，它是不连续的，而 $[\![M]\!]_\eta$（θ， s）$= \sigma_\eta$（$s + \theta$）。

然而，如果 $\boldsymbol{\theta} \mid \Sigma \vdash M$： $\iota^{(g,\Delta)}$那么 $[\![M]\!]_\eta$会 一致几乎一致地收敛到 $[\![M]\!]$，即在 $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ 上一致收敛，并在 $\mathbf{s} \in \mathbb{R}^n$上几乎一致收敛。形式化地，我们定义：

定义4. 设 f, f_η： $\boldsymbol{\Theta} \times \mathbb{R}^n \to \mathbb{R}$, μ 是 $\mathbb{R}^n$上的一个测度。若存在序列 $(\delta_k)_{k\in\mathbb{N}}$、$(\epsilon_k)_{k\in\mathbb{N}}$ 和 $(\eta_k)_{k\in\mathbb{N}}$，使得 $\lim_{k\to\infty} \delta_k = 0 = \lim_{k\to\infty} \epsilon_k$；并且对每个 $k \in \mathbb{N}$ 和 $\boldsymbol{\theta} \in \boldsymbol{\Theta}$，都存在 $U \subseteq \mathbb{R}^n$，使得 f_η一致几乎一致地收敛 到 f（记作： $f_\eta \xrightarrow{u.a.u.} f$），则称之为

1. $\mu(U) < \delta_k$ *and*
2. *for every* $0 < \eta < \eta_k$ *and* $\mathbf{s} \in \mathbb{R}^n \setminus U$, $|f_\eta(\boldsymbol{\theta}, \mathbf{s}) - f(\boldsymbol{\theta}, \mathbf{s})| < \epsilon_k$.

若 f, f_η与 $\boldsymbol{\theta}$相互独立，这一概念便与标准的几乎一致收敛相同。对于示例 M 中的 11 $[\![M]\!]_\eta \xrightarrow{\text{u.a.u.}} [\![M]\!]$成立，尽管一致收敛不成立。

然而，一致几乎一致收敛意味着期望的一致收敛：

引理 6. 设 f, f_η： $\boldsymbol{\Theta} \times \mathbb{R}^n \to \mathbb{R}$ 具有有限矩。

If $f_\eta \xrightarrow{u.a.u.} f$ *then* $\mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[f_\eta(\boldsymbol{\theta}, \mathbf{s})] \xrightarrow{unif.} \mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[f(\boldsymbol{\theta}, \mathbf{s})]$.

因此，只需证明 $[\![M]\!]_\eta \xrightarrow{\text{u.a.u.}} [\![M]\!]$。为此，我们在 **VectFr** 中的态射序列（对应于平滑化）与 **QBS** 中的态射（对应于可测标准语义）之间建立无穷逻辑关系。随后，我们证明基本引理17（详见附录D）。不出所料，条件表达式的情形最为有趣。这里运用了假设 3，并利用了这样一点：对于类型规则断言其守卫安全标志为$\mathbf{t}$的守卫，其值仅可能在零测集上为 0。由此得出：

定理1. 如果 θ_1： $\iota_1^{(\mathbf{f},\emptyset)}, \ldots, \theta_m$： $\iota_m^{(\mathbf{f},\emptyset)} \mid \Sigma \vdash_{\text{unif}} M$： $R^{(g,\Delta)}$ 那么 $[\![M]\!]_\eta$ 几乎处处 $\longrightarrow [\![M]\!]$。特别是，如果 $[\![M]\!]_\eta$和 $[\![M]\!]$ 也具有有限矩，那么

$$\mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[[\![M]\!]_\eta(\boldsymbol{\theta}, \mathbf{s})] \xrightarrow{unif.} \mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[[\![M]\!](\boldsymbol{\theta}, \mathbf{s})] \qquad as\ \eta \searrow 0\ for\ \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

最后我们指出，通过为 a-重（针对 $a \in \mathbb{N}_{>0}$）加法 $a \cdot M \equiv M + \cdots + M$和乘法 $M^a \equiv M \cdot \cdots \cdot M$添加语法糖，可以放宽对 $\vdash_{\text{unif}}$ 的限制。这会允许更多项作为守卫，但仍然安全（见图 10）。

6 相关工作

[23] 既是我们工作的出发点，也是最自然的比较对象。他们通过额外且非平凡的边界 项，修正了针对不可微模型的（有偏）重参数化梯度估计器。他们提出了

一种仅适用于仿射 守卫的高效方法。此外，他们并不关心基于梯度的优化过程是否收敛；也没有讨论其所作假设如何在编程语言中体现。

在重参数化梯度的背景下， [25]和 [18]以连续的方式放松离散随机变量，从而有效处理一类特定的不连续模型。 [41] 对不连续优化采用类似的平滑方法，但并未考虑完整的编程语言。

为了保证绝对连续性（这是例如变分推断正确性的必要但非充分条件）， [24] 采用了与我们的轨迹类型类似的方法来跟踪抽取的样本。它们不支持标准条件表达式，但其“变通方法”同样是急切的，因为它会合并两个分支的轨迹。此外，它们不支持完整的高阶语言，在这种语言中，高阶项可以抽取样本。因此，它们无需考虑在求值期间跟踪所抽取样本的函数 类型。

7 实证评估

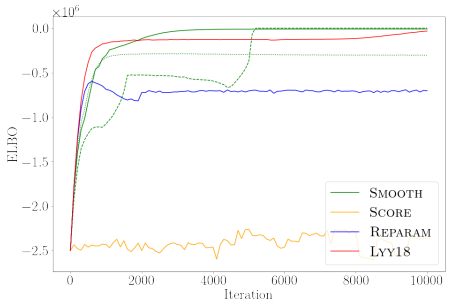
我们将平滑梯度估计器（Smooth）与有偏的重参数化估计器（Reparam）、 [23], 提出的无偏校正方法（LYY18），以及 [31,40,27]提出的无偏估计器（Score）进行比较。实验设置基于 [23]。实现采用 Python 编写，并使用自动微分（由 jax 库提供）为任意概率程序实现上述各估计器。对于每个估计器和模型，我们使用 Adam [19] 优化器进行 10,000 次迭代，学习率为 0.001；对于 xornet，则使用 0.01。每个模型的初始模型参数 $\boldsymbol{\theta}_0$ 在所有运行中均保持固定。每次迭代中，我们从梯度估计器中使用 $N = 16$个蒙特卡罗样本。对于 Lyy18 估计器，每次估计都使用一个边界项子样本。对于我们的平滑估计器，我们使用精度系数$\eta \in \{0.1, 0.15, 0.2\}$。更多细节见附录 E.1。

一阶程序的编译。 我们的所有基准测试都是一阶的。我们使用以下对象的相容闭包，将一个可能不连续的程序编译为一个平滑程序（由σ_η参数化）：

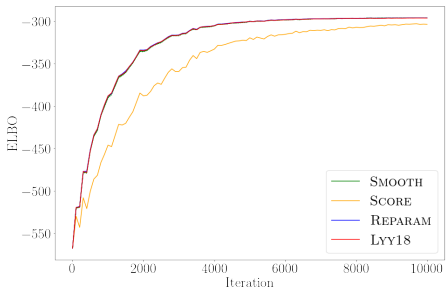
$$\textbf{if } L < 0 \textbf{ then } M \textbf{ else } N \rightsquigarrow (\lambda w.\, \sigma_\eta(-w) \cdot M + \sigma_\eta(w) \cdot N)\, L$$

请注意，规模只会线性增长；通过使用抽象而非复制守卫 L，我们避免了指数级膨胀。

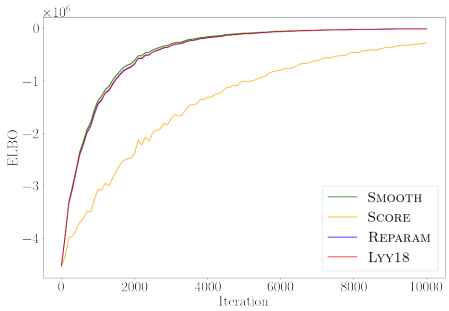
模型。 我们纳入了 [23],一个差分隐私示例[11] 中的模型，以及一个神经网络；对于该神经网络，我们的主要竞争方法—— [23],的估计器——并不 适用（更多详情请参见附录 E.2）。



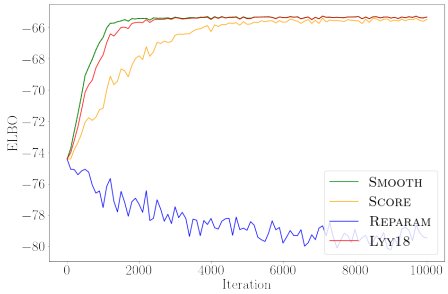
(a) 温度



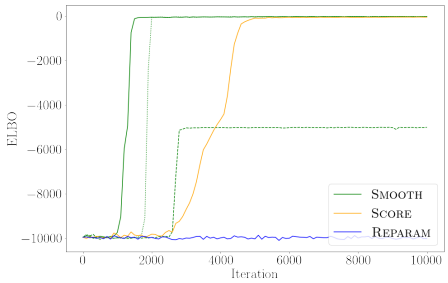
(b) textmsg



(c) 流感



(d) 作弊



(e) xornet

图5：各模型的ELBO轨迹。每种估计器使用一种颜色；Smooth 的精度系数 $\eta = 0.1$、0。15、0。2 分别以虚线、实线和点线表示。

结果分析

我们在图 5 中绘制了证据下界轨迹，并在附录 E.3 的表2中提供了计算成本和方差的数据。

图 5a 中温度模型的 ELBO图和图 5d 中作弊模型的 ELBO图显示，与平滑和 Lyy18 估计量相比，Reparam估计器存在偏差，并收敛到次优值。对于温度模型，我们还可以从图和表 Ta-

ble 2a 中的数据看到，Score估计量的方差极高，而且不会收敛。

最后，xornet模型展现了训练基于阶跃函数的神经网络的难度。由于存在非仿射条件分布，Lyy18估计器不适用于此处。在图 5e 中，Reparam估计器毫无进展，而其他估计器则成功收敛到接近 0 的证据下界，这表明它们学会了一个能够正确分类所有点的网络。具体而言，Smooth估计量收敛得最快。

总而言之，结果揭示了 Reparam 估计器存在偏差的情形，也表明 Smooth 估计量不受同样的限制。在 Lyy18 估计器有定义的情况下，它们收敛到大致相同的目标函数值；此外，平滑方法可推广至更复杂的模型，例如具有非线性边界的神经网络。我们提出的 Smooth 估计量始终具有显著更低的工作归一化方差，最多可低三个数量级。

8 结论与未来方向

我们讨论了一种简单的概率编程语言，用以形式化例如概率编程中的变分推断所产生的优化问题。我们为该语言赋予了指称（可测）值语义，并提供了对可能不连续程序的平滑近似，该近似由精度系数参数化。我们提出了类型系统，以确保在优化问题背景下具有良好的性质：对于固定的精度系数，即使使用重参数化梯度（它是无偏的），随机梯度下降也会收敛到驻点。此外，随着精度提高，平滑目标函数一致收敛到真实目标函数。

我们的类型系统可用于分别 检查这两项性质，从而即使其中一个系统存在不完备性，也能获得部分理论保证。我们还强调，随机梯度下降和平滑后的无偏梯度估计器甚至可应用于不可 进行类型检查的程序。

我们的原型实现所进行的实验证实了降低方差和无偏性的优势。与 [23],提出的重参数化梯度估计器的无偏校正相比，我们的估计器收敛表现相近，但更简单、速度更快，并将工作归一化方差降低了数个数量级（2 至 3,000 倍）。

