有向无环图的信息梯度：用于端到端互信息最大化的基于得分函数的框架

Tadashi Wadayama 名古
屋工业大学
wadayama@nitech.ac.jp

arXiv:2601.01789v1 [cs.IT] 5 Jan 2026

摘要——**本文提出一种通用框架，用于在由随机有向无环图（DAG）表示的通信与感知系统中实现端到端互信息最大化。我们利用边缘得分函数和条件得分函数，推导出关于任意内部参数的（互）信息梯度的统一公式。我们展示了如何在标准自动微分框架中利用向量-雅可比积（VJP）高效计算该梯度，从而能够在全局资源约束下优化复杂网络。在线性多径有向无环图和非线性信道上的数值实验验证了所提出的框架；结果证实，利用通过去噪得分匹配学习得到的得分函数构建的估计器，能够准确复现真实梯度，并成功最大化端到端互信息。除最大化之外，我们还将基于得分函数的框架拓展至一种新颖的无监督范式：通过费舍尔散度最小化进行数字孪生校准。**

索引词—**互信息、非线性高斯信道、得分函数、去噪得分匹配、基于梯度的优化、有向无环图、信息梯度**

一、引言

设计能够最大化信息流的复杂通信与感知系统，是现代工程中的核心挑战。从旨在实现语义通信 [11] 的下一代无线网络，到物联网中的分布式传感器融合，设计范式正从优化单个模块转向端到端的系统设计。在这些场景中，互信息（MI） [1] 是衡量目标函数的黄金标准，能够捕捉简单相关性所无法揭示的非线性依赖关系。然而，在此类高维、多阶段系统中最大化互信息，会带来巨大的计算障碍。与通过反向传播即可轻松计算梯度的确定性损失函数（例如均方误差）不同，互信息取决于联合概率密度，而该密度会随着系统动力学以复杂的方式演变。现有的变分界主要侧重于估计互信息的数值；利用这些界对系统内部参数进行基于梯度的优化，往往会导致高方差和不稳定性。从根本上说，我们缺少一种可扩展的机制，来通过复杂的随机图“反向传播”信息度量。

在一篇配套论文 [2], 中，我们通过引入信息梯度框架，迈出了应对这一挑战的第一步。该理论建立了一个精确恒等式，将互信息的梯度与得分函数 [1], [10]——即对数似然的梯度联系起来。借助现代得分匹配技术 [3]–[5], ，该框架无需闭式似然即可高效估计梯度。然而， [2] 的适用范围仅限于形式为 $Y = f_{\eta}(X) + Z$的单阶段信道。现实系统很少如此简单；它们具有复杂的拓扑结构，包含分支路径、合并信号和多跳处理，实际上构成了随机网络。将基本公式应用于这类架构并非易事，因为单个参数可能通过多个相互作用的路径影响最终输出，这使得得分函数及相关链式法则的推导更加复杂。

本文将信息梯度框架推广至由有向无环图（*DAG*）表示的任意随机系统。我们将系统视为计算图，其中节点表示随机变量，边表示参数化的非线性变换。这意味着，现代自动微分框架（例如 PyTorch、TensorFlow 和 JAX）中可实现的任何随机网络系统，现在都可以使用我们的基于得分的估计器来优化互信息。

我们的主要贡献是提出了一个统一的梯度公式，适用于有向无环图中的任意内部参数，不受其深度或拓扑复杂度的限制。更重要的是，我们弥合了信息论与深度学习实现之间的鸿沟：我们证明，广义信息梯度可以使用标准向量-雅可比积（*VJP*） [6]高效计算。此外，我们证明，多址信道（MAC）和广播信道（BC）等经典多用户模型都可以自然地表示为有向无环图。因此，我们的方法可作为统一数值求解器，用于探索复杂非线性信道的可达速率区域；对于这类信道，闭式解析表征难以求得。为严格验证这一能力，我们证明，我们的基于得分的信息梯度能够精确复现高斯多址信道的已知容量区域。

这一泛化拓展了广泛的实际应用。它使我们能够在未知的非线性信道上端到端优化发射机和接收机神经网络，这是语义通信的一项关键要求。此外，我们将该框架拓展至一种新颖的无监督范式：数字孪生校准。通过最小化 Fisher散度（这是我们基于得分的表述自然产生的结果），我们提出了一种仅利用输出观测、使仿真参数与物理现实相匹配的方法。

II. 预备知识

本节将介绍基本符号，并回顾与我们在配套论文 [2] 中确立的基本信息梯度恒等式相关的关键概念。

A. 符号说明

我们用粗体小写字母（例如 $\boldsymbol{x}$）表示向量，用粗体大写字母（例如 $\boldsymbol{A}$）表示矩阵。随机向量用大写字母表示（例如 X，Y），其实现值用小写字母表示（例如 $\boldsymbol{x}$，$\boldsymbol{y}$）连续随机向量 X 的概率密度函数（PDF）记为 $p_X(\boldsymbol{x})$。期望算子记为$\mathbb{E}[\cdot]$。我们用 $h(X) \equiv -\mathbb{E}[\text{对数 } p_X(X)]$ 表示 X的微分熵，用 $I(X;Y) \equiv h(Y) - h(Y|X)$表示 X 与 Y [1]之间的互信息。向量值可微函数 $\boldsymbol{f}(\boldsymbol{x})$ 关于 $\boldsymbol{x}$的雅可比矩阵记为 $D_{\boldsymbol{x}}\boldsymbol{f} \equiv \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$。

—

B. 得分函数

对数似然关于数据变量的梯度称为得分函数 [1]。对于随机向量$Y \in \mathbb{R}^d$，其概率密度函数为 $p_Y(\boldsymbol{y})$，其（边缘）得分函数定义为

$$\boldsymbol{s}_Y(\boldsymbol{y}) \equiv \nabla_{\boldsymbol{y}} \log p_Y(\boldsymbol{y}). \tag{1}$$

类似地，给定 X 时，Y 的条件得分函数为$\boldsymbol{s}_{Y|X}(\boldsymbol{y}|\boldsymbol{x}) \equiv \nabla_{\boldsymbol{y}} \log p_{Y|X}(\boldsymbol{y}|\boldsymbol{x})$。得分函数的一个基本性质是，只要概率密度函数在边界处趋于零，其在相应分布下的期望就为零：

$$\begin{aligned}\mathbb{E}[\boldsymbol{s}_Y(Y)] &= \int p_Y(\boldsymbol{y}) \nabla_{\boldsymbol{y}} \log p_Y(\boldsymbol{y}) d\boldsymbol{y} \\ &= \int \nabla_{\boldsymbol{y}} p_Y(\boldsymbol{y}) d\boldsymbol{y} = \boldsymbol{0}.\end{aligned} \tag{2}$$

这一性质对于简化梯度推导至关重要。

C. 相关工作

我们的框架（包括 [9]）建立在信息论与热力学之间的深刻对应关系之上，具体而言，是将噪声注入解释为扩散过程。这一观点源自 Stam [14]，他利用 de Bruijn 恒等式（该恒等式将熵的导数与费希尔信息联系起来）建立了熵功率不等式。在这一基于扩散的观点之上，Brown [15]和 Barron [16] 提出了熵中心极限定理，证明了费希尔信息的单调收敛性。在高斯信道的背景下，Guo、Shamai 和 Verd´u[17] 通过 I-MMSE 关系统一了这些概念，指出互信息的导数等于最小均方误差。

尽管这些经典研究主要将这些关系用于不等式的理论证明，我们的工作却将其重新用于实际计算。我们将这些基本恒等式的用途从理论分析拓展到数据驱动优化。通过将这些联系推广至任意随机有向无环图，并从数据中估计所需的得分函数，我们将这些经典的热力学联系转化为一个可扩展的端到端互信息最大化引擎。

D. 信息梯度恒等式

在我们的配套论文 [2],中，我们推导出了一个将互信息的梯度与得分函数联系起来的基本恒等式。考虑一个一般的参数化系统，其输出由 $Y = f_{\boldsymbol{\eta}}(X) + Z$给出，输入为 $X \in \mathbb{R}^n$，其中$f_{\boldsymbol{\eta}}$： $\mathbb{R}^n \to \mathbb{R}^m$ 是一个由$\boldsymbol{\eta} \in \mathbb{R}^d$参数化的可微函数，$Z \in \mathbb{R}^m$ 是一个加性噪声向量。在标准正则性条件下，互信息相对于 $\boldsymbol{\eta}$ 的梯度可表示为

$$\nabla_{\boldsymbol{\eta}} I(X;Y) = \mathbb{E}\left[(D_{\boldsymbol{\eta}} Y)^\top (\boldsymbol{s}_{Y|X}(Y|X) - \boldsymbol{s}_Y(Y))\right]. \tag{3}$$

该恒等式表明，信息梯度是输出关于参数的雅可比矩阵与“分数差”向量的期望内积。论文 [2] 着重于针对基本信道模型对这一恒等式的理论推导及其性质进行研究，而本文则将该公式推广到由一般有向无环图表示的复杂随机网络。

E. 去噪分数匹配

由于在复杂系统中真实的数据分布通常是未知的，我们必须从数据中估计得分函数。去噪得分匹配（DSM）[3]–[5] 是一种非常有效的技术。当数据 $\boldsymbol{y}$ 通过向干净样本 $\tilde{\boldsymbol{y}}$添加高斯噪声 $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0},\ \sigma^2 \boldsymbol{I})$生成时，即 $\boldsymbol{y} = \tilde{\boldsymbol{y}} + \boldsymbol{z}$，可以训练一个由参数向量 $\boldsymbol{\phi}$ 参数化的得分模型 $s_{\boldsymbol{\phi}}(\boldsymbol{y})$，使其最小化以下目标函数：

$$\mathcal{J}_{\text{DSM}}(\boldsymbol{\phi}) = \mathbb{E}_{\tilde{\boldsymbol{y}},\boldsymbol{z}}\left[\left\| \boldsymbol{s}_{\boldsymbol{\phi}}(\tilde{\boldsymbol{y}} + \boldsymbol{z}) + \frac{\boldsymbol{z}}{\sigma^2} \right\|^2\right]. \tag{4}$$

最小化此目标函数可确保 $\boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{y})$ 逼近真实得分 $\nabla_{\boldsymbol{y}} \log p_Y(\boldsymbol{y})$。我们利用这一技术学习梯度估计所需的得分函数。

尽管本文主要使用去噪分数匹配（DSM），但值得注意的是，也可以采用速度场框架（例如流匹配 [18] 和整流流 [19]），以数据驱动的方式推导得分函数。由于得分函数与速度场之间存在一一对应的代数对应关系，因此我们可以从速度场推导出得分函数。在实践中，训练速度场可能比其基于得分函数的对应方法稳定得多，尤其是在低噪声区域。在这种情况下，采用速度场框架估计得分函数可能更有优势。

III. 级联信道的信息梯度

将级联情形视为由有向无环图表示的更复杂系统的构建模块是有益的。在本节中，我们推导具有加性高斯噪声的确定性函数级联的信息梯度。

A. 问题设置

不失一般性，我们考虑一个由 N 个阶段组成的级联非线性高斯信道。令 $X \in \mathbb{R}^{d_0}$ 为随机输入源向量，其先验概率密度函数（PDF）为 p_X。该级联信道由一系列 N 个计算模块组成，其中第 i 个模块的输出记为 $Y_i \in \mathbb{R}^{d_i}$，并作为第（$i+1$）个模块的输入。

我们假设每个模块 $i \in \{1,\dots,N\}$ 由一个参数化确定性映射和一个加性高斯噪声信道组成。具体而言，相邻中间变量之间的关系为

$$Y_i = f_{\boldsymbol{\eta}_i}(Y_{i-1}) + Z_i, \quad i = 1,\dots,N, \tag{5}$$

其中，$Y_0 \triangleq X$ 是系统输入，$Y \equiv Y_N$ 表示级联信道的最终输出。这里，$f_{\boldsymbol{\eta}_i} : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$是由参数向量 $\boldsymbol{\eta}_i$参数化的确定性可微函数。假设噪声向量 $Z_i \in \mathbb{R}^{d_i}$ 由独立同分布（i.i.d.）的高斯随机变量构成，其均值为零，协方差矩阵为$\sigma_i^2 \boldsymbol{I}$ ($\sigma_i > 0$)，并且与 X 和 $\{Z_j\}_{j\neq i}$相互独立。图 1 展示了级联信道的示意图。

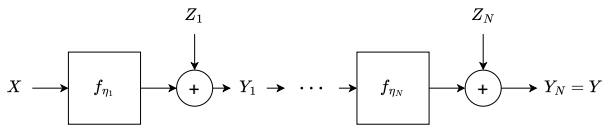

图 1. N级级联非线性高斯信道的示意图。每个阶段 i由确定性映射 $f_{\boldsymbol{\eta}_i}$和加性高斯噪声 Z_i 组成。$X \to Y_1 \to \cdots \to Y_N = Y$ 该系统构成一个马尔可夫链。