未来方向。 未来研究的一个自然方向是提高语言和类型系统的完备性，也就是说，支持更多行为良好的程序，尤其是涉及递归的程序。

此外，精度系数的选择也为进一步研究留有空间。我们认为，不预先固定精度系数，而是在优化过程中逐步提高它，可能会很有成效在这一过程中；提高方式可以是采用预先确定的计划（取决于程序的结构属性），也可以是自适应调整。

参考文献

1. Aumann, R.J.：函数空间的博雷尔结构。《伊利诺伊数学杂志》**5** (1961) 2. Bertsekas, D.：凸优化算法。Athena Scientific (2015) 3. Bertsekas, D.P., Tsitsiklis, J.N.：存在误差时梯度方法的梯度收敛。SIAM J. Optim. **10**(3), 627–642 (2000) 4. Bingham, E., Chen, J.P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P.A., Horsfall, P., Goodman, N.D.：Pyro：深度通用概率编程。《机器学习研究杂志》**20**, 28:1–28:6 (2019) 5. Bishop, C.M.：模式识别与机器学习，第5版。信息科学与统计，Springer (2007) 6. Blei, D.M., Kucukelbir, A., McAuliffe, J. D.：变分推断：面向统计学家的综述。《美国统计学会杂志》**112**(518), 859–877 (2017 1285773 7 10 1080 01621459 2017). https://doi.org/ . / . . . Borgström, J., Lago, U.D., Gordon, A.D., Szymczak, M.：通用概率编程的λ演算基础。载于：第21届ACM SIGPLAN函数式编程国际会议论文集，ICFP 2016，日本奈良，2016年9月18日至22日。第33–46页 (2016) 8. Botev, Z., Ridder, A.：方差缩减。载于：Wiley StatsRef：在线统计学参考资料，第1–6页 (2017) 9. Cusumano-Towner, M.F., Saad, F.A., Lew, A.K., Mansinghka, V.K.：Gen：具有可编程推断功能的通用概率编程系统。载于：McKinley, K.S., Fisher, K. (编) 第40届ACM SIGPLAN编程语言设计与实现会议论文集, PLDI 2019，美国亚利桑那州菲尼克斯，2019年6月22日至26日。第221–236页。ACM (2019)。https://doi. org/10.1145/3314221.3314642, https://doi.org/10.1145/3314221.3314642 10. Dahlqvist, F., Kozen, D.：带条件化的高阶概率程序语义。载于 ACM 程序语言 **4**(POPL), 57:1–57:29 (2020) 11. Davidson-Pilon, C.：《黑客的贝叶斯方法：概率编程与贝叶斯推断》。Addison-Wesley Professional (2015) 12. Ehrhard, T., Tasson, C., Pagani, M.：概率相干空间对概率PCF完全抽象。载于：第41届ACM SIGPLAN-SIGACT编程语言原理年会, POPL '14，美国加利福尼亚州圣迭戈，2014年1月20日至21日。第309–320页 (2014) 13. Frölicher, A., Kriegl, A.：《线性空间与微分理论》。Interscience, J. Wiley and Son，纽约 (1988) 14. Glynn, P.W., Whitt, W.：模拟估计量的渐近效率。《运筹学》**40**(3), 505–520 (1992) 15. Heunen, C., Kammar, O., Staton, S., Yang, H.：一个适用于高阶概率论的范畴。计算机科学逻辑研讨会论文集 (2017) 16. Heunen, C., Kammar, O., Staton, S., Yang, H.：一个适用于高阶概率论的范畴。载于：第32届ACM/IEEE计算机科学逻辑研讨会，LICS 2017，冰岛雷克雅未克，2017年6月20日至23日。第1–12页 (2017) 17. Hur, C., Nori, A.V., Rajamani, S.K., Samuel, S.：概率程序的可证明正确的采样器。载于：第35届IARCS软件技术与理论计算机科学基础年会，FSTTCS 2015，2015年12月16日至18日，印度班加罗尔。第475–488页 (2015)

18. Jang, E., Gu, S., Poole, B.：采用 Gumbel-Softmax 的分类重参数化。载于：第5届国际学习表征会议，ICLR 2017，法国土伦，2017年4月24日至26日，会议论文集（2017） 19. Kingma, D.P., Ba, J.：Adam：一种随机优化方法。载于：Bengio, Y., LeCun, Y.（编）第3届国际学习表征会议，ICLR 2015，美国加利福尼亚州圣迭戈，2015年5月7日至9日，会议论文集（2015） 20. Kingma, D.P., Welling, M.：自编码变分贝叶斯。载于：Bengio, Y., LeCun, Y.（编）第2届国际学习表征会议，ICLR 2014，加拿大阿尔伯塔省班夫，2014年4月14日至16日，会议论文集（2014） 21. Klenke, A.：《概率论：综合课程》。Universitext，Springer London（2014） 22. Lee, W., Yu, H., Rival, X., Yang, H.：面向概率程序的经验证随机变分推断。PACMPL 4(POPL)（2020） 23. Lee, W., Yu, H., Yang, H.：用于不可微模型的重参数化梯度。载于：神经信息处理系统进展第31卷：神经信息处理系统2018年度会议，NeurIPS 2018，加拿大蒙特利尔，2018年12月3日至8日。第5558–5568页（2018） 24. Lew, A.K., Cusumano-Towner, M.F., Sherman, B., Carbin, M., Mansinghka, V.K.：概率编程语言中可靠可编程推断的轨迹类型与指称语义。ACM 程序语言论文集 4(POPL)，19:1–19:32（2020） 25. Maddison, C.J., Mnih, A., Teh, Y.W.：Concrete 分布：离散随机变量的连续松弛。载于：第5届国际学习表征会议，ICLR 2017，法国土伦，2017年4月24日至26日，会议论文集（2017） 26. Mak, C., Ong, C.L., Paquet, H., Wagner, D.：几乎必然终止的概率程序的密度几乎处处可微。载于：Yoshida, N.（编）《编程语言与系统——第30届欧洲编程研讨会》，ESOP 2021，作为欧洲软件理论与实践联合会议ETAPS 2021的一部分举行，卢森堡卢森堡市，2021年3月27日至4月1日，论文集。计算机科学讲义，第12648卷，第432–461页。Springer（2021） 27. Minh, A., Gregor, K.：信念网络中的神经变分推断与学习。载于：第31届国际机器学习会议论文集，ICML 2014，中国北京，2014年6月21日至26日。JMLR 研讨会及会议论文集，第32卷，第1791–1799页。JMLR.org（2014） 28. Mityagin, B.：实解析函数的零点集（2015） 29. Munkres, J.R.：《拓扑学》。Prentice Hall，印度新德里，第2版（1999） 30. Murphy, K.P.：《机器学习：概率论视角》。MIT出版社（2012） 31. Ranganath, R., Gerrish, S., Blei, D.M.：黑盒变分推断。载于：第十七届人工智能与统计学国际会议论文集，AISTATS 2014，冰岛雷克雅未克，2014年4月22日至25日。第814–822页（2014） 32. Rezende, D.J., Mohamed, S., Wierstra, D.：深度生成模型中的随机反向传播与近似推断。载于：第31届国际机器学习会议论文集，ICML 2014，中国北京，2014年6月21日至26日。JMLR 研讨会及会议论文集，第32卷，第1278–1286页。JMLR.org（2014） 33. Shumway, R.H., Stoffer, D.S.：《时间序列分析及其应用》。Springer 统计学丛书，Springer-Verlag（2005）

34. Soudjani, S.E.Z., Majumdar, R., Nagapetyan, T.：用于混合系统统计模型检验的多层蒙特卡洛方法。载于：Bertrand, N., Bortolussi, L.（编）《系统的定量评估——第14届国际会议》，QEST 2017，德国柏林，2017年9月5日至7日，论文集。计算机科学讲义，第10503卷，第351–367页。Springer（2017） 35. Stacey, A.：比较光滑学。《范畴理论与应用》**25**(4)，64–117（2011） 36. Staton, S.：概率编程的交换语义。载于：《编程语言与系统——第26届欧洲编程研讨会》，ESOP 2017，作为欧洲软件理论与实践联合会议ETAPS 2017的一部分举行，瑞典乌普萨拉，2017年4月22日至29日，论文集。第855–879页（2017） 37. Staton, S., Yang, H., Wood, F.D., Heunen, C., Kammar, O.：概率编程的语义：高阶函数、连续分布与软约束。载于：第31届ACM/IEEE计算机科学逻辑研讨会论文集，LICS '16，美国纽约，2016年7月5日至8日。第525–534页（2016） 38. Titsias, M.K., Lázaro-Gredilla, M.：用于非共轭推断的双随机变分贝叶斯。载于：第31届国际机器学习会议论文集，ICML 2014，中国北京，2014年6月21日至26日。第1971–1979页（2014） 39. Vákár, M., Kammar, O., Staton, S.：统计概率编程的域理论。PACMPL **3**(POPL)，36:1–36:29（2019） 40. Wingate, D., Weber, T.：概率编程中的自动变分推断。CoRR **abs/1301.1299**（2013） 41. Zang, I.：通过平滑实现不连续优化。《运筹数学》**6**(1)，140–152（1981） 42. Zhang, C., Butepage, J., Kjellstrom, H., Mandt, S.：变分推断进展。《IEEE模式分析与机器智能汇刊》**41**(8)，2008–2026（2019）

A 第 2 节的补充材料

A.1 第2.2节的补充材料

引理 1. 如果 $\Gamma \mid \Sigma \vdash M\colon \tau$ 且 $\Gamma \mid \Sigma' \vdash M\colon \tau'$，则 $\Sigma = \Sigma'$。

证明（概要）。我们定义类型上的等价关系 $\approx$：

1. $\iota \approx \iota'$
2. $(\tau_1 \bullet \Sigma \to \tau_2) \approx (\tau_1' \bullet \Sigma' \to \tau_2')$ iff $\tau_1 \approx \tau_1'$ implies $\Sigma = \Sigma'$ and $\tau_2 \approx \tau_2'$

直观来说，如果对于（按归纳定义）相关的参数，它们生成相同的样本，并且返回类型也相关，那么这两种类型就由 $\approx$ 关联。我们将该关系扩展到上下文：$\Gamma \approx \Gamma'$，如果对于所有 $x\colon \tau$（属于 Γ）以及 $x\colon \tau'$（属于 Γ'），都有 $\tau \approx \tau'$。

接着，我们通过归纳法证明：若 $\Gamma \mid \Sigma \vdash M\colon \tau, \Gamma' \mid \Sigma' \vdash M\colon \tau'$ 且 $\Gamma \approx \Gamma'$，则 $\Sigma = \Sigma'$ 且 $\tau \approx \tau'$。最后，这个加强后的陈述使我们能够证明引理中棘手的情形：应用。

A.2 第2.3节的补充材料

与可测空间 (X, Σ_X) 类似，拟博雷尔空间（QBS）是一个二元组 (X, M_X)，其中 X 是一个集合；但 QBS 并不公理化可测子集 Σ_X，而是公理化可容许随机元素 M_X。集合 M_X 是函数 $\mathbb{R} \to X$ 的一个集合，必须满足以下闭包性质：

– 若 $\alpha \in M_X$ 且 $f\colon \mathbb{R} \to \mathbb{R}$ 可测，则 $\alpha \circ f \in M_X$ – 若 $\alpha\colon \mathbb{R} \to X$ 为常数，则 $\alpha \in M_X$ – 给定实数的一个可数划分 $\mathbb{R} = \biguplus_{i\in\mathbb{N}} S_i$，其中每个 S_i 都是博雷尔集，且 $\{\alpha_i\}_{i\in\mathbb{N}} \subseteq M_X$，函数 $r \mapsto \alpha_i(r)$，其中 $r \in S_i$ 属于 M_X。

QBS 态射 $(X, M_X) \to (Y, M_Y)$ 是函数 $f\colon X \to Y$，满足 $f \circ \alpha \in M_Y$ 当且仅当 $\alpha \in M_X$。

引理 7（替换）。设 $\Gamma, x\colon \tau' \mid \Sigma \vdash M\colon \tau$ 且 $\Gamma \mid [] \vdash N\colon \tau'$。则 $[\![M]\!]\big((\gamma, [\![N]\!](\gamma, [])), \mathbf{s}\big) = [\![M[N/x]]\!](\gamma, \mathbf{s})$。

A.3 章节 2.4 的补充材料

以下内容可通过对 M 进行结构归纳来验证：

引理 8（替换）。若 $\Gamma, x\colon \tau' \mid \Sigma \vdash M\colon \tau$ 且 $\Gamma \mid [] \vdash N\colon \tau'$，则 $\Gamma \mid \Sigma \vdash M[N/x]\colon \tau$。

请注意，$\Gamma, x\colon \tau' \mid \Sigma \vdash M\colon \tau$ 和 $\Gamma \mid \Sigma' \vdash N\colon \tau'$ 并不一定意味着 $\Gamma \mid \Sigma +\!\!+ \Sigma' \vdash M[N/x]\colon \tau$。取 $M \equiv x + x$ 并对 $N \equiv$ **进行采样** $_{\mathcal{N}}$。然后注意到

$$x : R \mid [] \vdash M : R \qquad \mid [\mathcal{N}] \vdash N : R \qquad \mid [\mathcal{N}, \mathcal{N}] \vdash M[N/x].$$

$$\frac{}{V \Downarrow_1^{[]} V} \qquad \frac{}{\mathbf{sample}_{\mathcal{D}} \Downarrow_{\mathrm{pdf}_{\mathcal{D}}(s)}^{[s]} \underline{s}}$$

$$\frac{L \Downarrow_{w_1}^{\mathbf{s}_1} \underline{r} \quad M \Downarrow_{w_2}^{\mathbf{s}_2} V \quad N \Downarrow_{w_3}^{\mathbf{s}_3} V'}{\mathbf{if}\, L < 0\, \mathbf{then}\, M\, \mathbf{else}\, N \Downarrow_{w_1 \cdot w_2 \cdot w_3}^{\mathbf{s}_1 +\!\!+ \mathbf{s}_2 +\!\!+ \mathbf{s}_3} V} \; r < 0$$

$$\frac{L \Downarrow_{w_1}^{\mathbf{s}_1} \underline{r} \quad M \Downarrow_{w_2}^{\mathbf{s}_2} V \quad N \Downarrow_{w_3}^{\mathbf{s}_3} V'}{\mathbf{if}\, L < 0\, \mathbf{then}\, M\, \mathbf{else}\, N \Downarrow_{w_1 \cdot w_2 \cdot w_3}^{\mathbf{s}_1 +\!\!+ \mathbf{s}_2 +\!\!+ \mathbf{s}_3} V'} \; r \geq 0$$

$$\frac{M_1 \Downarrow_{w_1}^{\mathbf{s}_1} \underline{r_1} \quad M_2 \Downarrow_{w_2}^{\mathbf{s}_2} \underline{r_2}}{M_1 \underline{\circ} M_2 \Downarrow_{w_1 \cdot w_2}^{\mathbf{s}_1 +\!\!+ \mathbf{s}_2} \underline{r_1 \circ r_2}} \; \circ \in \{+, \cdot\}$$

$$\frac{M \Downarrow_w^{\mathbf{s}} \underline{r}}{\underline{\mathrm{op}}\, M \Downarrow_w^{\mathbf{s}} \underline{\mathrm{op}(r)}} \; \mathrm{op} \in \{-, ^{-1}, \exp, \log\}, r \in \mathrm{dom}(\mathrm{op})$$

$$\frac{M \Downarrow_{w_1}^{\mathbf{s}_1} \lambda x.\, M' \quad N \Downarrow_{w_2}^{\mathbf{s}_2} V' \quad M'[V'/x] \Downarrow_{w_3}^{\mathbf{s}_3} V}{M\, N \Downarrow_{w_1 \cdot w_2 \cdot w_3}^{\mathbf{s}_1 +\!\!+ \mathbf{s}_2 +\!\!+ \mathbf{s}_3} V}$$

图 6：基于采样的操作性大步语义

讨论。引理 8 是按值调用语言中通常替换引理的一个略强版本：若 $\Gamma, x : \tau' \mid \Sigma \vdash M : \tau$ 且 $\Gamma \mid \Sigma' \vdash V : \tau$，则 $\Gamma \mid \Sigma +\!\!+ \Sigma' \vdash M[V/x] : \tau$；注意，$\Sigma' = []$ 必然成立，而且我们还有 $\Gamma \mid \Sigma +\!\!+ \Sigma' \vdash (\lambda x.M)\ V : \tau$。因此，对于 CBV β-归约，主题归约成立。

B 第 3 章节的补充材料

注 *1.* 设 $\phi : X \to Y$ 是一个函数，且 $(X, \mathcal{C}_X, \mathcal{F}_X)$ 和 $(Y, \mathcal{C}_Y, \mathcal{C}_X)$ 是向量 Frölicher 空间，其中前者由 $\mathcal{C}^0 \subseteq$ **集合**$(\mathbb{R}, X)$ 生成。则 ϕ 是一个态射，当且仅当对所有 $f \in \mathcal{F}_Y$ 和 $c \in \mathcal{C}^0$，均有 $f \circ \phi \circ c \in C^\infty(\mathbb{R}, \mathbb{R})$（即无需检查 $c \in \mathcal{C}_X \setminus \mathcal{C}^0$）。

~(注意 $\mathcal{C} \subseteq \mathcal{C}_X \subseteq \mathcal{C}_{X\circ}$。因此，若 $f : X \to \mathbb{R}$ 满足对所有 $c \in \mathcal{C}_X \supseteq \mathcal{C}$，均有 $f \circ c \in C^\infty(\mathbb{R}, \mathbb{R})$，则 $f \in \mathcal{F}_{X\circ}$。)

命题 2。 VectFr 是笛卡尔闭的。

证明。 1. 单元素向量空间是终对象。
2. 假设 $(X_1, \mathcal{C}_{X_1}, \mathcal{F}_{X_1})$ 和 $(X_2, \mathcal{C}_{X_2}, \mathcal{F}_{X_2})$ 是向量 Frölicher 空间。考虑由 $\{\langle c_1, c_2\rangle \mid c_1 \in \mathcal{C}_{X_1}, c_2 \in \mathcal{C}_{X_2}\}$ 在 $X_1 \times X_2$ 上生成的向量 Frölicher 空间。根据构造，$(X_1 \times X_2, \mathcal{C}_{X_1 \times X_2}, \mathcal{F}_{X_1 \times X_2})$ 是一个向量 Frölicher 空间，并且 $\pi_i : (X_1 \times X_2, \mathcal{C}_{X_1 \times X_2}, \mathcal{F}_{X_1 \times X_2}) \to (X_i, \mathcal{C}_{X_i}, \mathcal{F}_{X_i})$ 是态射。现在，假设 Z 和 $f : Z \to X_1$ 以及 $g : Z \to X_2$ 是态射。显然，$h := \langle f, g\rangle$ 是唯一满足 $\pi_1 \circ h = f$ 和 $\pi_2 \circ h = g$ 的态射 $Z \to X_1 \times X_2$。

3. 最后，假设 $(X, \mathcal{C}_X, \mathcal{F}_X)$ 和 $(Y, \mathcal{C}_Y, \mathcal{F}_Y)$ 是向量 Frölicher 空间。考虑由 $\{c\colon \mathbb{R} \to \mathsf{VectFr}(X,Y) \mid (r,x) \mapsto c(r)(x)) \in \mathsf{Fr}(\mathbb{R}\times X, Y)\}$ 在同态集 $\mathsf{VectFr}(X,Y)$ 上生成的向量 Frölicher 空间。定义 $\mathrm{eval} : \mathsf{VectFr}(X,Y) \times X \to Y$，其中 $\mathrm{eval}(f, x) \coloneqq f(x)$。根据注 1，要证明这是一个态射，只需考虑 $c_1 : \mathbb{R} \to C_{X\Rightarrow Y}$，使得 $((r,x) \mapsto c_1(r)(x)) \in \mathsf{Fr}(\mathbb{R}\times X, Y)$，$c_2 \in \mathcal{C}_X$ 和 $g \in \mathcal{F}_Y$。注意

$$g \circ \mathrm{eval} \circ \langle c_1, c_2\rangle = g \circ \underbrace{((r,x) \mapsto c_1(r)(x))}_{\in \mathbf{Fr}(\mathbb{R}\times X, Y)} \circ \underbrace{\langle \mathrm{id}, c_2\rangle}_{\in \mathcal{C}_{\mathbb{R}\times X}}$$

根据态射的定义，它属于 $C^\infty(\mathbb{R}, \mathbb{R})$。显然，这满足所要求的泛性质。

C 第4节补充材料

C.1 第4.1节补充材料

以下结论可直接由一个关于交换微分与积分次序的著名结果得出，而该结果是支配收敛定理的推论 [21,定理6.28]:

引理9。 设 $U \subseteq \mathbb{R}$ 为开集。假设 $g\colon \mathbb{R}\times\mathbb{R}^n \to \mathbb{R}$ 满足

1. 对每个 $x \in \mathbb{R}$，$\mathbf{s} \mapsto g(x、\mathbf{s})$ 可积；*2.* g 处处连续可微；*3.* 存在可积的 $h : \mathbb{R}^n \to \mathbb{R}$，使得对于所有 $x \in U$ 和 $\mathbf{s} \in \mathbb{R}^n$，$|\frac{\partial g}{\partial x}(x、\mathbf{s})| \le h(\mathbf{s})$。

那么对于所有 $x \in U$，$\frac{\partial}{\partial x}\int g(x、\mathbf{s})\,\mathrm{d}\mathbf{s} = \int \frac{\partial g}{\partial x}(x、\mathbf{s})\,\mathrm{d}\mathbf{s}$。

推论1。 设 $i \in \{1,\dots,m\}$、$M > 0$ 和 $U \coloneqq B_M(\mathbf{0}) \subseteq \mathbb{R}^m$ 为开M球。假设 $g\colon \mathbb{R}^m\times\mathbb{R}^n \to \mathbb{R}$ 满足

1. 对每个 $\mathbf{x} \in \mathbb{R}^m$，$\mathbf{s} \mapsto g(\mathbf{x}、\mathbf{s})$ 可积；*2.* g 处处连续可微；*3.* 存在可积的 $h : \mathbb{R}^n \to \mathbb{R}$，使得对于所有 $\mathbf{x} \in U$ 和 $\mathbf{s} \in \mathbb{R}^n$，$|\frac{\partial g}{\partial x_i}(\mathbf{x}、\mathbf{s})| \le h(\mathbf{s})$。

那么，对于所有 $\mathbf{x} \in U$，$\frac{\partial}{\partial x_i}\int g(\mathbf{x},\mathbf{s})\,\mathrm{d}\mathbf{s} = \int \frac{\partial g}{\partial x_i}(\mathbf{x},\mathbf{s})\,\mathrm{d}\mathbf{s}$。