由于每个阶段都注入噪声，从X 到任意中间输出 Y_i（包括最终输出$Y = Y_N$）的映射都是随机的。整个级联构成一个马尔可夫链：

$$X \to Y_1 \to Y_2 \to \cdots \to Y_N = Y. \tag{6}$$

我们假设 P_Y 和 $P_{Y|X}$ 存在且可微。

我们的主要目标是通过优化参数集合 $\{\boldsymbol{\eta}_1,\dots,\boldsymbol{\eta}_N\}$，最大化初始输入 X 与最终输出 Y 之间的端到端互信息。优化问题定义为

$$\text{maximize}_{\{\boldsymbol{\eta}_i\}} \quad I(X;Y). \tag{7}$$

为了采用基于梯度的优化方法，我们需要计算该目标函数相对于任意中间参数 $\boldsymbol{\eta}_k$（$1 \le k \le N$）的梯度，即 $\nabla_{\boldsymbol{\eta}_k} I(X;Y)$。

B. 梯度公式的推导

现在，我们推导第 k 个块中关于参数向量 $\boldsymbol{\eta}_k$ 的互信息$I(X;Y)$ 梯度。根据互信息的定义，有

$$I(X;Y) = h(Y) - h(Y|X), \tag{8}$$

其中，$h(\cdot)$ 表示微分熵。因此，梯度可分解为两项：

$$\nabla_{\boldsymbol{\eta}_k} I(X;Y) = \nabla_{\boldsymbol{\eta}_k} h(Y) - \nabla_{\boldsymbol{\eta}_k} h(Y|X). \tag{9}$$

首先，考虑边缘熵$h(Y) = -\mathbb{E}[\log p_Y(Y)]$的梯度。需要注意的是，最终输出 Y 是一个随机变量，其取值通过级联中的确定性映射依赖于参数 $\boldsymbol{\eta}_k$。我们可以将 Y 视为输入X、所有噪声变量 $\{Z_i\}_{i=1}^N$以及参数 $\{\boldsymbol{\eta}_i\}_{i=1}^N$的可微函数。对期望应用链式法则（假设满足允许交换微分与期望运算的标准正则性条件），可得：

$$\begin{aligned}\nabla_{\boldsymbol{\eta}_k} h(Y) &= -\mathbb{E}\left[\nabla_{\boldsymbol{\eta}_k} \log p_Y(Y)\right] \\ &= -\mathbb{E}\left[(D_{\boldsymbol{\eta}_k} Y)^\top \nabla_{\boldsymbol{y}} \log p_Y(\boldsymbol{y})\big|_{\boldsymbol{y}=Y}\right] \\ &\quad - \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\eta}_k} \log p_Y(\boldsymbol{y};\boldsymbol{\eta}_k)\big|_{\boldsymbol{y}=Y}\right],\end{aligned} \tag{10}$$

其中，$D_{\boldsymbol{\eta}_k} Y \equiv \frac{\partial Y}{\partial \boldsymbol{\eta}_k}$ 是 Y相对于 $\boldsymbol{\eta}_k$的雅可比矩阵。右侧的第二项为零，因为

$$\mathbb{E}_Y\left[\nabla_{\boldsymbol{\eta}_k} \log p_Y(Y;\boldsymbol{\eta}_k)\right] = \int \nabla_{\boldsymbol{\eta}_k} p_Y(\boldsymbol{y};\boldsymbol{\eta}_k)\, d\boldsymbol{y} \tag{11}$$

$$= \nabla_{\boldsymbol{\eta}_k} \int p_Y(\boldsymbol{y})\, d\boldsymbol{y} \tag{12}$$

$$= \nabla_{\boldsymbol{\eta}_k} 1 = 0. \tag{13}$$

这个方程体现了一个众所周知的性质：边缘分布的得分函数的期望为零。引入边缘得分函数 $s_Y(\boldsymbol{y}) \equiv \nabla_{\boldsymbol{y}} \log p_Y(\boldsymbol{y})$ 后，我们得到

$$\nabla_{\boldsymbol{\eta}_k} h(Y) = -\mathbb{E}\left[(D_{\boldsymbol{\eta}_k} Y)^\top s_Y(Y)\right]. \tag{14}$$

类似地，对于条件熵

$$h(Y|X) = -\mathbb{E}[\log p_{Y|X}(Y|X)], \tag{15}$$

我们可以通过对 X进行条件化来推导其梯度。令

$$s_{Y|X}(\boldsymbol{y}|\boldsymbol{x}) \equiv \nabla_{\boldsymbol{y}} \log p_{Y|X}(\boldsymbol{y}|\boldsymbol{x}) \tag{16}$$

为条件得分函数。按照上述相同的推理，我们得到：

$$\nabla_{\boldsymbol{\eta}_k} h(Y|X) = -\mathbb{E}\left[(D_{\boldsymbol{\eta}_k} Y)^\top s_{Y|X}(Y|X)\right]. \tag{17}$$

将 (14) 和 (17) 代入 (9)，可得以下关键结果。

定理 1（级联信道的信息梯度）。端到端互信息 $I(X;Y)$ 关于第 k 个（$k \in \{1,\dots,N\}$）块的参数 $\boldsymbol{\eta}_k$ 的梯度为

$$\nabla_{\boldsymbol{\eta}_k} I(X;Y) = \mathbb{E}\left[(D_{\boldsymbol{\eta}_k} Y)^\top (s_{Y|X}(Y|X) - s_Y(Y))\right]. \tag{18}$$

这个公式有一个直观的解释：梯度是雅可比矩阵 $D_{\boldsymbol{\eta}_k} Y$ 与一个量的乘积的期望；雅可比矩阵表示参数 $\boldsymbol{\eta}_k$ 如何影响最终输出，而这个量是

“分数差”项（$s_{Y_N|X} - s_{Y_N}$），它表示信息论学习信号。

注 1 (后验分数表示**)**。 一个关键观察是，分数差项可以等价地用后验分布表示：

$$s_{Y|X}(\boldsymbol{y}|\boldsymbol{x}) - s_Y(\boldsymbol{y}) = \nabla_{\boldsymbol{y}} \log p_{X|Y}(\boldsymbol{x}|\boldsymbol{y}). \tag{19}$$

根据贝叶斯公式，$\log p(\boldsymbol{y}|\boldsymbol{x})$ –$\log\ p(\boldsymbol{y})$ $=\log\ p(\boldsymbol{x}|\boldsymbol{y})-\log\ p(\boldsymbol{x})$，对 $\boldsymbol{y}$ 求梯度（注意到 $\nabla_{\boldsymbol{y}} \log p(\boldsymbol{x}) = 0$），即可轻易得到这一恒等式。因此，信息梯度可以重新表示为

$$\nabla_{\boldsymbol{\eta}_k} I(X;Y) = \mathbb{E}\left[(D_{\boldsymbol{\eta}_k} Y)^\top \nabla_{\boldsymbol{y}} \log p_{X|Y}(X|\boldsymbol{y})\big|_{\boldsymbol{y}=Y}\right]. \tag{20}$$

这种“后验分数”形式对于采用近似推断模型的实现尤其有用。

注 2 (化简为单阶段梯度**)**。 考察最简单的情形——单阶段信道（$N = 1$）——颇具启发性，该信道由 $Y_1 = f_{\boldsymbol{\eta}_1}(X) + Z_1$ 给出。在这种情况下，条件熵 $h(Y_1|X) = h(Z_1)$ 相对于 $\boldsymbol{\eta}_1$ 为常数，这意味着 $\nabla_{\boldsymbol{\eta}_1} h(Y_1|X) = 0$。因此，式（18）中的条件得分项在期望下消失

$$\mathbb{E}[(D_{\boldsymbol{\eta}_1} Y_1)^\top s_{Y_1|X}(Y_1|X)] = 0. \tag{21}$$

定理 1 因而可化简为

$$\nabla_{\boldsymbol{\eta}_1} I(X;Y_1) = -\mathbb{E}\left[(D_{\boldsymbol{\eta}_1} Y_1)^\top s_{Y_1}(Y_1)\right], \tag{22}$$

这恰好恢复了 [2] 推导的标准非线性高斯信道的信息梯度公式。我们的广义定理表明，对于更深层级联（$N > 1$），额外项 $s_{Y_N|X}$ 是必不可少的，因为中间的噪声注入会使 $h(Y_N|X)$ 依赖于较早的参数。

C. 高效梯度计算

定理 1 中推导的公式涉及雅可比矩阵 $D_{\boldsymbol{\eta}_k} Y$，若显式构造该矩阵，计算开销可能很大，尤其是在高维系统中。然而，利用级联结构和现代自动微分（AD）技术，我们可以高效地计算该梯度，而无需构造完整的雅可比矩阵。

1) 级联雅可比的链式法则：由于级联具有马尔可夫结构，雅可比矩阵 $D_{\boldsymbol{\eta}_k} Y$ 可以利用链式法则进行分解。回顾一下，Y_N 仅通过中间序列 Y_k、Y_{k+1}、……、Y_{N-1} 依赖于 $\boldsymbol{\eta}_k$。该雅可比矩阵可以显式写成一系列纯确定性映射的雅可比矩阵之积：

$$D_{\boldsymbol{\eta}_k} Y = \left(\prod_{i=N}^{k+1} \frac{\partial f_{\boldsymbol{\eta}_i}}{\partial Y_{i-1}}(Y_{i-1})\right) \cdot \frac{\partial f_{\boldsymbol{\eta}_k}}{\partial \boldsymbol{\eta}_k}(Y_{k-1}). \tag{23}$$

这里，$\frac{\partial f_{\boldsymbol{\eta}_i}}{\partial Y_{i-1}}$ 是第 i 个块的函数相对于其输入的雅可比矩阵，而 $\frac{\partial f_{\boldsymbol{\eta}_k}}{\partial \boldsymbol{\eta}_k}$ 是相对于其参数的雅可比矩阵。这种乘积结构与用于训练深度神经网络的标准反向传播算法中的反向传播过程完美匹配。

2) 向量-雅可比积（*VJP*）：更重要的是，梯度公式（18）不需要完整的雅可比矩阵本身，而只需要其转置与一个向量的乘积。我们将信息分数向量 $\boldsymbol{v} \in \mathbb{R}^{d_N}$ 定义为

$$\boldsymbol{v}(\boldsymbol{x},\boldsymbol{y}) \equiv s_{Y|X}(\boldsymbol{y}|\boldsymbol{x}) - s_Y(\boldsymbol{y}). \tag{24}$$

期望内部的梯度项因此为 $(D_{\boldsymbol{\eta}_k} Y)^\top \boldsymbol{v}$。这一运算称为向量-雅可比积（VJP）[6]。在标准自动微分框架（如 PyTorch、TensorFlow 或 JAX）中，将 $\boldsymbol{v}$ 作为输出层的初始梯度（或“伴随量”），即可通过一次反向传播高效地计算 VJP。

在实践中，只需定义一个标量替代损失函数，即可轻松计算 VJP。令 stop(·) 表示停止梯度算子；在自动微分框架中，它在前向传播时表现为恒等函数，而在反向传播时梯度为零。我们将 VJP 损失定义为

$$\mathcal{L}_{\text{VJP}}(\boldsymbol{\eta}) \equiv -\langle \text{stop}(\boldsymbol{v}), Y \rangle, \tag{25}$$

其中，$\langle \cdot,\ \cdot \rangle$ 表示内积。对该损失应用标准自动微分，可得

$$\nabla_{\boldsymbol{\eta}_k} \mathcal{L}_{\text{VJP}} = -(D_{\boldsymbol{\eta}_k} Y)^\top \boldsymbol{v}. \tag{26}$$

该项的期望恰好恢复了式（18）中推导出的信息梯度的负值。在实际实现中，我们可以通过对小批量中的$\nabla_{\boldsymbol{\eta}_k} \mathcal{L}_{\text{VJP}}$ 取样本均值，获得真实信息梯度的蒙特卡洛估计量。对于由 B 个样本组成的小批量{（$\boldsymbol{x}^{(b)}$，$\boldsymbol{y}^{(b)}$）而言，信息梯度的蒙特卡洛估计量为

$$\widehat{\nabla_{\boldsymbol{\eta}_k} I} = -\frac{1}{B} \sum_{b=1}^{B} \nabla_{\boldsymbol{\eta}_k} \mathcal{L}_{\text{VJP}}(\boldsymbol{\eta}_k; \boldsymbol{x}^{(b)}, \boldsymbol{y}^{(b)}). \tag{27}$$

该估计器允许通过标准随机梯度下降（SGD）最大化互信息。

D. 分数模型近似

信息梯度公式（18）依赖于真实的边缘分数 s_Y（$\boldsymbol{y}$）和条件得分函数$s_{Y|X}$（$\boldsymbol{y}|\boldsymbol{x}$），而对于复杂的非线性级联，它们通常是未知的。为了使所提出的框架具有实用性，我们采用参数化的得分模型（通常为深度神经网络）来近似这些函数。