C.2 第4.2节的补充材料

引理 3. 如果 $f, g\colon \mathbb{R}^n \to \mathbb{R}$ 具有（全部）有限矩，那么 $-f, f+g, f\cdot g$ 也具有有限矩。

证明。 对于取负，结论显然。对于加法，可以如下验证:

$$\begin{aligned}\mathbb{E}[|(f+g)(\mathbf{s})|^p] &\le \mathbb{E}\left[|2f(\mathbf{s})|^p + |2g(\mathbf{s})|^p\right] \\ &\le 2^p\cdot\mathbb{E}\left[|f(\mathbf{s})|^p\right] + 2^p\cdot\mathbb{E}\left[|g(\mathbf{s})|^p\right] < \infty\end{aligned}$$

对于乘法，由柯西-施瓦茨不等式可得:

$$\mathbb{E}[|(f\cdot g)(\mathbf{s})|^p] = \mathbb{E}\left[|f(\mathbf{s})|^p\cdot|g(\mathbf{s})|^p\right] \le \sqrt{\mathbb{E}\left[|f(\mathbf{s})|^{2p}\right]\cdot\mathbb{E}\left[|g(\mathbf{s})|^{2p}\right]} < \infty$$

$$\frac{\Gamma \mid \Sigma \vdash_{\mathrm{poly}} M : \tau}{\Gamma' \mid \Sigma \vdash_{\mathrm{poly}} M : \tau'}\ \Gamma \sqsubseteq_{\mathrm{poly}} \Gamma', \tau \sqsubseteq_{\mathrm{poly}} \tau' \qquad \frac{}{x : \tau \mid [] \vdash_{\mathrm{poly}} x : \tau}$$

$$\frac{}{\mid [] \vdash_{\mathrm{poly}} \underline{r} : R}\ r \in \mathbb{R} \qquad \frac{}{\mid [] \vdash_{\mathrm{poly}} \underline{r} : R_{>0}}\ r \in \mathbb{R}_{>0}$$

$$\frac{}{\mid [] \vdash_{\mathrm{poly}} \underline{\circ} : \iota \to \iota \to \iota}\ \circ \in \{+, \cdot\} \qquad \frac{}{\mid [] \vdash_{\mathrm{poly}} \underline{-} : R \to R}$$

$$\frac{\Gamma \mid \Sigma \vdash_{\mathrm{poly}} L : R \quad \Gamma \mid \Sigma' \vdash_{\mathrm{poly}} M : \sigma \quad \Gamma \mid \Sigma'' \vdash_{\mathrm{poly}} N : \sigma}{\Gamma \mid \Sigma +\!\!+ \Sigma' +\!\!+ \Sigma'' \vdash_{\mathrm{poly}} \mathbf{if}\, L < 0\, \mathbf{then}\, M\, \mathbf{else}\, N : \sigma}$$

$$\frac{}{\mid [s_j \sim \mathcal{D}] \vdash_{\mathrm{poly}} \mathbf{sample}_{\mathcal{D}} : R}\ \mathcal{D} \text{ has finite moments}$$

$$\frac{\Gamma, y : \tau_1 \mid \Sigma \vdash_{\mathrm{poly}} M : \tau_2}{\Gamma \mid [] \vdash_{\mathrm{poly}} \lambda y.\, M : \tau_1 \bullet \Sigma \to \tau_2} \qquad \frac{\Gamma \mid \Sigma_1 \vdash_{\mathrm{poly}} M : \tau_1 \bullet \Sigma_3 \to \tau_2 \quad \Gamma \mid \Sigma_2 \vdash_{\mathrm{poly}} N : \tau_1}{\Gamma \mid \Sigma_1 +\!\!+ \Sigma_2 +\!\!+ \Sigma_3 \vdash_{\mathrm{poly}} M\, N : \tau_2}$$

图 7: $\vdash_{\mathrm{poly}}$ 的类型判断。

引理 5（基本）。如果 $\boldsymbol{\theta},\ x_1 : \tau_1, \ldots, x_\ell : \tau_\ell \mid \Sigma \vdash_{\mathrm{poly}} M : \tau$，$n \in \mathbb{N},\ \xi_1 \in \mathcal{P}^{(n)}_{\tau_1}, \ldots, \xi_\ell \in \mathcal{P}^{(n)}_{\tau_\ell}$那么 $[\![M]\!] * \langle \xi_1, \ldots, \xi_\ell \rangle \in \mathcal{P}^{(n+|\Sigma|)}_{\tau}$，其中

$$[\![M]\!] * \langle \xi_1, \ldots, \xi_\ell \rangle : \boldsymbol{\Theta} \times \mathbb{R}^{n+|\Sigma|} \to [\![\tau]\!]$$
$$(\boldsymbol{\theta}, \mathbf{s} +\!\!+ \mathbf{s}') \mapsto [\![M]\!]((\boldsymbol{\theta}, \xi_1(\boldsymbol{\theta}, \mathbf{s}), \ldots, \xi_\ell(\boldsymbol{\theta}, \mathbf{s})), \mathbf{s}')$$

证明。我们通过对 M 进行归纳法来证明该结论。

1. 对于常量 r 和变量 x_i，这是显而易见的；对于参数 θ_i，则由假设 1 保证。

2. $[\![\mathbf{采样}_{\mathcal{D}}]\!]() \ [s]) = s$ 显然具有有限矩，因为 $\mathcal{D}$ 具有有限矩。

3. 接下来，为了证明 $[\![+]\!] \in \mathcal{P}^{(0)}_{\iota \to \iota \to \iota}$（乘法的验证方法类似）令$n_1,\ n_2 \in \mathbb{N}$, $f_1 \in \mathcal{P}^{(n_1)}_{\iota}$, $f_2 \in \mathcal{P}^{(n_1+n_2)}_{\iota}$。根据定义，$f_1$ 和 f_2 分别被某个 g_1 和 g_2一致受控，且具有有限矩。根据引理 3，$g_1 + g_2$ 具有有限矩，并且

$$\begin{aligned}|([\![+]\!] \odot f_1 \odot f_2)(\boldsymbol{\theta}, \mathbf{s}_1 +\!\!+ \mathbf{s}_2)| &\le |f_1(\boldsymbol{\theta}, \mathbf{s}_1)| + |f_2(\boldsymbol{\theta}, \mathbf{s}_1 +\!\!+ \mathbf{s}_2)| \\ &\le g_1(\mathbf{s}_1) + g_2(\mathbf{s}_1 +\!\!+ \mathbf{s}_2)\end{aligned}$$

4. 关于 $-$ 的推理很直接，而且 $^{-1}$、exp 和 log 不会出现。 ⌟

5. 关于条件表达式的结论可由引理 4 得出。

6. 对于应用，由归纳假设和定义可立即得出。

假设 $\boldsymbol{\theta}, x_1 : \tau_1, \ldots, \boldsymbol{\theta}, x_\ell : \tau_\ell \mid \Sigma_1 +\!\!+ \Sigma_2 +\!\!+ \Sigma_3 : \tau_\ell \vdash_{\mathrm{poly}} M\, N : \tau$因为 $\boldsymbol{\theta}, x_1 : \tau_1, \ldots, \boldsymbol{\theta}, x_\ell : \tau_\ell \mid \Sigma_1 : \tau_\ell \vdash_{\mathrm{poly}} M : \tau' \bullet \Sigma_3 \to \tau$并且$\boldsymbol{\theta}, x_1 : \tau_1, \ldots, \boldsymbol{\theta}, x_\ell : \tau_\ell \mid \Sigma_2 : \tau_\ell \vdash_{\mathrm{poly}} M\, N : \tau'$。

设 $n \in \mathbb{N}$和 $\xi_1 \in \mathcal{P}^{(n)}_{\tau_1}, \ldots, \xi_\ell \in \mathcal{P}^{(n)}_{\tau_\ell}$。根据归纳假设，

$$[\![M]\!] * \langle \xi_1, \ldots, \xi_\ell \rangle \in \mathcal{P}^{(n+|\Sigma_1|)}_{\tau' \bullet \Sigma_3 \to \tau} \qquad [\![N]\!] * \langle \xi_1, \ldots, \xi_\ell \rangle \in \mathcal{P}^{(n+|\Sigma_1|+|\Sigma_2|)}_{\tau'}$$

根据 $\mathcal{P}^{(n+|\Sigma_1|}_{)\ \tau'\bullet\Sigma_3\to\tau}$ 的定义,

$$(\llbracket M\rrbracket * \langle \xi_1, \ldots, \xi_\ell\rangle) \odot (\llbracket N\rrbracket * \langle \xi_1, \ldots, \xi_\ell\rangle) \in \mathcal{P}_\tau^{(n+|\Sigma_1|+|\Sigma_2|+|\Sigma_3|)}$$

根据 $\odot$ 和 $*$的定义,

$$(\llbracket M\rrbracket * \langle \xi_1, \ldots, \xi_\ell\rangle) \odot (\llbracket N\rrbracket * \langle \xi_1, \ldots, \xi_\ell\rangle) = \llbracket M\ N\rrbracket * \langle \xi_1, \ldots, \xi_\ell\rangle$$

7. 对于抽象，设 $\boldsymbol{\theta}, x_1\colon \tau_1, \ldots, x_\ell\colon \tau_\ell \mid [] \vdash_{\text{poly}} \lambda y.\, M\colon \tau \bullet \Sigma \to \tau'$，因为 $\boldsymbol{\theta}, x_1\colon \tau_1, \ldots, x_\ell\colon \tau_\ell, y\colon \tau \mid \Sigma \vdash_{\text{poly}} M\colon \tau'$；令 $n \in \mathbb{N}$和 $\xi_1 \in \mathcal{P}_{\tau_1}^{(n)}, \ldots, \xi_\ell \in \mathcal{P}_{\tau_\ell}^{(n)}$。

为证明该结论，设 $n_2 \in \mathbb{N}$ 和 $g \in \mathcal{P}_\tau^{(n+n_2}$)。根据逻辑谓词的定义，我们需要验证 $(\llbracket M\rrbracket * \langle \xi_1, \ldots, \xi_\ell\rangle) \odot g \in \mathcal{P}_{\tau'}^{(n+n_2+|\Sigma|}$)。
Call $\widehat{\xi_i}$ ξ_i 到 $\boldsymbol{\Theta} \times \mathbb{R}^{n+n_2} \to \mathbb{R}$的扩展。根据归纳假设,

$$\llbracket M\rrbracket * \langle \widehat{\xi_1}, \ldots, \widehat{\xi_\ell}, g\rangle \in \mathcal{P}_{\tau'}^{(n+n_2+|\Sigma|)}$$

最后，只需观测到

$$(\llbracket M\rrbracket * \langle \xi_1, \ldots, \xi_\ell\rangle) \odot g = \llbracket M\rrbracket * \langle \widehat{\xi_1}, \ldots, \widehat{\xi_\ell}, g\rangle$$

C.3 第 4.3 节的补充材料

见图 8。

C.4 第 4.4 节的补充材料

我们在 **VectFr** 上定义逻辑谓词 $\mathcal{Q}_\tau^{(n)}\ \boldsymbol{\Theta} \times \mathbb{R}^n \to \llbracket\tau\rrbracket$：

1. $f \in \mathcal{Q}_{\iota^{(e)}}^{(n)}$ 如果 (a) $\log^e \circ f$ 的二阶及以下偏导数一致受控于一个具有有限矩的函数；(b) 如果 $\iota^{(e)}$ 是 $R_{>0}^{(0)}$，那么 f 受控于一个正的常值函数