令 $s_{\boldsymbol{\phi}}$（$\boldsymbol{y};\boldsymbol{\eta}$）和 $s_{\boldsymbol{\psi}}$（$\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\eta}$）分别为由 $\boldsymbol{\phi}$ 和 $\boldsymbol{\psi}$参数化的得分模型，用于近似真实的分数：

$$s_{\boldsymbol{\phi}}(\boldsymbol{y};\boldsymbol{\eta}) \approx \nabla_{\boldsymbol{y}} \log p_Y(\boldsymbol{y};\boldsymbol{\eta}), \tag{28}$$

$$s_{\boldsymbol{\psi}}(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\eta}) \approx \nabla_{\boldsymbol{y}} \log p_{Y|X}(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\eta}). \tag{29}$$

请注意，这些模型隐式地依赖于级联参数 $\boldsymbol{\eta} = \{\boldsymbol{\eta}_i\}_{i=1}^N$，因为 Y 的分布会随 $\boldsymbol{\eta}$ 变化。关键的是，这些评分模型可以通过样本高效地训练

，采用 DSM [3]–[5] 或其变体，无需了解底层似然。对于最终阶段采用标准高斯噪声信道的情形，DSM 提供了一个简单且易于处理的目标函数。

用学习得到的分数替代真实分数，可得到针对样本 $(\boldsymbol{x}, \boldsymbol{y})$ 的实用替代向量-雅可比积损失：

$$\mathcal{L}_{\text{VJP}}(\boldsymbol{\eta}_k; \boldsymbol{x}, \boldsymbol{y}) \equiv -\langle \text{stop}(s_{\boldsymbol{\psi}}(\boldsymbol{y}|\boldsymbol{x}) - s_{\boldsymbol{\phi}}(\boldsymbol{y})), \boldsymbol{y} \rangle. \tag{30}$$

利用由该替代损失推导出的梯度优化级联参数 $\boldsymbol{\eta}$，同时（或交替地）训练评分模型 $\boldsymbol{\phi}$, $\boldsymbol{\psi}$ 以跟踪不断变化的分布，这构成了我们提出的端到端优化框架的核心。

E. 解析示例：线性高斯级联信道

为验证所提出的信息梯度公式（定理 1），我们考虑一个简单的两阶段线性高斯级联，其精确梯度可以通过解析方法求得。设输入为 $X \sim \mathcal{N}(0, \sigma_X^2)$。考虑一个包含标量变量的两阶段级联 $(N = 2)$：

$$Y_1 = \eta_1 X + Z_1, \quad Z_1 \sim \mathcal{N}(0, \sigma_1^2), \tag{31}$$

$$Y_2 = Y_1 + Z_2, \qquad Z_2 \sim \mathcal{N}(0, \sigma_2^2), \tag{32}$$

其中，η_1 是第一阶段中关注的标量参数。总有效噪声为 $Z_{eff} = Z_1 + Z_2 \sim \mathcal{N}(0, \sigma_N^2)$，其中$\sigma_N^2 = \sigma_1^2 + \sigma_2^2$。最终输出为 $Y_2 = \eta_1 X + Z_{eff}$，其方差为 $v_{Y_2} \equiv \text{Var}(Y_2) = \eta_1^2 \sigma_X^2 + \sigma_N^2$。

这个标量高斯信道的互信息已知为

$$I(X; Y_2) = \frac{1}{2} \log \left(1 + \frac{\eta_1^2 \sigma_X^2}{\sigma_N^2} \right) = \frac{1}{2} \log \left(\frac{v_{Y_2}}{\sigma_N^2} \right). \tag{33}$$

直接对 η_1 求导，可得真实梯度：

$$\frac{\partial I(X; Y_2)}{\partial \eta_1} = \frac{1}{2} \frac{\sigma_N^2}{v_{Y_2}} \cdot \frac{2\eta_1 \sigma_X^2}{\sigma_N^2} = \frac{\eta_1 \sigma_X^2}{v_{Y_2}}. \tag{34}$$

现在应用定理 1。最终输出Y_2 关于 η_1 的雅可比矩阵为 $D_{\eta_1} Y_2 = \frac{\partial}{\partial \eta_1}(\eta_1 X + Z_1 + Z_2) = X$。边缘分数为 $s_{Y_2}(y) = -y / v_{Y_2}$，条件得分函数为 $s_{Y_2|X}(y|x) = -(y - \eta_1 x)/\sigma_N^2$。将它们代入式 (18)：

$$\begin{aligned} \nabla_{\eta_1} I &= \mathbb{E}\left[X \cdot \left(-\frac{Y_2 - \eta_1 X}{\sigma_N^2} - \left(-\frac{Y_2}{v_{Y_2}} \right) \right) \right] \\ &= -\frac{\mathbb{E}[XY_2] - \eta_1 \mathbb{E}[X^2]}{\sigma_N^2} + \frac{\mathbb{E}[XY_2]}{v_{Y_2}}. \end{aligned} \tag{35}$$

利用 $\mathbb{E}[X^2] = \sigma_X^2$ 和 $\mathbb{E}[XY_2] = \mathbb{E}[X(\eta_1 X + Z_{eff})] = \eta_1 \sigma_X^2$，第一项为零：

$$-\frac{\eta_1 \sigma_X^2 - \eta_1 \sigma_X^2}{\sigma_N^2} + \frac{\eta_1 \sigma_X^2}{v_{Y_2}} = \frac{\eta_1 \sigma_X^2}{v_{Y_2}}. \tag{36}$$

这与文献（34）中的经典结果完全一致，验证了我们推广后的信息梯度公式的有效性。

IV. 推广至有向无环图

在确立了级联信道的信息梯度后，我们现在将这一框架扩展到由有向无环图（DAG）表示的一般随机网络。这种泛化使我们能够优化涉及分支、合并和并行处理路径的复杂系统。

A. 问题设置

考虑一个由有向无环图（DAG） $\mathcal{G} = (\mathcal{V}, \mathcal{E})$表示的通信或感知系统，该图包含 M个节点。令 V_1, ..., V_M 表示按照拓扑序排列的各节点所关联的随机向量。网络输入由根节点的一个子集表示，统称为X。随后，每个节点 V_j 通过参数化确定性映射，根据其父节点 $\text{Pa}(V_j)$计算得到，并可选择性地受到独立加性噪声的干扰。具体而言，节点 V_j 处的局部计算为

$$V_j = f_j(\text{Pa}(V_j); \boldsymbol{\eta}_j) + Z_j, \tag{37}$$

其中，$f_j : \mathbb{R}^{d_j^{\text{in}}} \to \mathbb{R}^{d_j^{\text{out}}}$ 是一个由 $\boldsymbol{\eta}_j$参数化的可微函数，Z_j 是零均值的独立高斯噪声向量（如果节点 V_j 是确定性的，则 $Z_j \equiv 0$）。所有网络参数的集合为 $\boldsymbol{\eta} \equiv \{\boldsymbol{\eta}_j\}_j$，所有噪声源的集合为 $\mathcal{Z} \equiv \{Z_j\}_j$。

系统的最终输出记为 Y，是叶节点的一个子集。得益于 $\mathcal{G}$的无环结构，在（37）中递归替换父节点必然会终止。因此，从根节点 X 到输出 Y 的端到端关系可以明确表示为输入 X 与所有相关噪声源 $\mathcal{Z}$的确定性复合：

$$Y = g(X, \mathcal{Z}; \boldsymbol{\eta}). \tag{38}$$

图 2 展示了有向无环图网络的示意图。为确保所需得分函数的存在，我们假设向网络中注入了足够的噪声（通常在最终输出节点处），使得 Y 具有可微的概率密度函数。

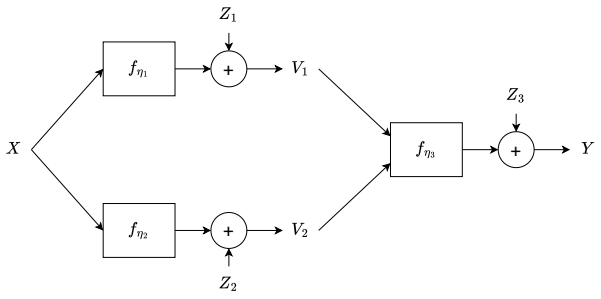

图 2. 有向无环图网络示意图。输入 X 分成两条并行路径，分别由不同的函数 $f_{\boldsymbol{\eta}_i}$ 处理，并受到噪声 Z_i 的干扰，从而生成中间节点 V_1, V_2。随后，这些节点由函数 $f_{\boldsymbol{\eta}_3}$ 合并，并进一步受到噪声 Z_3 的干扰，最终得到最终输出 Y。

B. DAG 信道的信息梯度公式

现在，我们推导互信息$I(X;Y)$关于有向无环图中任意参数向量 $\boldsymbol{\eta}_k \in \boldsymbol{\eta}$ 的梯度。值得注意的是，梯度公式保留了与级联情形相同的优雅结构，展现了其普适性。

定理 2（有向无环图的信息梯度）**：** 在标准正则性条件下，端到端互信息 $I(X;\ Y)$关于有向无环图中任意内部参数 $\boldsymbol{\eta}_k$ 的梯度为

$$\nabla_{\boldsymbol{\eta}_k} I(X;Y) = \mathbb{E}\left[(D_{\boldsymbol{\eta}_k} Y)^\top (s_{Y|X}(Y|X) - s_Y(Y))\right], \quad (39)$$

其中 $D_{\boldsymbol{\eta}_k} Y \equiv \frac{\partial g}{\partial \boldsymbol{\eta}_k}(X,\ \mathcal{Z};\ \boldsymbol{\eta})$ 是最终输出关于 $\boldsymbol{\eta}_k$ 的雅可比矩阵，期望取遍网络中的所有随机变量（X、$\mathcal{Z}$）。

证明：互信息为 $I(X;\ Y) = h(Y) - h(Y|X)$。我们首先分析边缘熵 $h(Y) = -\mathbb{E}[\log p_Y(Y)]$ 的梯度。与级联情形不同，现在的期望取自有向无环图中所有输入和噪声源的联合分布。令 $\boldsymbol{\omega} = (X,\ \mathcal{Z})$ 表示网络中所有原始随机变量的集合。输出 Y 是一个确定性函数 $Y = g(\boldsymbol{\omega};\ \boldsymbol{\eta})$。假设满足允许交换 $\nabla_{\boldsymbol{\eta}_k}$ 和 $\mathbb{E}_{\boldsymbol{\omega}}$ 的正则性条件，则有：

$$\begin{aligned}\nabla_{\boldsymbol{\eta}_k} h(Y) &= -\mathbb{E}_{\boldsymbol{\omega}}\left[\nabla_{\boldsymbol{\eta}_k} \log p_Y(g(\boldsymbol{\omega};\boldsymbol{\eta});\boldsymbol{\eta})\right] \quad (40)\\ &= -\mathbb{E}_{\boldsymbol{\omega}}\left[(D_{\boldsymbol{\eta}_k} Y)^\top \nabla_{\boldsymbol{y}} \log p_Y(\boldsymbol{y})\big|_{\boldsymbol{y}=Y}\right]\\ &\quad - \mathbb{E}_{\boldsymbol{\omega}}\left[\frac{\partial}{\partial \boldsymbol{\eta}_k} \log p_Y(\boldsymbol{y};\boldsymbol{\eta})\big|_{\boldsymbol{y}=Y}\right]. \quad (41)\end{aligned}$$

第二项为 $\mathbb{E}_Y[\nabla_{\boldsymbol{\eta}_k} \log p_Y(Y;\boldsymbol{\eta})]$,，由于得分函数性质 $\int \nabla p = \nabla \int p = 0$，该项为零。因此，$\nabla_{\boldsymbol{\eta}_k} h(Y) = -\mathbb{E}\left[(D_{\boldsymbol{\eta}_k} Y)^\top s_Y(Y)\right]$。对条件熵项采用完全类似的论证（始终以 X 为条件），可得 $\nabla_{\boldsymbol{\eta}_k} h(Y|X) = -\mathbb{E}\left[(D_{\boldsymbol{\eta}_k} Y)^\top s_{Y|X}(Y|X)\right]$。结合这些结果即可得到定理的结论。■

有向无环图中的雅可比矩阵 $D_{\boldsymbol{\eta}_k} Y$ 可通过沿图执行标准的反向模式自动微分（反向传播）高效计算，这与现代深度学习框架中的做法完全相同。这意味着第 III-C 节所述的基于向量-雅可比积的高效计算策略可直接应用于一般有向无环图，无需修改。

注 3（无环性的数学重要性）。假设图 $\mathcal{G}$ 是无环的（有向无环图）对于我们的框架至关重要，原因有以下两点：

1) **定义良好的复合：** 不存在环可保证递归计算在有限步内终止。这确保端到端映射$Y = g(X,\ \mathcal{Z};\boldsymbol{\eta})$ 是定义良好的复合函数。如果存在环，系统将具有自指性，需要采用不动点表述。2) **链式法则的适用性：** 由于 Y 是一个标准复合函数，其雅可比矩阵 $D_{\boldsymbol{\eta}_k} Y$ 在（局部可微性假设下）理论上保证存在，并且可以通过标准多元链式法则进行计算。这一理论保证使得使用反向传播（向量-雅可比积）进行高效计算具有充分依据。

注 4（重参数化与可微性）。关于随机的输出 Y 相对于参数 $\boldsymbol{\eta}$的可微性，自然会产生一个问题。我们的框架依赖于重参数化技巧 [20]（也称为路径导数估计量）。正如公式 (38) 中明确表述的那样，我们不将系统输出视为一个普通的随机变量，而是将其视为参数和固定独立噪声源 $\mathcal{Z}$的确定性函数 $Y = g(X,\ \mathcal{Z};\ \eta)$。由于随机性被外化到不依赖于 $\boldsymbol{\eta}$的 $\mathcal{Z}$ 中，因此对于 $(X,\ \mathcal{Z})$ 的任意固定实现，雅可比矩阵 $D_\eta Y$ 都是定义良好的。这一视角有效地将随机性与参数依赖性解耦，使梯度能够通过标准反向传播在随机节点中“流动”。

注释 5（噪声假设的普遍性）。虽然本文主要假设噪声为高斯噪声，以便高效地进行分数匹配（去噪分数匹配），但推导得到的信息梯度公式理论上适用于任何具有可微概率密度函数的噪声分布。此外，无需在每个节点注入噪声。只要有足够的噪声进入有向无环图（通常在最终输出处注入，或沿路径累积），以确保输出 Y 具有光滑密度，中间变换就可以是纯确定性的（$Z_j \equiv 0$）。这种灵活性使我们的框架能够对随机与确定性组件混合的系统进行建模，例如将数字逻辑与模拟传输相结合的系统。

C. 解析示例：多路径有向无环图

为进一步验证有向无环图的广义信息梯度（定理 2），我们考虑一个多路径网络，其中参数通过不同的拓扑结构影响输出。考虑图 2 所示的有向无环图，其线性高斯设定如下：

$$V_1 = \eta_1 X + Z_1, \quad Z_1 \sim \mathcal{N}(0, \sigma_1^2) \quad (42)$$

$$V_2 = X + Z_2, \quad Z_2 \sim \mathcal{N}(0, \sigma_2^2) \quad (43)$$

$$Y = \eta_2 V_1 + V_2 + Z_3, \quad Z_3 \sim \mathcal{N}(0, \sigma_3^2) \quad (44)$$

其中 $X \sim \mathcal{N}(0,\ \sigma_X^2)$。这里，$\eta_1$ 是一个仅影响一条路径的深层参数，而 η_2 是合并节点处的混合参数。关键的是，η_2 同时影响信号增益和有效噪声方差。端到端关系为 $Y = G_{\text{eff}} X + Z_{\text{eff}}$，其中 $G_{\text{eff}} = \eta_1\eta_2 + 1$ 是有效增益，$Z_{\text{eff}} = \eta_2 Z_1 + Z_2 + Z_3$ 是方差为 $\sigma_N^2 = \eta_2^2\sigma_1^2 + \sigma_2^2 + \sigma_3^2$的有效噪声。输出方差为 $v_Y = G_{\text{eff}}^2 \sigma_X^2 + \sigma_N^2$。

互信息可简单表示为 $I(X;Y) = \frac{1}{2}\log(v_Y/\sigma_N^2)$。直接求导可得真实的解析

梯度：

$$\frac{\partial I}{\partial \eta_1} = \frac{\eta_2 G_{\text{eff}} \sigma_X^2}{v_Y} \tag{45}$$

$$\frac{\partial I}{\partial \eta_2} = \frac{\eta_1 G_{\text{eff}} \sigma_X^2 + \eta_2 \sigma_1^2}{v_Y} - \frac{\eta_2 \sigma_1^2}{\sigma_N^2} \tag{46}$$

现在应用定理 2。雅可比矩阵 $D_{\eta_k} Y$ 通过对有向无环图应用链式法则推导得出：

$$D_{\eta_1} Y = \eta_2 X, \quad D_{\eta_2} Y = \underset{s_Y}{V_1} = \underset{y}{\eta_1 X + Z_1}. \qquad \underset{s_{Y|X}}{=} \tag{47}$$

使用高斯得分函数 $_$

$() = -y/v_Y$和 $(y|x)\,(y - G_{\text{eff}} x)/\sigma_N^2 = -Z_{\text{eff}}/\sigma_N^2$，我们验证该定理：对于$\eta_1$，由于 $\mathbb{E}[X Z_{\text{eff}}] = 0$：

$$\nabla_{\eta_1} I = \mathbb{E}\left[\eta_2 X \left(s_{Y|X} - s_Y\right)\right] = \eta_2 \mathbb{E}\left[X \frac{Y}{v_Y}\right] = \frac{\eta_2 G_{\text{eff}} \sigma_X^2}{v_Y}. \tag{48}$$

这与式 (45) 完全一致。对于 η_2，雅可比矩阵 V_1 通过 Z_1与 Z_{eff} 相关联：

$$\begin{aligned} \nabla_{\eta_2} I &= \mathbb{E}\left[V_1 \left(-\frac{Z_{\text{eff}}}{\sigma_N^2} + \frac{Y}{v_Y}\right)\right] \\ &= -\frac{\mathbb{E}[(\eta_1 X + Z_1)(\eta_2 Z_1 + \dots)]}{\sigma_N^2} + \frac{\mathbb{E}[V_1 Y]}{v_Y} \\ &= -\frac{\eta_2 \sigma_1^2}{\sigma_N^2} + \frac{\eta_1 G_{\text{eff}} \sigma_X^2 + \eta_2 \sigma_1^2}{v_Y}. \end{aligned} \tag{49}$$

这完美地还原了式 (46)。这证实了我们的方法能够正确处理复杂依赖关系，其中参数会同时影响信号路径和噪声特性。

D. 信息梯度估计算法

信息梯度公式 (39) 使我们能够使用蒙特卡洛方法估计有向无环图中任意参数的信息梯度。算法 1 概述了针对目标参数 $\boldsymbol{\eta}_k$，获取梯度 $\nabla_{\boldsymbol{\eta}_k} I(X;\, Y)$ 小批量估计的过程。我们假设所需的得分函数模型 $s_{\boldsymbol{\phi}}(\boldsymbol{y}) \approx \nabla_{\boldsymbol{y}} \log p_Y(\boldsymbol{y})$ 和 $s_{\boldsymbol{\psi}}(\boldsymbol{y}|\boldsymbol{x}) \approx \nabla_{\boldsymbol{y}} \log p_{Y|X}(\boldsymbol{y}|\boldsymbol{x})$ 已预先训练，或正在同步训练。

这一估计过程能够在一次反向传播中，高效地同时计算有向无环图中所有参数的梯度，而不仅仅是某个特定 $\boldsymbol{\eta}_k$的梯度。

算法 1 通过向量-雅可比积进行信息梯度估计 要求：有向无环图 $\mathcal{G}$、目标参数 $\boldsymbol{\eta}_k$、学习得到的评分模型 $s_{\boldsymbol{\phi}}$、 $s_{\boldsymbol{\psi}}$、输入的小批量 $\{\boldsymbol{x}^{(i)}\}_{i=1}^{B}$ 确保：梯度估计 $\nabla_{\boldsymbol{\eta}_k} I$。1：// 1. 前向传播 2：对 $i = 1$ 到 B 执行 3：从输入 $\boldsymbol{x}^{(i)}$ 开始，在有向无环图上运行前向采样，以获得输出 $\boldsymbol{y}^{(i)}$ 和所有中间节点值。4：结束循环 5：// 2. 得分差计算 6：对 $i = 1$ 到 B 执行 7：计算边缘分数：$\boldsymbol{s}_{\text{marg}}^{(i)} \leftarrow s_{\boldsymbol{\phi}}(\boldsymbol{y}^{(i)})$ 8：计算条件得分函数：$\boldsymbol{s}^{(i)}$ cond $\leftarrow s_{\boldsymbol{\psi}}(\boldsymbol{y}^{(i)}|\boldsymbol{x}^{(i)})$ 9：计算得分差向量：$\boldsymbol{v}^{(i)} \leftarrow \boldsymbol{s}^{(i)}$ cond $- \boldsymbol{s}_{\text{marg}}^{(i)}$ 10：结束循环 11：// 3. 反向传播（向量-雅可比积）12：定义标量向量-雅可比积损失：$\mathcal{L} \leftarrow -\frac{1}{B}\sum_{i=1}^{B} \langle \text{stop}(\boldsymbol{v}^{(i)}) \boldsymbol{y}^{(i)} \rangle$13：从 $\mathcal{L}$ 开始，通过有向无环图运行反向传播。14：提取目标参数的梯度：$\nabla_{\boldsymbol{\eta}_k} I \leftarrow -\nabla_{\boldsymbol{\eta}_k} \mathcal{L}$

E. 互信息的积分表示

1）路径积分表示：虽然我们的主要关注点是基于梯度的优化，但所提出的框架也能估计互信息值$I(X;Y)$本身。利用线积分基本定理，我们可以根据互信息的梯度重建互信息：

$$I(X;Y;\boldsymbol{\eta}_1) = I(X;Y;\boldsymbol{\eta}_0) + \int_0^1 \langle \nabla_{\boldsymbol{\eta}} I(\boldsymbol{\eta}(t)), \dot{\boldsymbol{\eta}}(t) \rangle dt, \tag{50}$$

其中，$\boldsymbol{\eta}(t)$ 是参数空间中的任意光滑路径，连接参考配置 $\boldsymbol{\eta}_0 = \boldsymbol{\eta}(0)$ 与目标配置 $\boldsymbol{\eta}_1 = \boldsymbol{\eta}(1)$。在许多实际场景中，我们可以选择一个互信息已知的参考配置 $\boldsymbol{\eta}_0$ 设其值为零（例如，将参数设为零，使信号完全阻断）。通过沿着这样的路径对算法 1 得到的梯度估计进行数值积分（例如，使用梯形法则），我们可以估计任意一般有向无环图的端到端互信息绝对值。

2）费希尔积分表示：利用微分熵与费希尔信息之间的联系（即 de Bruijn恒等式），可以直接估计互信息$I(X;Y)$。设 $Y_0 \equiv Y$ 为有向无环图的输出，并考虑其受噪声扰动的版本$Y_t = Y_0 + \sqrt{t} Z'$，其中$Z' \sim \mathcal{N}(\mathbf{0}, I)$是一个独立的标准高斯噪声向量，$t \geq 0$ 是添加的噪声方差。回顾一下，互信息等于微分熵之差：$I(X;Y_t) = h(Y_t) - h(Y_t|X)$。de Bruijn恒等式指出，微分熵关于噪声方差 t 的导数与费希尔信息成正比：

$$\frac{d}{dt} h(Y_t) = \frac{1}{2} J(Y_t), \quad \frac{d}{dt} h(Y_t|X) = \frac{1}{2} J(Y_t|X), \tag{51}$$

其中，$J(Y_t) \equiv \mathbb{E}[\|\nabla_{\boldsymbol{y}_t} \log p(Y_t)\|^2]$ 和 $J(Y_t|X) \equiv \mathbb{E}[\|\nabla_{\boldsymbol{y}_t} \log p(Y_t|X)\|^2]$ 分别是无条件费希尔信息和条件费希尔信息。因此，互信息的导数为：

$$\frac{d}{dt} I(X;Y_t) = \frac{1}{2}\left(J(Y_t) - J(Y_t|X)\right). \tag{52}$$

利用条件分数的条件期望为边缘分数这一性质，即 $\mathbb{E}[s_{Y_t|X}(Y_t|X)|Y_t] = s_{Y_t}(Y_t)$，我们得到如下恒等式

$$J(Y_t|X) - J(Y_t) = \mathbb{E}\left[\|s_{Y_t|X}(Y_t|X) - s_{Y_t}(Y_t)\|^2\right]. \tag{53}$$

将其代入（52），并从 $t=0$ 积分到 ∞（注意到 $\lim_{t\to\infty} I(X;Y_t)=0$），可得以下积分表示：

$$I(X;Y)=\frac{1}{2}\int_0^\infty [J(Y_t|X)-J(Y_t)]dt \tag{54}$$

$$=\frac{1}{2}\int_0^\infty \mathbb{E}\left[\|s_{Y_t|X}(Y_t|X)-s_{Y_t}(Y_t)\|^2\right]dt. \tag{55}$$

该公式通过针对不同噪声水平 t 训练评分模型并执行数值积分，实现对 $I(X;Y)$ 的估计。值得注意的是，该方法不需要像路径积分方法那样，预先设置具有已知互信息的参考参数配置 $\boldsymbol{\eta}_0$。这一积分表示从根本上建立在 [9] 中提出的分数到费希尔桥（SFB）框架之上。

F. 与既有工作的关系

大量传统互信息估计和表示学习工作依赖变分界或对比界（例如 InfoNCE、CLUB、MINE），这些方法会优化判别器或密度比模型，因此也继承了评论器训练所带来的偏差—方差权衡和调参问题。基于核密度估计（KDE）[7]或 k 近邻（kNN）[8]的非参数方法也广泛用于互信息估计。然而，这些方法往往会受到维数灾难的影响，并且并不天然适用于基于梯度的优化，因为相对于系统参数对这些估计器进行微分在计算上代价过高。