2. $f \in \mathcal{P}_{\tau_1\bullet\Sigma_3\to\tau_2}^{(n)}$ 如果对于所有 $n_2 \in \mathbb{N}$ 和 $g \in \mathcal{Q}_{\tau_1}^{(n+n_2}$)，$f \odot g \in \mathcal{Q}_{\tau_2}^{(n+n_2+|\Sigma_3|}$)。

引理 10（基本）。如果 $\theta_1\colon \iota_1^{(0)}, \ldots, \theta_m\colon \iota_m^{(0)}, x_1\colon \tau_1, \ldots, x_\ell\colon \tau_\ell \mid \Sigma \vdash_{\text{SGD}} M\colon \tau$, $n \in \mathbb{N}$, $\xi_1 \in \mathcal{Q}_{\tau_1}^{(n)}, \ldots, \xi_\ell \in \mathcal{Q}_{\tau_\ell}^{(n)}$那么 $\llbracket M\rrbracket_\eta * \langle \xi_1, \ldots, \xi_\ell\rangle \in \mathcal{Q}_\tau^{(n+|\Sigma|)}$。

证明。与引理 5 类似，利用对数和偏导数的标准规则。

$$\frac{}{\iota^{\alpha} \sqsubseteq_? \iota^{\alpha'}}\ \text{(cond. subt. 1)} \qquad \frac{}{R^{\alpha}_{>0} \sqsubseteq_? R^{\alpha'}}\ \text{(cond. subt. 2)}$$

$$\frac{\tau_1' \sqsubseteq_? \tau_1 \quad \tau_2 \sqsubseteq_? \tau_2'}{(\tau_1 \bullet \Sigma \to \tau_2) \sqsubseteq_? (\tau_1' \bullet \Sigma \to \tau_2')}$$

(a) Subtyping

$$\frac{\Gamma \mid \Sigma \vdash_? M : \tau}{\Gamma' \mid \Sigma \vdash_? M : \tau'}\ \Gamma \sqsubseteq_? \Gamma', \tau \sqsubseteq_? \tau' \qquad \frac{}{x : \tau \mid [] \vdash_? x : \tau}$$

$$\frac{\Gamma, y : \tau_1 \mid \Sigma \vdash_? M : \tau_2}{\Gamma \mid [] \vdash_? \lambda y.\, M : \tau_1 \bullet \Sigma \to \tau_2} \qquad \frac{\Gamma \mid \Sigma_2 \vdash_? M : \tau_1 \bullet \Sigma_3 \to \tau_2 \quad \Gamma \mid \Sigma_1 \vdash_? N : \tau_1}{\Gamma \mid \Sigma_1 +\!\!+ \Sigma_2 +\!\!+ \Sigma_3 \vdash_? M\, N : \tau_2}$$

$$\frac{}{\theta_i : \iota^{\alpha} \mid [] \vdash_? \theta_i : \iota^{\alpha}}\ \text{(cond. Para)} \qquad \frac{}{\mid [] \vdash_? \underline{r} : \iota^{\alpha}}\ \text{(cond. Const)}, r \in [\![\iota]\!]$$

$$\frac{}{\mid [] \vdash_? \underline{+} : \iota^{\alpha_1} \to \iota^{\alpha_2} \to \iota^{\alpha}}\ \text{(cond. Add)}$$

$$\frac{}{\mid [] \vdash_? \underline{\cdot} : \iota^{\alpha_1} \to \iota^{\alpha_2} \to \iota^{\alpha}}\ \text{(cond. Mul)}$$

$$\frac{}{\mid [] \vdash_? \underline{-} : R^{\alpha} \to R^{\alpha}}\ \text{(cond. Min)} \qquad \frac{}{\mid [] \vdash_? \underline{^{-1}} : R^{\alpha}_{>0} \to R^{\alpha}_{>0}}\ \text{(cond. Inv)}$$

$$\frac{}{\mid [] \vdash_? \underline{\exp} : R^{\alpha} \to R^{\alpha'}_{>0}}\ \text{(cond. Exp)} \qquad \frac{}{\mid [] \vdash_? \underline{\log} : R^{\alpha}_{>0} \to R^{\alpha'}}\ \text{(cond. Log)}$$

$$\frac{\Gamma \mid \Sigma \vdash_? L : \iota^{\alpha} \quad \Gamma \mid \Sigma' \vdash_? M : \sigma \quad \Gamma \mid \Sigma'' \vdash_? N : \sigma}{\Gamma \mid \Sigma +\!\!+ \Sigma' +\!\!+ \Sigma'' \vdash_? \mathbf{if}\, L < 0\, \mathbf{then}\, M\, \mathbf{else}\, N : \sigma}\ \text{(cond. If)}$$

$$\frac{}{\mid [s_j \sim \mathcal{D}] \vdash_? \mathbf{sample}_{\mathcal{D}} : R^{\alpha}}\ \text{(cond. Sample)}$$

(b) Typing rules for $\vdash_?$

图 8：带注释的通用类型系统。

$$\frac{\Gamma \mid \Sigma \vdash_{\mathrm{SGD}} M : \tau}{\Gamma' \mid \Sigma \vdash_{\mathrm{SGD}} M : \tau'}\ \Gamma \sqsubseteq_{\mathrm{SGD}} \Gamma', \tau \sqsubseteq_{\mathrm{SGD}} \tau' \qquad \frac{}{x : \tau \mid [] \vdash_{\mathrm{SGD}} x : \tau}$$

$$\frac{\Gamma, y : \tau_1 \mid \Sigma \vdash_{\mathrm{SGD}} M : \tau_2}{\Gamma \mid [] \vdash_{\mathrm{SGD}} \lambda y.\, M : \tau_1 \bullet \Sigma \to \tau_2}$$

$$\frac{\Gamma \mid \Sigma_2 \vdash_{\mathrm{SGD}} M : \tau_1 \bullet \Sigma_3 \to \tau_2 \quad \Gamma \mid \Sigma_1 \vdash_{\mathrm{SGD}} N : \tau_1}{\Gamma \mid \Sigma_1 +\!\!+ \Sigma_2 +\!\!+ \Sigma_3 \vdash_{\mathrm{SGD}} M\, N : \tau_2}$$

$$\frac{}{\theta_i : \iota^{(0)} \mid [] \vdash_{\mathrm{SGD}} \theta_i : \iota^{(0)}} \qquad \frac{}{\mid [] \vdash_{\mathrm{SGD}} \underline{r} : \iota^{(0)}}\ r \in [\![\iota]\!]$$

$$\frac{}{\mid [] \vdash_{\mathrm{SGD}} \underline{+} : \iota^{(0)} \to \iota^{(0)} \to \iota^{(0)}} \qquad \frac{}{\mid [] \vdash_{\mathrm{SGD}} \underline{\cdot} : \iota^{(e)} \to \iota^{(e)} \to \iota^{(e)}}$$

$$\frac{}{\mid [] \vdash_{\mathrm{SGD}} \underline{-} : R^{(0)} \to R^{(0)}} \qquad \frac{}{\mid [] \vdash_{\mathrm{SGD}} \underline{^{-1}} : R^{(e)}_{>0} \to R^{(e)}_{>0}}$$

$$\frac{}{\mid [] \vdash_{\mathrm{SGD}} \underline{\exp} : R^{(0)} \to R^{(1)}_{>0}} \qquad \frac{}{\mid [] \vdash_{\mathrm{SGD}} \underline{\log} : R^{(e)}_{>0} \to R^{(0)}}$$

$$\frac{\Gamma \mid \Sigma \vdash_{\mathrm{SGD}} L : \iota^{(0)} \quad \Gamma \mid \Sigma' \vdash_{\mathrm{SGD}} M : \sigma \quad \Gamma \mid \Sigma'' \vdash_{\mathrm{SGD}} N : \sigma}{\Gamma \mid \Sigma +\!\!+ \Sigma' +\!\!+ \Sigma'' \vdash_{\mathrm{SGD}} \mathbf{if}\, L < 0\, \mathbf{then}\, M\, \mathbf{else}\, N : \sigma}$$

$$\frac{}{\mid [s_j \sim \mathcal{D}] \vdash_{\mathrm{SGD}} \mathbf{sample}_{\mathcal{D}} : R^{(0)}}\ \mathcal{D} \text{ has finite moments}$$

Fig. 9: Typing rules for $\vdash_{\mathrm{SGD}}$

D 第5章节的补充材料

例*12*（发散）。假设 $M \equiv \mathbf{如果}\ ((\lambda z.\, z^3 + \theta)\ \mathbf{采样}\ _{\mathcal{N}}) < 0\ \mathbf{则}\ 0\ \mathbf{否则}\ 1$。令 $\phi_\theta(z) := z^3 + \theta$。注意，尽管 $\phi_\theta : \mathbb{R} \to \mathbb{R}$ 是双射，但它不是微分同胚，因为 $\phi_\theta^{-1}(\alpha) = \sqrt[3]{\alpha - \theta}$ 在 $\alpha = \theta$ 处不可微。接着

$$\mathbb{E}_{z\sim\mathcal{N}}[[\![M]\!](\theta, z)] = \int_{-\sqrt[3]{-\theta}}^{\infty} \mathcal{N}(z \mid 0, 1)\, \mathrm{d}z$$

$$\frac{\partial}{\partial\theta}\mathbb{E}_{z\sim\mathcal{N}}[[\![M]\!](\theta, z)] = \frac{1}{3} \cdot \mathcal{N}(-\sqrt[3]{-\theta} \mid 0, 1) \cdot \theta^{-\frac{2}{3}}$$

因此，$\theta \mapsto \mathbb{E}_{z\sim\mathcal{N}}[[\![M]\!](\boldsymbol{\theta}, z)]$ 在0处不可微。此外，对于 $\theta = 0$,

$$\mathbb{E}_{z\sim\mathcal{N}}\left[\frac{\partial}{\partial\theta}[\![M(\boldsymbol{\theta}, z)]\!]_\eta\right] = \mathbb{E}_{z\sim\mathcal{N}}\left[\sigma'_\eta(z^3)\right] \to \infty$$

D.1 一致收敛与几乎一致收敛的性质

令 $\mu(U) = \mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[[\mathbf{s} \in U]]$，其中 $\mathcal{D}$ 具有有限矩，且 $\boldsymbol{\phi_\theta}$ 是微分同胚。我们继续假设 $\boldsymbol{\Theta}$具有紧致性。