相比之下，我们的方法不使用变分下界，而是将精确信息梯度表示为分数差与一个 VJP 的乘积，使其可直接用于自动微分，并避免使用对抗式/对比目标。经典恒等式（I-MMSE [17]，de Bruijn恒等式 [14]）将 $I(X;Y)$ 的导数与加性高斯场景下的均方误差或费希尔信息联系起来；我们以这些见解为基础，通过统一的梯度公式将其扩展到一般非线性有向无环图和任意内部参数，该公式可通过 VJP 计算。当无法获得闭式分数时，我们利用 DSM（及条件/后验变体）学习所需的分数，这与似然比估计器或基于评论器的估计器形成对比。对于互信息值（而不仅是梯度）的估计，我们的路径积分表示和费希尔积分表示提供了密度比估计器和变分估计器之外的另一种选择，使我们能够对易于估计的梯度/费希尔项进行积分。

V. 全局约束下的互信息最大化

所推导出的信息梯度能够在各种实际约束下优化有向无环图网络。在许多工程场景中，例如功率受限传感器网络或带宽受限通信系统，网络必须在全局资源约束下进行优化，而不是受单独的局部约束限制。

A. 问题表述

我们考虑在网络参数 $\boldsymbol{\eta}$ 上满足全局约束的条件下，最大化端到端互信息的问题。设 $C(\boldsymbol{\eta})\colon\ \mathbb{R}^{\dim(\boldsymbol{\eta})}\to\mathbb{R}$ 是表示网络总资源使用量的可微凸成本函数（例如，由 L_2 范数 $C(\boldsymbol{\eta})=\|\boldsymbol{\eta}\|_2^2$ 定义的总功耗）。该优化问题表述为

$$\text{maximize}_{\boldsymbol{\eta}}\ I(X;Y)\ \text{subject to}\ C(\boldsymbol{\eta})\le P, \tag{56}$$

其中，$P>0$ 是最大允许资源预算。

B. 投影梯度上升

为求解 (56)，我们可以采用投影梯度上升（PGA）方法。PGA算法是一种标准的迭代优化算法，适用于投影到可行集 $\mathcal{S}=\{\boldsymbol{\eta}\mid C(\boldsymbol{\eta})\le P\}$ 在计算上可行的约束问题。

在迭代 t 时，更新规则包含两个步骤：使用估计的信息梯度进行梯度上升，然后将结果投影到 $\mathcal{S}$：

$$\tilde{\boldsymbol{\eta}}^{(t+1)}\leftarrow\boldsymbol{\eta}^{(t)}+\alpha_t\widehat{\nabla_{\boldsymbol{\eta}}I}(\boldsymbol{\eta}^{(t)}) \tag{57}$$

$$\boldsymbol{\eta}^{(t+1)}\leftarrow\mathcal{P}_{\mathcal{S}}(\tilde{\boldsymbol{\eta}}^{(t+1)}). \tag{58}$$

这里，$\alpha_t>0$ 是学习率（步长），$\widehat{\nabla_{\boldsymbol{\eta}}I}$ 是使用算法 1 在一个小批量数据上估计得到的梯度。算子 $\mathcal{P}_{\mathcal{S}}(\cdot)$ 表示投影到可行集 $\mathcal{S}$上的欧几里得投影：

$$\mathcal{P}_{\mathcal{S}}(\boldsymbol{v})\triangleq\arg\min_{\boldsymbol{\eta}'\in\mathcal{S}}\|\boldsymbol{\eta}'-\boldsymbol{v}\|_2. \tag{59}$$

对于全局功率约束（$C(\boldsymbol{\eta})=\|\boldsymbol{\eta}\|_2^2\le P$）等常见约束，这种投影有简单的闭式解（若向量的范数超过 $\sqrt{P}$，则对其进行缩放）。重复这些步骤即可更新网络参数 $\boldsymbol{\eta}$，在严格满足全局资源约束的同时最大化端到端信息流。

备注 6（收敛性与非凸性）。一般而言，互信息目标函数 $I(X;Y)$ 关于网络参数 $\boldsymbol{\eta}$ 是非凸的，尤其是在有向无环图涉及深度神经网络等复杂非线性映射时。因此，标准的基于梯度的优化方法（包括所提出的 PGA算法）只能保证收敛到目标函数的某个驻点（局部最优点或鞍点），而不一定是全局最大值。尽管存在这一理论局限，基于梯度的互信息最大化方法已被证明在各种实际应用中非常有效。

注释 7（非凸性缓解策略）。为降低陷入较差局部最优点的风险，可以采用几种常见的实用策略。一种有效的方法是多轮优化：使用不同的 $\boldsymbol{\eta}$ 随机初始化，多次执行梯度上升，并选取估计互信息最高的解。

另一种有望奏效的方法是噪声退火。通过向 Y 人为注入额外的高斯噪声，并在优化过程中逐渐将其方差降至零，可以在初始阶段平滑目标函数的地形，从而可能使优化器摆脱较浅的局部最优点，最终收敛到一个解。

C. 实际考量：得分漂移

上述 PGA 算法在实现中面临的一项关键挑战是，更新网络参数 $\boldsymbol{\eta}$ 会内在地改变底层数据分布 $p_Y(\boldsymbol{y};\boldsymbol{\eta})$ 和 $p_{Y|X}(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\eta})$。因此，真实的得分函数会在优化过程中发生“漂移”。使用基于旧参数配置训练的过时得分模型来估计新配置下的梯度，可能会导致更新方向出现偏差或错误。

为缓解这一问题，可以采用两种主要策略：

- **迭代式重新训练：** 最简单的方法是在得分函数学习和参数优化之间交替进行。针对 $\boldsymbol{\eta}$ 执行几步梯度上升后，使用从当前有向无环图生成的新鲜样本对评分模型进行微调。
- **条件得分网络：** 一种更先进的方法是训练条件评分模型 $s_\phi(\boldsymbol{y};\boldsymbol{\eta})$ 和$s_\psi(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\eta})$，使其显式地将网络参数 $\boldsymbol{\eta}$ 作为额外输入。通过在相关的 $\boldsymbol{\eta}$范围内预训练这些模型，我们可以在无需频繁重新训练的情况下进行优化，从而显著加快这一过程。

这两种策略都需要在实现复杂度与计算效率之间进行权衡。迭代式重新训练易于实现，并能确保局部得分估计具有很高的准确性，但在优化阶段，由于需要频繁重新训练，可能会成为计算瓶颈。相比之下，条件得分网络无需重新训练，因而能显著加快在线优化阶段的速度。不过，为确保模型能够在$\boldsymbol{\eta}$相关参数空间内良好泛化，它们需要更复杂且可能更加依赖数据的预训练阶段。

D. 潜在影响与应用

所提出的框架能够在全局约束下对一般有向无环图进行端到端的信息论优化，在涉及复杂随机系统的多个领域都具有显著的潜在影响。

1) 下一代无线系统（语义通信）：不同于旨在最大化单纯数据传输速率的传统通信，语义通信 [11]等下一代通信范式专注于最大化下游任务的成功执行率。我们的框架将整个通信链路视为一个有向无环图，因此能够自然地支持这一目标。关键在于，这使得发射机（编码器）和接收机（解码器）可以进行联合优化。通过将编码器 $\boldsymbol{\eta}_{\text{TX}}$、噪声信道和解码器 $\boldsymbol{\eta}_{\text{RX}}$ 建模为一个统一的有向无环图，我们可以同时优化两者 $\boldsymbol{\eta}_{\text{TX}}$ 和 $\boldsymbol{\eta}_{\text{RX}}$ ，以最大化与任务相关的互信息 $I(T;\ Y_{\text{final}})$在接收机决策点处的值，而不只是最大化信道容量。当信道涉及复杂的非线性失真、难以通过经典模块化设计进行处理时，这种端到端方法尤其有价值。

2) 物联网与分布式传感器网络：考虑一个大型传感器网络，其中大量受电池容量限制的传感器观测同一种物理现象 X ，并将处理后的数据传输至中心枢纽 Y。该系统构成一个具有全局功率约束 $C(\{\boldsymbol{\eta}_i\}) \le P_{\text{total}}$的大型有向无环图。我们的框架提供了一种有原则的方法，用于确定最优运行策略（哪些传感器应当启用、数据应压缩到何种程度，以及传输时应分配多少功率），以最大化中心处的聚合信息 $I(X;Y)$。

3）深度学习与表示学习：深度神经网络（DNN）可视为确定性有向无环图（DAG）（若采用丢弃法或噪声激活等技术，则可视为随机有向无环图）。我们的方法提供了一种直接途径，可将信息论原理（例如信息瓶颈（*IB*）原理 [12], ）融入深度神经网络训练。利用为信息瓶颈目标函数 $\mathcal{L}_{\text{IB}} = I(T;\ Y)\ -\beta I(X;Y)$ 推导出的梯度，我们可以训练网络学习表示 Y，使其在最大程度上包含与目标任务 T 相关的信息，同时尽可能少地包含原始输入 X 的信息，从而提升泛化能力和鲁棒性。

VI. 与网络信息论的关系 EORY

所提出的有向无环图框架不仅可用于任意计算图，还自然涵盖了网络信息论 [1]中研究的许多基本信道模型。通过从有向无环图的视角审视这些经典模型，我们基于梯度的方法为其端到端互信息、容量或可达速率最大化问题提供了统一的数值求解器，即使在复杂的非线性或非高斯假设下（此时解析解难以求得）也适用。

A. 多用户信道模型作为有向无环图

许多典型的多用户信道都可以直接映射为特定的有向无环图拓扑：

- **多址信道 (MAC)：** 一个包含两个发送端 $X_1,\ X_2$ 和一个接收机 Y 的多址信道，可表示为一张有向无环图，其中两个根节点汇合为一个子节点。

$$Y = f_{\text{MAC}}(X_1, X_2; \boldsymbol{\eta}) + Z. \tag{60}$$

在 $X_1,\ X_2$ 各自的功率约束下最大化 $I(X_1,\ X_2;Y)$，即可得到和容量点。

- **广播信道 (BC)：** 一个包含一个发送端 X 和两个接收机 $Y_1,\ Y_2$ 的广播信道，对应一张有向无环图，其中一个根节点分支为两个叶节点。

$$Y_1 = f_1(X;\boldsymbol{\eta}_1) + Z_1, \quad Y_2 = f_2(X;\boldsymbol{\eta}_2) + Z_2. \tag{61}$$

容量区域涉及 $I(X;Y_1)$ 和 $I(X;Y_2)$ 之间的权衡，通常通过优化输入分布 p_X 或叠加编码参数 $\boldsymbol{\eta}$ 来协调。

这两种信道模型如图 3 所示。

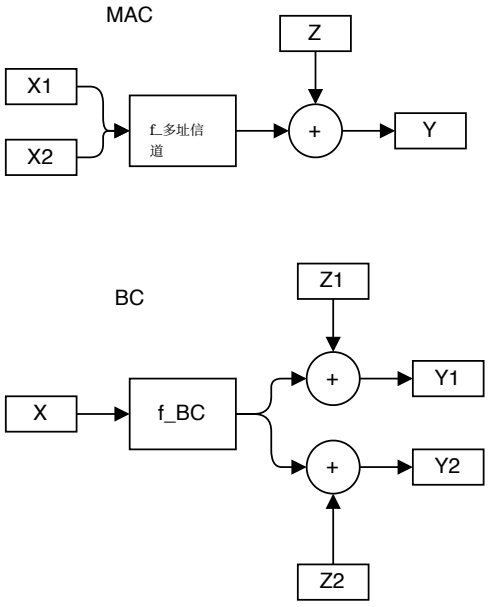

图 3. 多址信道模型和广播信道模型。

B. 统一数值求解器

传统上，要找出这些模型的容量区域，通常需要针对特定模型的解析技术（例如，熵不等式、逆向证明）；当组件为非线性或非高斯时，这些技术会变得极其困难。我们基于有向无环图的信息梯度框架提供了一种统一的数值方法：通过指定图拓扑并确定可调参数 $\boldsymbol{\eta}$（这些参数可以用于参数化编码器、解码器或中继函数），即可应用同一种通用优化算法（算法 1 +PGA），以数值方式发现各种网络场景中的可达速率区域。

C. 输入分布优化

信道容量问题通常需要优化输入分布 p_X。我们的框架可以将输入生成过程本身视为有向无环图的一部分，从而自然地纳入这项优化。任何能够通过计算过程进行抽样的分布（例如逆变换抽样、高斯混合模型，或变分自编码器、整流流等深度生成模型 [19]）都可以表示为一个子图，其中根节点是原始噪声源（例如多维标准高斯分布或均匀随机变量）。将这个生成子图接到主信道有向无环图之前，输入分布的参数便成为标准的网络参数 $\boldsymbol{\eta}$。这样便能以统一的方式，端到端地基于梯度优化源分布以及信道/系统参数，从而为界定信道的容量区域提供一种强大的数值方法。

D. 速率区域探索的统一流程

本文提出的信息梯度框架为探索多用户信道的速率区域提供了一套统一的数值方法。对于输入分布和系统组件的给定参数化，由互信息表达式定义的速率区域构成了容量区域的一个界。当该参数化具有足够的表达能力来表示达到容量的输入分布时，此界在数值上与真实容量区域重合。我们概述了这一通用方法，它适用于任何可表示为有向无环图的信道，无论是否存在解析解。

1）加权和最大化：速率区域的边界可刻画为涉及多个互信息表达式的多目标优化问题的帕累托前沿。描绘该边界的一种标准方法是求解一系列加权优化问题，这些问题由 $\boldsymbol{\lambda} = (\lambda_1, \ldots\ldots, \lambda_K)$ 参数化，其中 $\lambda_i \geq 0$:

$$\max_{\boldsymbol{\theta}} \sum_{i=1}^{K} \lambda_i R_i(\boldsymbol{\theta}) \quad \text{subject to resource constraints,} \tag{62}$$

其中，$R_i(\boldsymbol{\theta})$ 表示第 i 个互信息表达式，可以是无条件互信息 $I(X_i; Y)$，也可以是条件互信息 $I(X_i; Y|X_j)$；$\boldsymbol{\theta}$ 表示系统中可调的参数（例如输入分布参数、编码器权重或功率分配）。在概率单纯形上遍历 $\boldsymbol{\lambda}$，并针对每种权重配置求解（62），即可获得一系列边界点，从而描绘出速率区域的一个界。

2）条件互信息的信息梯度：要对（62）应用基于梯度的优化，我们需要求条件互信息关于系统参数的梯度。定理 2 可自然地推广到条件互信息，如下所示。

推论 3（条件互信息的信息梯度）。对于输出 Y 取决于输入 X_1、X_2 和参数 $\boldsymbol{\theta}$ 的有向无环图信道，其条件互信息 $I(X_1; Y|X_2)$ 的梯度为

$$\begin{aligned}&\nabla_{\boldsymbol{\theta}} I(X_1; Y|X_2) \\ &= \mathbb{E}\left[(D_{\boldsymbol{\theta}} Y)^\top \left(s_{Y|X_1,X_2}(Y|X_1, X_2) - s_{Y|X_2}(Y|X_2)\right)\right],\end{aligned} \tag{63}$$

其中，$s_{Y|X_1,\ X_2}$ 和 $s_{Y|X_2}$ 为条件得分函数。

证明：这可直接由分解 $I(X_1; Y|X_2) = h(Y|X_2) - h(Y|X_1,\ X_2)$ 得出，并将定理 2 中的相同推导方法应用于每个条件熵项。 ■

该结果结合定理 2，可计算标准多用户信道特征描述中所有互信息表达式的梯度，

算法 2 速率区域边界探索（双用户情形）

输入： 有向无环图结构 $\mathcal{G}$、互信息表达式 R_1，R_2，权重样本数量 L，约束集 $\mathcal{C}$

输出： 可达速率区域 $\mathcal{R}$

```
B ← ∅                                   ▷ 边界点集合
对 l = 0 至 L 执行
    λ ← l/L                             ▷ 权重扫描
    初始化参数 θ
    重复
        从有向无环图 G 生成样本，使用当前 θ
        通过去噪分数匹配估计所需的得分函数
        使用 (63) 计算 ∇_θ R_1 和 ∇_θ R_2
        ∇_θ L ← λ∇_θ R_1 + (1 − λ)∇_θ R_2
        θ ← P_C(θ + α∇_θ L)             ▷ 投影梯度上升
    直到收敛
    评估 (R_1^*, R_2^*) ← (R_1(θ), R_2(θ))（例如，通过 Fisher 积分或解析表达式（如果有）
    B ← B ∪ {(R_1^*, R_2^*)}
结束循环
R ← 凸包(B)                             ▷ 时分共享可达性
返回 R
```

包括多址信道和速率 $I(X_1,\ X_2;Y)$ 和个体速率$I(X_i;Y|X_j)$。

3) 算法框架：为便于说明，我们以双用户信道（例如多址信道）为例描述该过程，其速率区域由两个互信息表达式 $R_1(\boldsymbol{\theta})$ 和 $R_2(\boldsymbol{\theta})$ 刻画。算法 2 概述了边界探索过程。该算法在均匀网格上扫描标量权重 $\lambda \in [0,1]$，针对每个 λ 求解加权优化问题 $\max_{\boldsymbol{\theta}} \lambda R_1 + (1-\lambda)R_2$，最后对所得点集计算凸包，以获得速率区域。

最终的凸包运算体现了这样一个事实：通过在相应编码策略之间进行时分共享，任何可达速率对的凸组合也都是可达的。

一个重要的实际考量是 得分漂移：随着 $\boldsymbol{\theta}$ 在优化过程中不断变化，底层分布也会随之改变，因此需要更新得分模型。对此，可以定期重新训练评分网络，或者采用条件得分模型 $s_\phi(y;\boldsymbol{\theta})$，将系统参数作为额外输入，如 V-C 节所述。

4) 范围与未来方向：本文提出的框架为通过信息梯度探索速率区域奠定了理论和算法基础。虽然线性高斯情形可以进行解析验证（见第 VI-E 节），但该方法的主要价值在于适用于无法获得闭式解的信道，例如具有 $Y = f(X_1,\ X_2)+Z$ 的非线性多址信道，或具有非高斯噪声的信道。对于此类信道，算法 2 提供了一种系统的数值方法，用于探索容量区域的界。针对特定非线性多用户信道开展全面的数值研究，是未来研究的一个有前景的方向。

E. 解析验证：高斯多址信道

为了验证算法 2，我们考虑具有总功率约束的线性高斯多址信道。我们通过两条独立路径推导速率区域，并证明两者等价：路径 1 使用信息梯度公式（推论 3）和解析得分函数；路径 2 则直接对已知的互信息表达式求导。所得速率区域的一致性证实了所提框架的正确性。

1) 信道模型：考虑由以下定义的双用户高斯多址信道

$$Y = X_1 + X_2 + Z, \tag{64}$$

其中，$X_1 \sim \mathcal{N}(0,\ P_1)$ 和 $X_2 \sim \mathcal{N}(0,\ P_2)$ 是相互独立的高斯输入，$Z \sim \mathcal{N}(0,\ \sigma^2)$ 是独立的高斯噪声，并施加总功率约束 $P_1+P_2 \le P$。速率表达式为：

$$R_1(P_1) = I(X_1;Y|X_2) = \frac{1}{2}\log\left(1+\frac{P_1}{\sigma^2}\right), \tag{65}$$

$$R_2(P_2) = I(X_2;Y|X_1) = \frac{1}{2}\log\left(1+\frac{P_2}{\sigma^2}\right). \tag{66}$$

2) 路径 *1*：基于得分函数的推导：我们使用信息梯度框架计算速率区域。该高斯信道的解析得分函数为：

$$s_{Y|X_1,X_2}(y|x_1,x_2) = -\frac{y-x_1-x_2}{\sigma^2}, \tag{67}$$

$$s_{Y|X_2}(y|x_2) = -\frac{y-x_2}{P_1+\sigma^2}. \tag{68}$$

使用参数化 $X_1 = \sqrt{P_1}\tilde{X}_1$，其中 $\tilde{X_1} \sim \mathcal{N}(0,\ 1)$，可得 $D_{P_1}Y = X_1/(2P_1)$。应用推论 3：

$$\nabla_{P_1} R_1 = \mathbb{E}\left[\frac{X_1}{2P_1}\left(s_{Y|X_1,X_2} - s_{Y|X_2}\right)\right] = \frac{1}{2(P_1+\sigma^2)}. \tag{69}$$

根据对称性，$\nabla_{P_2} R_2 = 1/(2(P_2+\sigma^2))$。

对于加权目标函数 $\mathcal{L} = \lambda R_1 + (1-\lambda)R_2$，约束条件为 $P_1+P_2=P$，使用这些梯度得到的 KKT 驻点条件为：

$$\frac{\lambda}{2(P_1+\sigma^2)} = \mu, \quad \frac{1-\lambda}{2(P_2+\sigma^2)} = \mu. \tag{70}$$

消去 μ，并使用 $P_1+P_2=P$ 求解：

$$P_1^{(1)}(\lambda) = \lambda(P+2\sigma^2) - \sigma^2, \tag{71}$$

$$P_2^{(1)}(\lambda) = (1-\lambda)(P+2\sigma^2) - \sigma^2. \tag{72}$$

代入速率表达式，边界点为：

$$R_1^{(1)}(\lambda) = \frac{1}{2}\log\left(\frac{\lambda(P+2\sigma^2)}{\sigma^2}\right), \tag{73}$$

$$R_2^{(1)}(\lambda) = \frac{1}{2}\log\left(\frac{(1-\lambda)(P+2\sigma^2)}{\sigma^2}\right). \tag{74}$$

3）路径 *2*：经典推导：我们通过直接对已知表达式 (65)–(66) 求导来计算速率区域：

$$\nabla_{P_1} R_1 = \frac{1}{2(P_1+\sigma^2)}, \quad \nabla_{P_2} R_2 = \frac{1}{2(P_2+\sigma^2)}. \tag{75}$$

$\mathcal{L} = \lambda R_1 + (1-\lambda) R_2$ 的 KKT 条件 $P_1 + P_2 = P$ 与 (70) 完全相同。求解后得到：

$$P_1^{(2)}(\lambda) = \lambda(P + 2\sigma^2) - \sigma^2, \tag{76}$$

$$P_2^{(2)}(\lambda) = (1-\lambda)(P + 2\sigma^2) - \sigma^2. \tag{77}$$

相应的边界点为：

$$R_1^{(2)}(\lambda) = \frac{1}{2}\log\left(\frac{\lambda(P+2\sigma^2)}{\sigma^2}\right), \tag{78}$$

$$R_2^{(2)}(\lambda) = \frac{1}{2}\log\left(\frac{(1-\lambda)(P+2\sigma^2)}{\sigma^2}\right). \tag{79}$$

4) 速率区域的等价性：比较 (73)–(74) 与 (78)–(79)，可确认：

$$\left(R_1^{(1)}(\lambda), R_2^{(1)}(\lambda)\right) = \left(R_1^{(2)}(\lambda), R_2^{(2)}(\lambda)\right), \quad \forall \lambda \in [0,1]. \tag{80}$$

两种路径得到的速率区域边界完全相同。加入位于 $\lambda \in \{0,1\}$ 的角点，其中 $(R_1, R_2) = (0, \frac{1}{2}\log(1+P/\sigma^2))$ 或 $(\frac{1}{2}\log(1+P/\sigma^2), 0)$，再取凸包，两种路径均得到相同的可达速率区域。

5) 启示：这项解析验证表明，以计算方式实现路径 1 的 Algorithm2 能够精确复现通过路径 2 推导出的已知容量区域。关键在于，基于得分函数的互信息梯度（推论 3）提供了正确的优化方向，从而得到相同的最优功率分配，进而得到相同的速率区域。对于非线性或非高斯信道，由于不存在互信息闭式表达式，路径 2 无法使用；此时，采用学习得到的得分函数的路径 1 是探索速率区域边界的唯一系统性方法。

VII. 数值实验

A. 高斯多路径有向无环图上的信息梯度估计

在转向更复杂的非线性和高维示例之前，我们首先对定理 2 在第 IV-C 节引入的线性高斯多路径有向无环图上进行数值合理性检查。在该小节中，给出了有向无环图结构

$$X \to (V_1, V_2) \to Y$$

并推导出相应的等效标量高斯信道 $Y = G_{\text{eff}} X + Z_{\text{eff}}$，以及互信息 $I(X;Y)$ 的闭式表达和关于 (η_1, η_2) 的梯度。特别是，解析梯度

$$\frac{\partial I}{\partial \eta_1}, \quad \frac{\partial I}{\partial \eta_2}$$

在式 (45) 和 (46) 中得到了这些结果，并通过解析方法验证了它们与定理 2 中有向无环图梯度公式的一致性

使用高斯得分函数和雅可比矩阵 $D_{\eta_k} Y$。在本小节中，我们通过对同一模型进行完全由数据驱动的蒙特卡洛研究，补充这项分析。

1) 实验设置：我们完全沿用第 IV-C 节中的线性高斯多路径有向无环图，包括方差参数以及 G_{eff}、σ_N^2 和 v_Y 的定义。除非另有说明，我们采用归一化设置

$$\sigma_X^2 = \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1, \tag{81}$$

使输入和所有噪声分支的方差均为单位方差。我们考虑参数的两种一维扫描：

- 在第一次扫描中，我们固定 $\eta_2 = \eta_{2,\ \text{固定}} = 1$ 并使用 K_1 个等间距网格点，在 η_1的$[\eta_{1,\min}\ \eta_{1,\ \max}] = [-2, 2]$ 范围内进行变化。

- 在第二次扫描中，我们固定 $\eta_1 = \eta_{1,\ \text{固定}} = 1$ 并使用 K_2 个等间距网格点，在 η_2的$[\eta_{2,\min}\ \eta_{2,\ \max}] = [-2, 2]$ 范围内进行变化。

对于每个网格点（η_1，η_2），我们用三种不同方法计算梯度 $\partial I/\partial \eta_1$和 $\partial I/\partial \eta_2$：