引理 11. $\lim_{k\in\mathbb{N}} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mu(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbb{R}^n \setminus \mathbf{B}_k(\mathbf{0}))) = 0$

$$\frac{\Gamma \mid \Sigma \vdash_{\text{unif}} M : \tau}{\Gamma' \mid \Sigma \vdash_{\text{unif}} M : \tau'}\ \Gamma \sqsubseteq_{\text{unif}} \Gamma', \tau \sqsubseteq_{\text{unif}} \tau' \qquad \frac{}{x : \tau \mid [] \vdash_{\text{unif}} x : \tau}$$

$$\frac{\Gamma, y : \tau_1 \mid \Sigma \vdash_{\text{unif}} M : \tau_2}{\Gamma \mid [] \vdash_{\text{unif}} \lambda y.\, M : \tau_1 \bullet \Sigma \to \tau_2} \qquad \frac{\Gamma \mid \Sigma_2 \vdash_{\text{unif}} M : \tau_1 \bullet \Sigma_3 \to \tau_2 \quad \Gamma \mid \Sigma_1 \vdash_{\text{unif}} N : \tau_1}{\Gamma \mid \Sigma_1 +\!\!+ \Sigma_2 +\!\!+ \Sigma_3 \vdash_{\text{unif}} M\, N : \tau_2}$$

$$\frac{}{\theta_i : \iota^{(\mathbf{f},\emptyset)} \mid [] \vdash_{\text{unif}} \theta_i : \iota^{(\mathbf{f},\emptyset)}} \qquad \frac{}{\mid [] \vdash_{\text{unif}} \underline{r} : \iota^{(\mathbf{f},\emptyset)}}\ r \in [\![\iota]\!]$$

$$\frac{}{\mid [] \vdash_{\text{unif}} \underline{\circ} : \iota^{(\mathbf{f},\Delta)} \to \iota^{(\mathbf{f},\Delta)} \to \iota^{(\mathbf{f},\Delta)}}\ \circ \in \{+, \cdot\}$$

$$\frac{}{\mid [] \vdash_{\text{unif}} \underline{\circ} : \iota^{(\mathbf{t},\Delta_1)} \to \iota^{(\mathbf{t},\Delta_2)} \to \iota^{(\mathbf{t},\Delta_1 \cup \Delta_2)}}\ \circ \in \{+, \cdot\}, \Delta_1 \cap \Delta_2 = \emptyset$$

$$\frac{}{\mid [] \vdash_{\text{unif}} \underline{-} : R^{(g,\Delta)} \to R^{(g,\Delta)}} \qquad \frac{}{\mid [] \vdash_{\text{unif}} \underline{^{-1}} : R^{(g,\Delta)}_{>0} \to R^{(g,\Delta)}_{>0}}$$

$$\frac{}{\mid [] \vdash_{\text{unif}} \underline{\exp} : R^{(g,\Delta)} \to R^{(g,\Delta)}_{>0}} \qquad \frac{}{\mid [] \vdash_{\text{unif}} \underline{\log} : R^{(g,\Delta)}_{>0} \to R^{(g,\Delta)}}$$

$$\frac{\Gamma \mid \Sigma \vdash_{\text{unif}} L : \iota^{(\mathbf{t},\Delta)} \quad \Gamma \mid \Sigma' \vdash_{\text{unif}} M : \sigma \quad \Gamma \mid \Sigma'' \vdash_{\text{unif}} N : \sigma}{\Gamma \mid \Sigma +\!\!+ \Sigma' +\!\!+ \Sigma'' \vdash_{\text{unif}} \mathbf{if}\, L < 0\, \mathbf{then}\, M\, \mathbf{else}\, N : \sigma}$$

$$\frac{}{\mid [s_j \sim \mathcal{D}] \vdash_{\text{unif}} \mathbf{sample}_{\mathcal{D}} : R^{(\mathbf{t},\{s_j\})}}$$

$$\frac{\boldsymbol{\theta} \mid [] \vdash_{\text{unif}} T : R^{\alpha} \to R^{\alpha}}{\boldsymbol{\theta} \mid [s_j \sim \mathcal{D}] \vdash_{\text{unif}} \mathbf{transform}\, \mathbf{sample}_{\mathcal{D}}\, \mathbf{by}\, T : R^{(\mathbf{t},\{s_j\})}}\ T \text{ diffeomorphic}$$

$$\frac{\Gamma \mid [] \vdash_{\text{unif}} M : \iota^{(\mathbf{t},\Delta)}}{\Gamma \mid [] \vdash_{\text{unif}} \underline{a \cdot} M : \iota^{(\mathbf{t},\Delta)}}\ a \in \mathbb{N}_{>0} \qquad \frac{\Gamma \mid [] \vdash_{\text{unif}} M : \iota^{(\mathbf{t},\Delta)}}{\Gamma \mid [] \vdash_{\text{unif}} M^{a}_{\_} : \iota^{(\mathbf{t},\Delta)}}\ a \in \mathbb{N}_{>0}$$

Fig. 10: Typing rules for $\vdash_{\text{unif}}$.

证明。 设$\mathbf{s}_0 \in \mathbb{R}^n$ 为任意值。

$$\begin{aligned}\delta_*^{(i)} &\coloneqq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\boldsymbol{\phi}_{\boldsymbol{\theta}}^{(i)}(\mathbf{s}_0)| \\ d_k^{(i)} &\coloneqq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sup_{\mathbf{s}\in\mathbf{B}_k(\mathbf{s}_0)} \|\nabla_{\boldsymbol{\theta}}\boldsymbol{\phi}_{\boldsymbol{\theta}}^{(i)}(\mathbf{s})\|\end{aligned}$$

因此，如果$\mathbf{s} \in \mathbf{B}_k(\mathbf{s}_0)$,

$$\begin{aligned}|\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s})^{(i)}| &\le \|\boldsymbol{\phi}_{\boldsymbol{\theta}}^{(i)}(0)\| + \langle\nabla\boldsymbol{\phi}_{\boldsymbol{\theta}}^{(i)}(\zeta), x\rangle \\ &\le \delta_*^i + \|\nabla\boldsymbol{\phi}_{\boldsymbol{\theta}}^{(i)}(\zeta)\| \cdot \|x\| \\ &\le \delta_*^i + d_k^{(i)} \cdot k\end{aligned}$$

Let

$$\delta_k^{(i)} \coloneqq \delta_*^i + d_k^{(i)} \cdot k \qquad\qquad \delta_k \coloneqq \sqrt{n} \cdot \max_{i=1}^{n} \delta_k^{(i)}$$

注意，对于$\mathbf{s} \in \mathbb{R}^n$，若 $\|\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s})\| > \delta_k$，则 $|\boldsymbol{\phi}_{\boldsymbol{\theta}}^{(i)}(\mathbf{s})| > \delta_r^{(i)}$对某个值成立，因此$\mathbf{s} \in \mathbb{R}^n \setminus \mathbf{B}_k(\mathbf{s}_0)$。由此可得，$\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbb{R}^n \setminus \mathbf{B}^{\delta_k}(\mathbf{0})) \subseteq \mathbb{R}^n \setminus \mathbf{B}_k(\mathbf{s}_0)$。

现在，只需注意到$\lim_k \mu(\mathbb{R}^n \setminus \mathbf{B}_k(\mathbf{s}_0)) = 0$。

引理 12。 对每个 $k \in \mathbb{N}$，都存在 $c > 0$，使得 $\mu(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(U \cap \mathbf{B}_k(\mathbf{0}))) \le c \cdot \mathrm{Leb}(U \cap \mathbf{B}_k(\mathbf{0}))$。

证明。 令 $f\colon \mathbb{R}^n \to \mathbb{R}$ 为 μ 的密度。则

$$\begin{aligned}\mu(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(U \cap \mathbf{B}_k(\mathbf{0}))) &= \int_{\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(U\cap\mathbf{B}_k(\mathbf{0}))} f(\mathbf{s})\,\mathrm{d}\mathbf{s} \\ &= \int_{U\cap\mathbf{B}_k(\mathbf{0})} f(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})) \cdot |\det \mathbf{J}\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})|\,\mathrm{d}\mathbf{z}\end{aligned}$$

引理 13。 假设 $f_\eta \circ \boldsymbol{\phi}_{(-)}(-)$ 几乎一致 $\longrightarrow f \circ \boldsymbol{\phi}_{(-)}(-)$ 且 $f \neq 0$ 几乎处处成立。

Then $\sigma_\eta \circ f_\eta \circ \boldsymbol{\phi}_{(-)}(-) \xrightarrow{u.a.u.} [f(\boldsymbol{\phi}_{(-)}(-)) > 0]$.

证明。 令 δ_k、ϵ_k 和 η_k 分别为 $f_\eta \circ \boldsymbol{\phi}_{(-)}(-)$ 几乎一致 $\longrightarrow f \circ \boldsymbol{\phi}_{(-)}(-)$的见证。

对于 $i \in \mathbb{N}$，定义 $V_i \coloneqq \{\mathbf{z} \in \mathbb{R}^n \mid |f(\mathbf{z})| < \frac{1}{i}\}$。对于每个 $k \in \mathbb{N}$，存在$i_k \in \mathbb{N}$，使得 $\mathrm{Leb}(V_{i_k} \cap \mathbf{B}_k(\mathbf{0})) < \frac{1}{k}$。（这是因为 $\mathrm{Leb}((-) \cap \mathbf{B}_k(\mathbf{0}))$ 是有限测度，且 $\cap_{i\in\mathbb{N}} V_i \subseteq f^{-1}(0)$ 和 $f \neq 0$几乎处处成立。）

此外，对于 $k \in \mathbb{N}$，令 $K_k \in \mathbb{N}$ 满足 $\epsilon_{K_k} < \frac{1}{2i_k}$。根据假设 3，存在$0 < \eta_k' < \eta_{K_k}$，使得对于所有 $0 < \eta < \eta_k'$ 和 $y > \frac{1}{2i_k}$，$\sigma_\eta(-y) < \frac{1}{k}$ 且 $\sigma_\eta(y) > 1 - \frac{1}{k}$。我们还定义

$$\delta_k' \coloneqq \delta_{K_k} + \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mu(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbb{R}^n \setminus \mathbf{B}_k(\mathbf{0}))) + \frac{1}{k} \qquad\qquad \epsilon_k' \coloneqq \frac{1}{k}$$

根据引理 11，$\lim \delta'_k = 0 = \lim \epsilon'_k$。

现在，假设 $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ 和 $k \in \mathbb{N}$。

定义 $U' \coloneqq U_{K_k} \cup \boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(V_{i_k})$，其中 $U_{K_k} \subseteq \mathbb{R}^n$ 是 $[f(\boldsymbol{\phi}_{(-)}(-)) > 0]$、$\boldsymbol{\theta}$ 和 K_k 对应的集合。成立

$$\begin{aligned}\mu(U') &\leq \mu(U_{K_k}) + \mu(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbb{R}^n \setminus \mathbf{B}_k(\mathbf{0}))) + \mu(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(V_{i_k} \cap \mathbf{B}_k(\mathbf{0})))\\ &\leq \mu(U_{K_k}) + \mu(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbb{R}^n \setminus \mathbf{B}_k(\mathbf{0}))) + c \cdot \mathrm{Leb}(V_{i_k} \cap \mathbf{B}_k(\mathbf{0}))\\ &\leq \delta'_k\end{aligned}$$

此外，对于 $0 < \eta < \eta'_k$ 和 $\mathbf{s} \in \mathbb{R}^n \setminus U'$，$|f_\eta(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s})) - f(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s}))| < \frac{1}{2i_k}$ 以及 $|f(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s}))| \geq \frac{1}{i_k}$，因此 $|f_\eta(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s}))| > \frac{1}{2i_k}$。

因此，$|\sigma_\eta(f_\eta(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s}))) - [f(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s})) > 0]| < \frac{1}{k}$。

引理 14. 若 $f\colon U_1 \times U_2 \to \mathbb{R}$（对于开连通的 U_1，$U_2 \subseteq \mathbb{R}$）连续可微，且 $g_\eta \xrightarrow{} g$ $\boldsymbol{\Theta} \times \mathbb{R}^n \to U_1$ $h_\eta \xrightarrow{} h$ $\boldsymbol{\Theta} \times \mathbb{R}^n \to_{u.a.u.}$ 以 U_2 g h $\mathbb{R}^n$ $f \circ \langle g_\eta, h_\eta \rangle \xrightarrow{}$ 及 $_{u.a.u.}$，和均在有界子集上有界，则 $_{u.a.u.} f \circ \langle g, h \rangle\colon \boldsymbol{\Theta} \times \mathbb{R}^n \to \mathbb{R}$。.