(i) 解析梯度：我们直接计算第 IV-C 节中推导出的闭式表达式 (45) 和 (46)；并将其视为真实值。

(ii) 使用解析得分函数的蒙特卡洛梯度：我们将定理 2 应用于多路径有向无环图，使用与第 IV-C 节相同的高斯得分函数 $s_Y(y)$ 和 $s_{Y|X}(y|x)$，并通过反向模式自动微分（向量-雅可比积）计算雅可比矩阵 $D_{\eta_k} Y$。具体而言，我们估计

$$\frac{\partial I}{\partial \eta_k} \approx \frac{1}{N}\sum_{n=1}^{N} \left(D_{\eta_k} Y^{(n)}\right)\left(s_{Y|X}(Y^{(n)}|X^{(n)}) - s_Y(Y^{(n)})\right), \tag{82}$$

其中（$X^{(n)}$，$Y^{(n)}$）是由有向无环图生成的独立同分布样本。

（iii）使用完全学习的得分函数计算蒙特卡洛梯度：我们再次使用定理 2，但这次将无条件得分函数 $s_Y(y)$ 和条件得分函数 $s_{Y|X}(y|x)$ 均替换为通过去噪得分匹配（DSM）获得的神经得分估计器 $\hat{s}_Y(y)$ 和 $\hat{s}_{Y|X}(y|x)$。该设置不依赖于真实密度的任何信息，从而验证了该方法作为黑盒估计器的有效性。

对于蒙特卡洛估计（ii）和（iii），我们使用 $N = 10^5$ 独立同分布样本（$X^{(n)}$，$Z_1^{(n)}$，$Z_2^{(n)}$，$Z_3^{(n)}$），通过多路径有向无环图生成相应的 $Y^{(n)}$，并使用小批量大小为 8，192 的设置计算梯度。

2）得分函数学习细节：对于每个网格点（η_1，η_2），我们都会针对相应的信道实现训练专用的评分模型。训练数据由从第 IV-C 节的多路径有向无环图中采样得到的干净样本组成。我们采用两个独立的多层感知机：一个用于无条件得分函数 $\hat{s}_Y(y)$，另一个用于条件得分函数 $\hat{s}_{Y|X}(y|x)$（以 y 和 x 的拼接结果作为输入）。每个网络均包含三个宽度为 128 的隐藏层，使用 SiLU 激活函数，并配有一个线性输出层。

我们通过去噪得分匹配，在固定噪声方差 $t_{\text{DSM}} = 0.05$下训练 $\hat{s}_Y$ 和 $\hat{s}_{Y|X}$。无条件得分函数经过训练以对 Y进行去噪，而条件得分函数经过训练

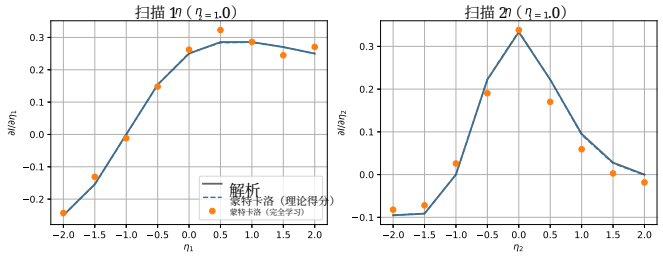

Fig. 4. 第 IV-C 节中线性高斯多路径有向无环图的信息梯度。左：$\partial I/\partial \eta_1$ 作为 η_1 的函数，η_2 固定。右：$\partial I/\partial \eta_2$ 作为 η_2 的函数，η_1 固定。在每个面板中，我们比较解析梯度（实线）、使用解析得分函数对定理 2 进行蒙特卡洛评估的结果（虚线），以及使用完全学习的得分函数对定理 2 进行蒙特卡洛评估的结果（标记）。

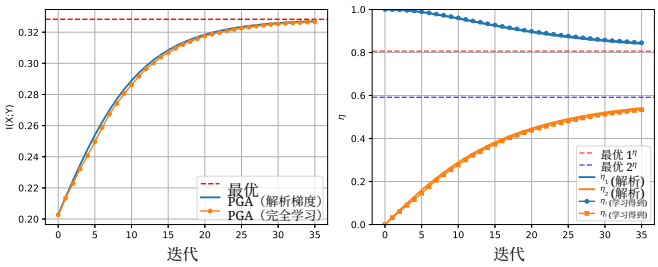

Fig. 5. 在约束 $\eta_1^2+\eta_2^2=1$ 下最大化多路径有向无环图的互信息。左：X(Y；η_1) 随 PGA 迭代的变化。右：参数 η_2和 s_Y 的轨迹。所提出的完全学习方法（标记点）通过迭代微调和双重斯坦校准在线学习$s_{Y|X}$ 和，能够完美跟踪通过解析梯度计算出的最优轨迹（实线）。

以在给定 X的情况下对 Y 进行去噪。在报告的实验中，我们使用批量大小 4,096、500 个梯度步，以及学习率为 10^{-3}的 AdamW 优化器。

训练后，我们应用标量斯坦校准，以纠正学习得到的无条件得分函数 $\hat{s}_Y$ 的整体缩放偏差。对于一维输出，这相当于按以下比例缩放

$$\tilde{s}_Y(y) = c\,\hat{s}_Y(y), \qquad c = -\frac{1}{\mathbb{E}[Y\,\hat{s}_Y(Y)]}, \tag{83}$$

这会强制满足 $\mathbb{E}[Y\tilde{s}_Y(Y)] = -1$。在蒙特卡洛梯度 (iii) 中，我们使用 $\tilde{s}_Y$ 代替 $\hat{s}_Y$ （并结合原始的 $\hat{}s_{Y|X}$）。

3) 结果： 图 4 汇总了结果。左图显示 $\partial I/\partial \eta_1$ 随 η_1 变化的情况，其中 η_2 固定；右图显示 $\partial I/\partial \eta_2$ 随 η_2 变化的情况，其中 η_1 固定。每幅图中我们绘制了：

- 解析梯度 (45) 或 (46)（实线），
- 基于定理 2、使用解析得分函数 $s_{Y|X}$ 和 s_Y 得到的蒙特卡洛估计（虚线），以及
- 基于定理 2、使用完全学习的得分函数 $\hat{}s_{Y|X}$ 和 $\tilde{s}_Y$ 得到的蒙特卡洛估计（标记）。

对于 η_1 和 η_2扫描，在图示尺度下，解析梯度与使用解析得分函数得到的蒙特卡洛梯度无法区分，这有力地验证了定理 2 中的梯度公式及其基于向量-雅可比积的实现。此外，由经去噪分数匹配训练的评分网络得到的数据驱动梯度，在整个 $(\eta_1,\ \eta_2)$ 范围内都紧密贴合解析曲线。这表明，即使将有向无环图视为生成 $(X,\ Y)$ 样本的完全黑盒生成器，所提出的得分学习方法仍能忠实地复现信息梯度；这也补充了第 IV-C 节中给出的解析验证。

B. 全局约束下多路径有向无环图的优化

我们将所提出的信息最大化框架应用于第 IV-C 节中定义的线性高斯多路径有向无环图。目标是在参数 η = 下最大化端到端互信息 $I(X;Y)$ $(\eta_1,\ \eta_2)$，并满足全局功率约束 $\|\eta\|_2^2 = 1$。此问题用于对我们的框架进行动态验证，评估在优化过程中底层数据分布发生变化时，梯度估计器是否仍然稳健。

1) 实验设置：我们将参数初始化为 $(\eta_1,\ \eta_2)$ =(1.0, 0.0)。我们采用步长固定为 $\alpha = 0.1$ 的 PGA 算法，进行 35 次迭代。每次迭代中，我们使用定理 2 估计梯度 $\nabla_\eta I(X;Y)$。关键的是，为了测试完全黑盒能力，我们假设不知道解析得分函数。相反，我们使用完全学习的得分函数来估计梯度，其中无条件得分函数 $s_Y(y)$ 和条件得分函数$s_{Y|X}(y|x)$ 均由神经网络近似。

2）迭代得分学习策略：由于 $\boldsymbol{\eta}$ 不断演变，Y（以及 $Y|X$）的分布也随之变化，因此每个 PGA 步骤都必须更新评分网络。为高效处理这一问题，我们采用热启动策略。我们维护两个评分网络$\hat{s}_Y$和$\hat{}s_{Y|X}$，并在各次迭代中持续使用它们。在首次迭代（$t=0$）中，我们使用去噪分数匹配（DSM）从头训练这些网络 1,000 步，以确保初始梯度准确。在后续迭代（$t>0$）中，我们从上一次迭代的权重开始，仅使用去噪分数匹配（DSM）进行 200 步更新。这种微调方法显著降低了计算成本，同时使得分模型能够有效跟踪不断漂移的分布。

两个网络均为多层感知机（MLPs），其架构与上一小节相同。为进行校准，我们在每次迭代中对无条件得分和条件得分应用斯坦校准。具体而言，对于条件得分函数$\hat{}s_{Y|X}$，我们计算一个标量校正因子 c，使其满足 $\mathbb{E}_X[\mathbb{E}_{Y|X}[\langle Y - \mathbb{E}[Y|X],\ c\,\hat{}s_{Y|X}(Y|X)\rangle]] = -1$，从而确保在整个优化过程中获得稳健的尺度估计。

3）结果：图 5 展示了优化结果。左图显示了互信息$I(X;Y)$随迭代变化的轨迹，对比了使用解析梯度的 PGA（实线）和使用完全学习的梯度的 PGA（标记）。右图则展示了参数 η_1 和 η_2对应的变化轨迹。

结果表明，所提出的方法与真实解析优化高度一致。完全学习的方法稳步提升互信息，并以与解析基线相同的速度收敛至理论最优值（红色虚线所示）。此外，右图证实，参数（η_1，η_2）沿着约束流形上的正确路径变化。这项实验表明，采用迭代微调策略的框架即使无法获取闭式似然，也能在动态环境中成功实现端到端优化。

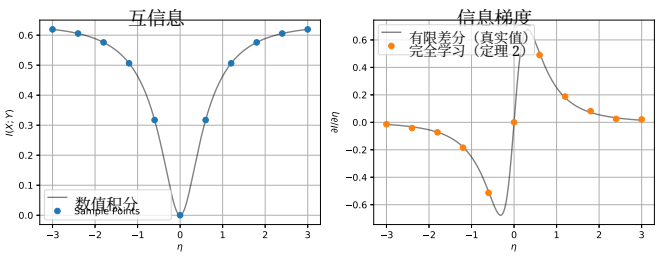

图 6. 非线性标量信道 $Y = \tanh(\eta X) + Z$，其中 $X \sim \mathcal{N}(0,1)$，且 $Z \sim \mathcal{N}(0, 0.25)$。左图：互信息 $I(X;Y)$ 随非线性参数 η 的变化，通过使用高斯–埃尔米特求积计算 $I(X;Y) = h(Y) - h(Z)$ 的数值积分得到，其中 $p_Y(y)$ 采用高斯–埃尔米特求积，y 使用均匀网格。右图：信息梯度 $\partial I/\partial\eta$ 随 η 的变化。我们比较了基于数值计算的 $I(\eta)$ 的有限差分近似（实线）与使用完全学习和 Stein 校准得分、根据定理 2 得到的梯度（标记）。

C. 非线性标量信道

这里，我们验证所提出的信息梯度估计器在一种真正的非线性信道中同样能够可靠工作，而该信道不存在闭式互信息。我们考虑如下标量模型：

$$X \sim \mathcal{N}(0,1), \quad Z \sim \mathcal{N}(0,\sigma_Z^2), \quad Y = \tanh(\eta X) + Z, \tag{84}$$

噪声方差固定为 $\sigma_Z^2 = 0.25$，标量参数为 $\eta \in [\eta_{\min}, \eta_{\max}] = [-3, 3]$。对于小网格中 η 的每个取值，我们根据定义 $I(X;Y) = h(Y) - h(Z)$，对 $I(X;Y)$ 进行数值计算。这里，$h(Z)$ 是解析的，而 $h(Y)$ 则首先通过在 x 上进行高阶高斯–埃尔米特求积来计算 $p_Y(y) = \int p_X(x) p_{Y|X}(y|x)\, dx$，然后在有限区间上使用均匀的 y 网格和梯形法则近似 $H(Y) = -\int p_Y(y) \log p_Y(y)\, dy$。

真实梯度 $\partial I/\partial\eta$ 通过对这些数值进行中心有限差分来近似

$$\frac{\partial I}{\partial \eta}(\eta) \approx \frac{I(\eta + h) - I(\eta - h)}{2h}, \tag{85}$$

步长较小，为 $h = 0.01$。与此同时，我们使用定理 2 和完全学习的得分函数实现来估计信息梯度：无条件得分函数 $s_Y(y)$ 和条件得分函数 $s_{Y|X}(y|x)$ 均由经去噪分数匹配训练的评分网络近似。此外，两种得分函数都经过标量Stein校准进行校正。定理 2 中的期望通过蒙特卡洛方法，使用 $N = 10^5$ 个样本进行计算。

图6总结了结果。[1] 左图展示了互信息 $I(X;Y)$ 随 η 变化的情况，呈现出预期的偶对称性，并且随着 $|\eta|$ 增大而趋于饱和。右图比较了有限差分梯度与使用完全学习的得分函数，通过定理 2 得到的信息梯度。在整个 η 范围内，两条曲线几乎无法区分，包括真实梯度在 $\eta = 0$ 附近为零的区域，以及 $I(X;Y)$ 趋于饱和的高增益区域。这项实验表明，所提出的

即使对于只能通过数值积分获得 $I(X;Y)$ 本身的非线性信道，基于得分的信息梯度估计器仍然准确且稳定。

[1] 另请参阅补充材料中的数值曲线。

VIII. 扩展：数字孪生校准

ION

到目前为止所介绍的信息最大化框架基于这样一个前提：有向无环图模型 $\mathcal{G}$ 能够准确表示真实物理系统。如果模型与现实之间存在显著差异，那么在仿真中优化得到的参数 $\boldsymbol{\eta}^*$ 可能无法在真实世界中达到预期效果。因此，将有向无环图模型校准至与现实相符（即数字孪生校准）是实际部署的先决条件。本节将说明，我们的基于得分函数的框架可以自然地扩展，以解决这一校准问题。

A. 问题表述

设真实物理系统由一个未知的随机过程控制，该过程根据输入 X 和内部噪声源 $\mathcal{Z}_{\text{real}}$ 产生输出 Y_{real}：

$$Y_{\text{real}} = g_{\text{real}}(X, \mathcal{Z}_{\text{real}}). \tag{86}$$

我们的数字孪生是式 (38) 中定义的有向无环图模型：

$$Y_{\boldsymbol{\eta}} = g(X, \mathcal{Z}; \boldsymbol{\eta}). \tag{87}$$

我们假设模型中的输入 X 和噪声源 $\mathcal{Z}$ 的分布与真实系统中的分布一致。校准的目标是找到参数 $\boldsymbol{\eta}^*$，使真实系统 g_{real} 与有向无环图模型 $g(\cdot;\, \boldsymbol{\eta}^*)$ 之间的差异最小。

B. 基于得分的校准策略

由于真实机制 g_{real} 未知，而我们只能观测到来自 Y_{real} 的样本，因此无法直接比较内部函数。相反，我们的目标是匹配它们的输出分布。基于 SFB 理念，我们将真实输出分布与模型输出分布之间的 Fisher散度作为校准目标：

$$\begin{aligned} \mathcal{J}_{\text{calib}}(\boldsymbol{\eta}) &\equiv D_F(p_{Y_{\text{real}}} \| p_{Y_{\boldsymbol{\eta}}}) \\ &= \mathbb{E}_{\boldsymbol{y} \sim p_{Y_{\text{real}}}} \left[\| s_{Y_{\text{real}}}(\boldsymbol{y}) - s_{Y_{\boldsymbol{\eta}}}(\boldsymbol{y}; \boldsymbol{\eta}) \|^2 \right]. \end{aligned} \tag{88}$$

最小化该目标函数会促使有向无环图模型生成与真实系统输出在统计上无法区分的输出（就其得分函数而言）。图 7 展示了基于得分的校准策略。

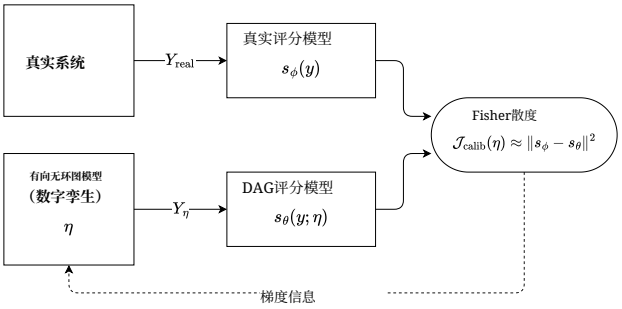

图 7. 基于得分的校准策略。使用来自物理系统的观测数据 Y_{real} 训练“真实评分模型”（s_ϕ）。使用来自数字孪生的仿真数据 $Y_{\boldsymbol{\eta}}$ 训练“DAG评分模型”（$s_{\boldsymbol{\theta}}$）。校准目标 $\mathcal{J}_{\text{calib}}(\eta)$ 是这两个得分函数之间的费舍尔散度（用平方差近似）。通过该目标函数的梯度更新 DAG 模型参数 $\boldsymbol{\eta}$，使其与现实相符。

C. 通过费舍尔散度最小化进行校准的算法

式 (88) 的优化可实际通过两种类型的得分模型来完成：

- **真实评分模型** $s_{\boldsymbol{\phi}}(y)$**：** 使用观测样本 $\{\boldsymbol{y}_{(i)\ \text{real}}\}$训练得到，采用切片评分匹配（SSM）等无监督技术；由于真实的g_{real} 未知，因此将其视为干净数据。
- **DAG评分模型** $s_{\boldsymbol{\theta}}(\boldsymbol{y};\boldsymbol{\eta})$**：** 在来自有向无环图 $Y_{\boldsymbol{\eta}} = g(X, \mathcal{Z}; \boldsymbol{\eta})$ 的仿真数据上使用标准去噪分数匹配预训练的条件得分网络。

随后，通过对估计的Fisher散度进行梯度下降来更新校准参数 $\boldsymbol{\eta}$：

$$\boldsymbol{\eta}_{k+1} \leftarrow \boldsymbol{\eta}_k - \alpha \nabla_{\boldsymbol{\eta}} \mathbb{E}_{\boldsymbol{y}\sim p_{Y_{\text{real}}}} \left[\| s_{\boldsymbol{\phi}}(\boldsymbol{y}) - s_{\boldsymbol{\theta}}(\boldsymbol{y};\boldsymbol{\eta}_k) \|^2 \right]. \quad (89)$$

该方法能够仅依据输出观测对复杂的非线性有向无环图模型进行校准，同时利用了为信息最大化而开发的相同基于得分函数的工具包。

备注 8（校准的无监督特性）。传统校准方法通常需要配对输入输出数据 $\{(\boldsymbol{x}^{(i)},\ \boldsymbol{y}^{(i)}_{\text{真实}})\}$，以最小化均方误差（MSE）等逐元素误差指标。从复杂物理系统中收集此类完全同步的配对数据，在操作上可能很困难或成本高昂。我们所采用的基于得分函数的方法有一个显著优势，即其无监督特性；它只需要来自真实系统的非配对观测样本 $\{\boldsymbol{y}^{(i)}_{\text{真实}}\}$（假设输入分布 p_X 相同）。这种灵活性显著拓宽了该框架在现实场景中的适用范围，尤其是在只能监测输出的情况下。

IX. 结论

本文为由有向无环图表示的复杂随机系统建立了端到端互信息最大化的一般框架。核心贡献是推导出广义信息梯度公式（定理 2），该公式统一并扩展了 [2]中的结果，使其适用于涉及分支、合并和非线性变换的任意网络拓扑结构。我们证明，可以将现代自动微分工具（尤其是向量-雅可比积）与得分匹配技术相结合，高效地计算这一梯度。

我们框架的一项关键优势在于其能够以完全黑箱的方式运行。正如我们在多路径线性有向无环图和非线性标量信道上的数值实验中所验证的那样，所提出的估计器利用通过去噪得分匹配训练并经斯坦校准细化的神经得分模型，在无需了解底层似然函数的情况下，仍能忠实地再现真实的信息梯度。在全局约束下成功优化网络参数，进一步证实了所提出的迭代得分学习策略对分布漂移的鲁棒性。

除了最大化之外，我们还展示了基于得分函数的视角能够自然地扩展到无监督模型校准。所提出的数字孪生校准方法仅利用输出样本，最小化物理系统与其模型之间的Fisher散度，为使仿真环境与现实相一致提供了一种实用方案。

虽然梯度估计的准确性自然取决于学习得到的评分模型的保真度，但所推导的定理本身提供了一个严谨且通用的接口，将互信息与得分函数联系起来。随着得分估计技术不断进步，该框架的适用性和精确度也将自然提升。

通过弥合基础信息论恒等式与可扩展深度学习计算之间的鸿沟，本研究提供了一套通用工具，可用于设计下一代信息系统，涵盖从语义通信到分布式传感器网络等领域；在这些领域中，优化复杂且未建模信道中的信息流至关重要。

未来工作包括将这一框架应用于率失真分析和分布式信源编码。尽管本文主要关注最大化信道编码中的互信息，但所提出的信息梯度框架本质上具有对偶性，同样适用于最小化问题。未来一个有前景的方向是将这一框架应用于率失真分析和分布式信源编码。通过将有向无环图解释为测试信道（编码器），并在失真约束下最小化互信息，我们的方法可以作为一种数据驱动的求解器，用于刻画复杂传感器网络的可达速率区域，从而补充本文讨论的信道容量问题。

致谢

本研究得到日本科学技术振兴机构（JST）CRONOS 项目的资助，资助编号为 JPMJCS25N5。

参考文献

[1] T. M. Cover and J. A. Thomas，《信息论基础》，第2版。美国新泽西州霍博肯：威利出版社，2006。ISBN：978-0-471-24195-9。[2] T. Wadayama，“非线性高斯信道的信息梯度及其在面向任务的通信中的应用”，https://arxiv.org/abs/2510.20179，2025。[3] P. Vincent，“分数匹配与去噪自编码器之间的联系”，《神经计算》，第23卷，第7期，第1661–1674页，2011年7月。doi：10.1162/NECO a 00142。[4] Y. Song和S. Ermon，“通过估计数据分布的梯度进行生成建模”，载于《神经信息处理系统进展》（神经信息处理系统大会），2019年，第11895–11907页。arXiv：1907.05600。[5] Y. Song、J. Sohl-Dickstein、D. P. Kingma、A. Kumar、S. Ermon和B. Poole，“通过随机微分方程进行基于分数的生成建模”，载于《神经信息处理系统进展》（神经信息处理系统大会），第34卷，2021年，第30470–30480页。arXiv：2011.13456。[6] A. G. Baydin、B. A. Pearlmutter、A. A. Radul和J. M. Siskind，“机器学习中的自动微分：综述”，《机器学习研究期刊》，第18卷，第5595–5637页，2018年。[7] Y.-I. Moon、B. Rajagopalan和U. Lall，“使用核密度估计器估计互信息”，《物理评论E》，第52卷，第3期，第2318–2321页，1995年。doi：10.1103/PhysRevE.52.2318。[8] A. Kraskov、H. St¨ogbauer和P. Grassberger，“互信息估计”，《物理评论E》，第69卷，第6期，第066138页，2004年6月。doi：10.1103/PhysRevE.69.066138。[9] T. Wadayama，“通过分数到费希尔桥估计非线性高斯噪声信道的互信息”，https://arxiv.org/abs/2510.05496，2025。[10] A. Hyv¨arinen，“通过分数匹配估计未归一化统计模型”，《机器学习研究期刊》，第6卷，709，2005 695pp. – 。[11] D. G¨und¨uz、Z. Qin、I.˜naki Estella Aguerri、H. S. Dhillon、Z. Yang、A. Yener、K.-K. Wong和C.-B. Chae，“超越比特传输：背景、语义与面向任务的通信”，《IEEE通信选题领域期刊》，第41卷，第1期，第5–41页，2023年1月。，doi：10.1109/JSAC.2022.3223408。[12] N. Tishby、F. C. Pereira和W. Bialek，“信息瓶颈方法”，载于第37届Allerton通信、控制与计算年度会议论文集，美国伊利诺伊州蒙蒂塞洛，1999年，第368–377页。[13] R. A. Fisher，“理论统计学的数学基础”，《伦敦皇家学会哲学汇刊：A辑》，第222卷，第309–368页，1922年。doi：10.1098/rsta.1922.0009。[14] A. J. Stam，“Fisher信息量与Shannon信息量满足的一些不等式”，《信息与控制》，第2卷，第2期，第101–112页，1959年6月。doi：10.1016/S0019-9958(59)90348-1。[15] L. D. Brown，“受Cram´er-Rao不等式启发的中心极限定理证明”，《统计与决策》，第1卷，第58–68页，1982年。[16] A. R. Barron，“熵与中心极限定理”，《概率年刊》，第14卷，第1期，第336–342页，1986年1月。doi：10.1214/aop/1176992632。[17] D. Guo、S. Shamai（Shitz）和S. Verd´u，“高斯信道中的互信息与最小均方误差”，《IEEE信息论汇刊》，第51卷，第4期，第1261–1282页，2005年4月。doi：10.1109/TIT.2005.844072。[18] Y. Lipman、R. T. Q. Chen、H. Ben-Hamu、M. Nickel和M. Le，“生成建模的流匹配”，载于国际学习表征会议论文集（国际学习表征会议），2023年。 [在线]。可从以下网址获取：arXiv:2210.02747。[19] X. Liu、C. Gong和Q. Liu，“流动笔直且快速：通过整流流学习生成和迁移数据”，载于国际学习表征会议论文集（国际学习表征会议），2023年。 [在线]。可从以下网址获取：arXiv:2209.03003。[20] D. P. Kingma和M. Welling，“自编码变分贝叶斯”，载于国际学习表征会议论文集（国际学习表征会议），2014年。 [在线]。可从以下网址获取：arXiv:1312.6114。