证明。首先注意到，由于 f 连续可微，且 g 和 h 也满足这一性质，因此 $f \circ \langle g, h \rangle$ 在 $\mathbb{R}^n$ 的有界子集上有界。

令 $\delta_k^{(i)}$、$\epsilon_k^{(i)}$ 和 $\eta_k^{(i)}$ $(i \in \{1, 2\})$ 分别为 g_η 一致几乎一致收敛$\xrightarrow{}$ g 和 h_η 一致几乎一致收敛$\xrightarrow{}$ h 的见证。不失一般性，所有 $\epsilon_k^{(i)} \leq 1$。观测到，对于 $k \in \mathbb{N}$，

$$M_k \coloneqq \sup_{(\boldsymbol{\theta}, \mathbf{s}) \in \boldsymbol{\Theta} \times \mathbf{B}_k(\mathbf{0})} \|(g(\boldsymbol{\theta}, \mathbf{s}), h(\boldsymbol{\theta}, \mathbf{s}))\| + \sqrt{2} < \infty$$

因为 $g(\boldsymbol{\Theta} \times \mathbf{B}_k(\mathbf{0}))$ 和 $h(\boldsymbol{\Theta} \times \mathbf{B}_k(\mathbf{0}))$ 根据假设（也包括假设 1）有界，因此

$$d_k \coloneqq \sup_{(x,y) \in M_k \cap (U_1 \times U_2)} \|\nabla f(x, y)\| < \infty$$

是良定义的。对于 $k \in \mathbb{N}$，存在 $K_k \geq k$，使得每个 $\sqrt{2} \cdot d_k \cdot \epsilon_{K_k}^{(i)} < \frac{1}{k}$。定义

$$\delta_k \coloneqq \mu(\mathbb{R}^n \setminus \mathbf{B}_k(\mathbf{0})) + \delta_{K_k}^{(1)} + \delta_{K_k}^{(2)} \qquad \epsilon_k \coloneqq \frac{1}{k} \qquad \eta_k \coloneqq \min\{\eta_{K_k}^{(1)}, \eta_{K_k}^{(2)}\}$$

注意，根据引理 11，$\lim \delta_k = 0 = \lim \epsilon_k$。

设 $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ 和 $k \in \mathbb{N}$。设 $V \coloneqq (\mathbb{R}^n \setminus \mathrm{B}_k(0)) \cup V(1) \cup V(2)$，其中 $V(1)$（分别还有 $V^{(2)}$）是 g（分别还有 h）的集合，$\boldsymbol{\theta}$ 和 K_k。注意，$\mu(V) \leq \delta_k$。此外，对于每个 $0 < \eta < \eta_k$ 和 $\mathbf{s} \in \mathbb{R}^n \setminus V$，$|g_\eta(\boldsymbol{\theta}, \mathbf{s})| \leq |g(\boldsymbol{\theta}, \mathbf{s})| + \epsilon_{K_k}^{(1)} \leq |g(\boldsymbol{\theta}, \mathbf{s})| + 1$（$h$ 也类似）。因此，$(g_\eta(\boldsymbol{\theta}, \mathbf{s})\ h_\eta(\boldsymbol{\theta}, \mathbf{s}))$ 与 $(g(\boldsymbol{\theta}, \mathbf{s})\ h(\boldsymbol{\theta}, \mathbf{s}))$ 之间的每个点都属于 $\mathrm{B}_{M_k}(0) \cap (U_1 \times U_2)$，因此根据中值

定理,

$$
\begin{aligned}
&|f(g_\eta(\boldsymbol{\theta},\mathbf{s}),h_\eta(\boldsymbol{\theta},\mathbf{s}))-f(g(\boldsymbol{\theta},\mathbf{s}),h(\boldsymbol{\theta},\mathbf{s}))|\\
&\leq \sup_{\zeta\in\mathbf{B}_{M_k}(\mathbf{0})\cap(U_1\times U_2)} |\langle\nabla f(\zeta),(g_\eta(\boldsymbol{\theta},\mathbf{s})-g(\boldsymbol{\theta},\mathbf{s}),h_\eta(\boldsymbol{\theta},\mathbf{s})-h(\boldsymbol{\theta},\mathbf{s}))\rangle|\\
&\leq \sup_{\zeta\in\mathbf{B}_{M_k}(\mathbf{0})\cap(U_1\times U_2)} \|\nabla f(\zeta)\|\cdot\|(g_\eta(\boldsymbol{\theta},\mathbf{s})-g(\boldsymbol{\theta},\mathbf{s}),h_\eta(\boldsymbol{\theta},\mathbf{s})-h(\boldsymbol{\theta},\mathbf{s}))\|\\
&< d_k\cdot\sqrt{2}\cdot\max\{\epsilon_{K_k}^{(1)},\epsilon_{K_k}^{(2)}\}\\
&\leq \epsilon_k
\end{aligned}
$$

第二步中使用了柯西–施瓦茨不等式。

引理 6. 设 $f, f_\eta:\ \boldsymbol{\Theta}\times\mathbb{R}^n\to\mathbb{R}$ 具有有限矩。

$$\textit{If } f_\eta \xrightarrow{u.a.u.} f \textit{ then } \mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[f_\eta(\boldsymbol{\theta},\mathbf{s})]\xrightarrow{unif.}\mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[f(\boldsymbol{\theta},\mathbf{s})].$$

证明。 只需证明 $\mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[|f_\eta(\boldsymbol{\theta},\mathbf{s})-f(\boldsymbol{\theta},\mathbf{s})|]$一致收敛到 0。

根据假设，存在 $M>0$，使得对于所有 $\eta>0$ 和 $\boldsymbol{\theta}\in\boldsymbol{\Theta}$，$\mathbb{E}_{\mathbf{s}\sim\mathcal{D}}\left[|f_\eta(\boldsymbol{\theta},\mathbf{s})-f(\boldsymbol{\theta},\mathbf{s})|^2\right]<M$。

设 $\epsilon>0$。由 f_η 到 f 的一致几乎一致收敛，存在 k，使得 $\delta_k\cdot M,\ \epsilon_k<\frac{\epsilon}{2}$。

假设 $\boldsymbol{\theta}\in\boldsymbol{\Theta}$ 和 $\eta<\eta_k$。令 $U\subseteq\mathbb{R}^n$ 为 f_η、k 和 $\boldsymbol{\theta}$几乎一致收敛的见证。特别地，$\mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[[\mathbf{s}\in U]]\cdot M<\delta_k\cdot M<\frac{\epsilon}{2}$，并且对每个 $\mathbf{s}\in\mathbb{R}^n\setminus U$，$|f_\eta(\boldsymbol{\theta},\mathbf{s})-f(\boldsymbol{\theta},\mathbf{s})|<\epsilon_k<\frac{\epsilon}{2}$。

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[|f_\eta(\boldsymbol{\theta},\mathbf{s})-f(\boldsymbol{\theta},\mathbf{s})|\leq \mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[[\mathbf{s}\in U]\cdot|f_\eta(\boldsymbol{\theta},\mathbf{s})-f(\boldsymbol{\theta},\mathbf{s})|]+\mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[[\mathbf{s}\in\mathbb{R}^n\setminus U]\cdot|f_\eta(\boldsymbol{\theta},\mathbf{s})-f(\boldsymbol{\theta},\mathbf{s})|\leq \mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[[\mathbf{s}\in U]]\cdot\mathbb{E}_{\mathbf{s}\sim\mathcal{D}}\left[|f_\eta(\boldsymbol{\theta},\mathbf{s})-f(\boldsymbol{\theta},\mathbf{s})|^2\right]+\mathbb{E}_{\mathbf{s}\sim\mathcal{D}}\left[[\mathbf{s}\in\mathbb{R}^n\setminus U]\cdot\frac{\epsilon}{2}\right]\leq\epsilon
\end{aligned}
$$

D.2 类型可靠性

为了汇总变换的影响，我们采用微分同胚列表（通常记作 Φ）。微分同胚列表 $[\phi_-^{(1)}()\ \cdots,\phi_-^{(n)}()]$ 定义了一个微分同胚

$$
\begin{aligned}
\boldsymbol{\phi}_{(-)}&:\boldsymbol{\Theta}\times\mathbb{R}^n\to\mathbb{R}^n\\
(\boldsymbol{\theta},[s_1,\ldots,s_n])&\mapsto\left(\phi_{\boldsymbol{\theta}}^{(1)}(s_1),\ldots,\phi_{\boldsymbol{\theta}}^{(n)}(s_n)\right)
\end{aligned}
$$

并使用连接记法。

我们提出如下无穷逻辑关系 $\mathcal{R}_\tau^\Phi$，用于描述**VectFr** 中的元素序列 $\boldsymbol{\Theta}\times\mathbb{R}^n\to[\![\tau]\!]$（对应于平滑化）与**QBS** 中的元素序列$\boldsymbol{\Theta}\times\mathbb{R}^n\to[\![\tau]\!]$（对应于可测标准语义）之间的关系:

40 BasimKhajwal、C.-H. Luke Ong 和 Dominik Wagner(✉)

1 $f_\eta\ f\ \in \mathcal{R}^{\Phi}_{\iota^{(\mathbf{f},\Delta)}}\quad f_\eta \longrightarrow f$. (,) 若u.a.u.. $(f_\eta, f) \in \mathcal{R}^{\Phi}_{\iota^{(\mathbf{t},\Delta)}}$若 f_ηu.a.u.$\longrightarrow f$,
2
$f_\eta = g_\eta \circ \phi_{(-)}$且 $f = g \circ \phi_{(-)}$，其中 (a) ϕ 按上述方式由 Φ 定义；(b) g : $\mathbb{R}^n \to \mathbb{R}$ 是分段解析且非恒定的；(c) 在每一段上， g 只能依赖于（变换后的） z_j ，若 $s_j \in \Delta$3. $(f_\eta, f) \in \mathcal{R}^{\Phi}_{\tau_1 \bullet \Sigma_3 \to \tau_2}$当且仅当对所有 Φ_2 以及满足 $\in \mathcal{R}^{\Phi \uplus \Phi_2}_{\tau_1}$ 的 (g_η, g)，都存在 Φ_3使得 $|\Phi_3| = |\Sigma_3|$ 且 $\in \mathcal{R}^{\Phi \uplus \Phi_2 \uplus \Phi_3}_{\tau_2}$的 $(f_\eta \odot g_\eta, f \odot g)$。

注意，条目 2b 蕴含 $f \neq 0$ 几乎处处成立，因为非常数解析函数仅在零测集上为零 [28] ，且微分同胚保持零测性。

引理 15. 若$(f_\eta, f) \in \mathcal{R}^{\Phi}_{R^{(\mathbf{t},\Delta)}}$且$(g_\eta, g)$，$(h_\eta, h) \in \mathcal{R}^{\Phi}_{\sigma}$则

$$((\sigma_\eta \circ (-f_\eta)) \cdot g_\eta + (\sigma_\eta \circ f_\eta) \cdot h_\eta, [f(-) < 0] \cdot g + [f(-) \geq 0] \cdot h) \in \mathcal{R}^{\Phi}_{\sigma}$$

证明。 我们重点讨论 σ 为带注释的基类型这一情形的论证，尤其是 $\iota^{(\mathbf{t},\Delta)}$，这是最有意思的情形；高阶情形可仿照引理 4 类似地推广。显然，条目 2b 和 2c 均成立，且由引理 13 和 14 可得 u.a.u. 收敛。

直观而言， Φ 描述了执行过程中可能 已被采样的样本如何变换。我们可以添加额外的样本，而这些样本会被忽略:

引理 *16.* 设 $(f_\eta, f) \in \mathcal{R}_{\Phi,\tau}$和 Φ' 是一个微分同胚列表。那么 $(g_\eta, g) \in \mathcal{R}_{\Phi \uplus \Phi',\tau}$，其中 $g_\eta(\boldsymbol{\theta}, s \uplus s') := f_\eta(\boldsymbol{\theta}, s)$，且 $g(\boldsymbol{\theta}, s \uplus s') := f(\boldsymbol{\theta}, s)$。

引理 17（基本）。若 $\theta_1 : \iota_1^{(\mathbf{f},\emptyset)}, \ldots, \theta_m : \iota_m^{(\mathbf{f},\emptyset)}, x_1 : \tau_1, \ldots, x_\ell : \tau_\ell . |\Sigma \vdash M : \quad \in \mathcal{R}^{\Phi}_{\tau_\ell} \qquad \Phi' \qquad |\Sigma| = |\Phi'| \qquad [\![M]\!]_\eta *$, Φ 是一个微分同胚列表，$(\xi_\eta^{(1)}, \xi^{(1)}) \in \mathcal{R}^{\Phi}_{\tau_1}$, $(\xi_\eta^{(\ell)}, \xi^{(\ell)})$，那么存在一个微分同胚列表，使得且 $(\langle \xi_\eta^{(1)}, \ldots, \xi_\eta^{(\ell)} \rangle, [\![M]\!] * \langle \xi^{(1)}, \ldots, \xi^{(\ell)} \rangle) \in \mathcal{R}^{\Phi \uplus \Phi'}_{\tau}$，其中 $*$的定义同引理 *5*。

证明。 通过对类型判定进行归纳法即可证明该结论。我们重点讨论最有意思的情形:

1. 对于条件表达式，我们利用归纳假设和引理 15。2. 假设 $\boldsymbol{\theta} \mid [s_j \sim \mathcal{D}] \vdash_{\text{unif}}$ **变换采样** $_{\mathcal{D}}$ **通过** T: $R^{(\mathbf{t},\{s_j\})}$，因为 T是微分同胚的。我们定义

$$g(s_j) := s_j \qquad \phi_{\boldsymbol{\theta}}(s) := [\![T]\!](\boldsymbol{\theta}, [])(s, []) = [\![T]\!]_\eta(\boldsymbol{\theta}, [])(s, [])$$

因此我们很容易看出

$$[\![\mathbf{transform\ sample}_{\mathcal{D}}\ \mathbf{by}\ T]\!]_\eta = g \circ \phi_{(-)} = [\![\mathbf{transform\ sample}_{\mathcal{D}}\ \mathbf{by}\ T]\!]$$

并且([**对采样值应用变换** $_{\mathcal{D}}$， $T]\!]_\eta$， [**对采样值应用变换** $_{\mathcal{D}}$， $T]\!]) \in \mathcal{R}^{[\phi_{(-)}]}_{R^{(\mathbf{t},\{s_j\})}}$ 立即可得。

3. 对于加法，我们关注有趣的情形 $|[]\vdash_{\text{unif}}+:\ \iota^{(\mathbf{t},\Delta_1)}\underset{\multimap}{\to}\iota^{(\mathbf{t},\Delta_2)}\to\iota^{(\mathbf{t},\Delta_1\cup\Delta_2)}$，其中 $\Delta_1\cap\Delta_2=\emptyset$。令 Φ、Φ_1 和 Φ_2 为微分同胚列表，$(f_{(1)\eta},\ f^{(1)})\in\mathcal{R}^{\Phi+\!\!+\Phi_1}_{\iota^{(\mathbf{t},\Delta_1)}}$ 和 $(f_{(2)\eta},\ f^{(2)})\in\mathcal{R}^{\Phi+\!\!+\Phi_1+\!\!+\Phi_2}_{\iota^{(\mathbf{t},\Delta_2)}}$。根据定义，存在分解

$$f^{(1)}_\eta=g^{(1)}_\eta\circ\boldsymbol{\phi}^{(1)}_{(-)}\quad f^{(1)}=g^{(1)}\circ\boldsymbol{\phi}^{(1)}_{(-)}\quad f^{(2)}_\eta=g^{(2)}_\eta\circ\boldsymbol{\phi}^{(2)}_{(-)}\quad f^{(2)}=g^{(2)}\circ\boldsymbol{\phi}^{(2)}_{(-)}$$

令 $g^{(1)}_\eta$和 $\widehat{g^{(1)}}$ 分别为 $g^{(1)}_\eta$和 $g^{(1)}$在 $\mathbb{R}^{|\Phi|+|\Phi_1|+|\Phi_2|}\to\mathbb{R}$上的扩展。注意，

$$[\![\underline{+}]\!]_\eta\odot f^{(1)}_\eta\odot f^{(2)}_\eta=(\widehat{g^{(1)}_\eta}+g^{(2)}_\eta)\circ\boldsymbol{\phi}^{(2)}_{\boldsymbol{\theta}}\quad [\![\underline{+}]\!]\odot f^{(1)}\odot f^{(2)}=(\widehat{g^{(1)}}+g^{(2)})\circ\boldsymbol{\phi}^{(2)}_{\boldsymbol{\theta}}$$

显然（利用引理 16），$g^{(1)}+g^{(2)}$ 仍是分段解析的，并且在每一段上，它取决于（经过变换的）样本，而 $g^{(1)}$ 或 $g^{(2)}$ 取决于这些样本。此外，在每一段上，$g^{(1)}+g^{(2)}$ 都不是常数，因为 $g^{(1)}$ 和 $g^{(2)}$均非常数，且依赖于不同的变量。

E 第7节的补充材料

E.1 实验设置

为生成图 5 中所示的证据下界轨迹，我们每隔 100 次迭代分别对证据下界采样 1000 次，并额外增加采样次数，以降低所呈现图表中的方差。所有估计量使用相同的随机样本，因此其轨迹中出现的噪声具有相关性。

表2比较了各估计量的平均方差，平均值取自单条优化轨迹。对于每种估计量，我们每隔100次迭代抽取1000个梯度的蒙特卡罗样本，以计算该次迭代时估计量的方差；随后对这些方差取平均。由于梯度是向量，方差通过两种方式测量：对各分量的方差取平均，以及计算 L_2 范数的方差。

随后，我们分别对每种估计量进行基准测试，测量其在固定时间预算内可完成的迭代次数，并将计算成本设为该次数的倒数。接着用它来计算工作归一化方差 [14,8]，即计算成本与方差的乘积。直观来说，我们用相对耗时作除数，因为多次运行速度更快的估计量，可以按相同比例降低方差。

E.2 模型

我们纳入了 [23], 中的模型，具体如下：

– 温度 [34] 模型描述了一个控制器，它使室温保持在设定范围内。不连续性源于控制器的离散状态

控制器只有开启或关闭两种状态，这会破坏表示室温的连续状态。给定一组带噪声的室温测量值，目标是推断 21 个时间步中每一步的控制器状态。该模型具有 41 维潜变量和 80 个 if语句。

– textmsg [11] 对每日短信发送率进行建模，目标是发现给定的 74 天数据期间发送率的变化。不可微性源于模型设定发送率发生变化的时点。该模型包含一个 3 维潜变量（两个发送率及其变化时点）和 37 个 if 语句。

– 流感 [33] 对 1969 年美国流感死亡率进行建模。每个月的死亡率取决于优势病毒株属于 1 型还是 2 型，因此每个月都会产生不连续性。给定死亡率数据，目标是推断每个月的优势病毒株。该模型包含一个 37 维潜变量和 24 个 if语句。

此外，我们还引入以下模型：

– 作弊 [11] 模拟一种差分隐私场景：对参加考试的学生进行调查，以确定作弊行为的普遍程度，同时不泄露任何个人的具体情况。学生需要抛一枚硬币：如果正面朝上，就如实回答（是否作弊）；如果反面朝上，就再抛第二枚硬币来决定如何作答。抛硬币是此处不连续性的来源。给定回答“是”的学生比例，目标是预测作弊率的后验分布。该模型包含 300 个 if语句和一个 301 维的潜在空间，不过我们只对一个维度进行优化，其余 300 个维度均为随机性来源。

– xornet 是一个简单的多层神经网络，经过训练以计算异或（XOR）函数。它采用 2-4-2-1 的网络架构，有两个输入和一个输出，且所有激活函数均为 Heaviside 阶跃函数。由于该函数在 0 处不连续，且在其他所有位置的梯度均为零，因此传统上无法对其进行基于梯度的优化。该模型具有 25 维潜在空间（对应所有权重和偏置）以及 28 个 if语句。请注意，由于分支条件并非在潜在空间中全部为仿射函数，因此此模型不适用于 Lyy18估计器。

E.3 结果分析

图 5a 中温度模型的 ELBO图显示，Reparam估计器存在偏差；与 Smooth 和 Lyy18估计器相比，它收敛到一个次优值。从该图以及表2a中的数据还可以看出，Score估计量的方差极高，且无法收敛。

图5b和图5c中的 textmsg 和流感 ELBO图均显示，所有估计量都收敛到大致相同的值，而 Score 的方差较大。在这两种模型中，Smooth估计量的工作归一化方差在两种方差度量下均最低。

对于图5d中的作弊模型，我们还可以通过另一种直观方式观察重参数化梯度的偏差。这里 Smooth 再次表现更优，工作归一化方差最低（忽略重参数化梯度的方差，因为它存在偏差）。

最后，xornet模型展示了训练基于阶跃函数的神经网络的难度。Lyy18估计器在此处不适用，因为对于非仿射条件分布（本例即属于这种情况），边界积分没有通用的高效估计器。在图5e中，Reparam估计器没有取得进展，而其他估计器则成功收敛到接近0的证据下界，表明它们学会了一个能够正确分类所有点的网络。特别是，Smooth估计量收敛得最快。

总而言之，结果展示了Reparam估计器存在偏差的情况，以及Smooth估计量不受同样限制的原因。在Lyy18估计器有定义的情况下，两者会收敛到大致相同的目标函数值；此外，平滑方法还可推广到更复杂的模型，例如具有非线性边界的神经网络。我们提出的Smooth估计量始终具有显著更低的工作归一化方差，最高可降低3个数量级。

表2：计算成本和工作归一化方差，均以相对于 Score估计量的比率给出（省略其数据，因为其对应的一行全是 1）。我们为平滑选择了 $\eta = 0.15$ 。

（a）温度

估计器	Cost	平均值(V(。))	$V(\|\|.\|\|_2)$
平滑	1.62e+00	3.17e-10	2.09e-09
重参数化	1.28e+00	1.48e-08	2.01e-08
Lyy18	9.12e+00	1.22e-06	4.76e-05

(b) textmsg

估计器	Cost	平均值(V(.))	$V(\|\|.\|\|_2)$
平滑	2.00e+00	2.29e-02	3.79e-02
重参数化	1.18e+00	1.43e-02	2.29e-02
Lyy18	4.00e+00	5.76e-02	8.46e-02

(c) 流感

估计器	Cost	平均值(V(.))	$V(\|\|.\|\|_2)$
平滑	1.47e+00	9.15e-03	4.58e-03
重参数化	1.17e+00	7.45e-03	3.68e-03
Lyy18	8.30e+00	5.88e-02	2.91e-02

(d) 作弊

估计器	Cost	平均值(V(。))	$V(\|\|.\|\|_2)$
平滑	1.59e+00	3.64e-03	5.94e-03
重参数化	9.66e-01	6.47e-19	1.74e-18
Lyy18	2.51e+00	5.39e-02	1.34e-01

(e) xornet

估计器	Cost	平均值(V(。))	$V(\|\|.\|\|_2)$
平滑	1.66e+00	9.57e-03	4.46e-02
重参数化	3.51e-01	7.55e-09	2.37e-09