

通过信息几何和量子度量重新思考大语言模型训练

Riccardo Di Sipio

Dayforce, HCM

riccardo.disipio@dayforce.com

2025 年 6 月 28 日

摘要

大型语言模型 (LLMs) 的优化在具有非欧几里得结构的高维参数空间中展开。信息几何使用费雪信息度量来构建这一景观，使通过自然梯度下降 [1, 2] 进行更原则的学习成为可能。尽管这通常不切实际 [10]，但这种几何视角澄清了诸如尖锐最小值、泛化和观察到的缩放定律 [7] 等现象。我们主张，曲率感知方法加深了我们对 LLM 训练的理解。最后，基于 Fubini – Study 度量和量子费雪信息 [3, 11]，我们推测量子类比，暗示在量子增强系统中实现高效优化。

1 介绍

大型语言模型 (LLMs) 的优化揭示了显著的成功和深刻的理论谜题。随着这些模型规模的扩大，它们表现出更平滑的损失景观、更好的泛化能力以及经验上可预测的表现。这些趋势被形式化为缩放定律，将计算量、数据量和参数数量与损失 [6, 7] 联系起来。然而，导致这些模式的原因尚未完全理解。是什么决定了大规模下的损失景观的形状？为什么某些架构比其他架构更高效地收敛？我们的优化工具是否适合学习过程的真实几何结构？

在本文中，我们探讨了一个推测但有结构的假设：大型语言模型训练动态中的某些特征——特别是涉及曲率、收敛性和泛化性的那些——可能通过量子几何的视角得到更清晰的理解。虽然量子力学和深度学习操作于不同的领域，但两者都描述了根据变分原理演化且对局部曲率敏感的高维系统。量子系统在配备有富比尼—斯图迪度量的流形中演化，这是一种定义纯量子状态间距离的黎曼几何。该度量诱导出量子费舍尔信息矩阵 (QFI)，它测量参数变化的局部灵敏度，并编码比其经典对应物 [5, 9] 更锐利、更具表现力的几何结构。

相比之下，经典深度学习仅在显式近似的情况下才使用感知曲率的方法如费舍尔信息矩阵，例如自然梯度下降 [2]。但这些方法在大规模应用中很少实用：计算或逆费雪矩阵在高维情况下是计算代价高昂且不稳定的。然而，量子系统实现其内在的优化几何结构。它们并不近似曲率；而是嵌入曲率。这种不对称性突显了一个关键差异：经典模型难以访问二阶结构，而量子系统本质上是在一个丰富弯曲的流形上进行优化。

从这个角度来看，训练一个大型语言模型与量子系统塌缩到某个测量结果并无不同。梯度下降作为一种有噪声的迭代投影过程，将模型导向更低损失的状态，这一过程受到随机数据暴露和局部

曲率的影响。类比于波函数坍缩、能量最小化和叠加原理，这些新的词汇为描述大型模型如何在参数空间中移动提供了新方式——或许也为改进它们的训练提供了一些新工具。

本文的其余部分将详细阐述这一论点。第 2 节介绍了优化、费雪几何和量子态空间的相关概念。第 3 节在经典和量子机器学习文献的大背景下定位我们的工作。第 4 节讨论了对优化、缩放行为和算法设计的影响——并概述了未来的研究方向。我们通过反思隐喻作为连接学科的工具的价值，以及量子几何可能教会我们训练更好模型的可能性来结束本文。

2 理论背景

本节介绍了优化大语言模型的关键思想以及量子力学的基础几何结构。尽管这些领域在应用和背景上有所不同，但它们共享了大量的数学框架。我们的目标不是形式化每一个元素，而是为后续章节探讨的类比建立直觉。

2.1 优化与几何在大语言模型中的应用

现代大型语言模型（LLMs）依赖于一阶优化技术，如随机梯度下降（SGD）及其变体 [4, 8]。虽然这些方法在高维设置中有效，但它们操作的参数空间可能无法用欧几里得几何很好地表示。相反，来自信息几何的工具为底层优化景观 [1, 2] 提供了更丰富的图景。

这种行为邀请了一种几何解释：训练神经网络可以被视为穿越由模型参数和损失函数塑造的高维流形。费雪信息矩阵在理解这一几何结构中扮演着核心角色。它定义了参数空间上的局部黎曼度量，捕捉到损失面的曲率以及模型输出分布对参数微小变化的敏感性。尽管费雪信息通常从概率或统计的角度引入，但其几何解释——形式上作为统计流形上的度量张量——将其直接与物理学中的变分原理联系起来。这一联系为即将到来的类比奠定了基础。

这种平行性进一步延伸。在广义相对论中，引力不是传统意义上的力，而是时空曲率的可见效应——物体之所以如此运动是因为它们所处的几何结构所致。在深度学习中，特别是在大语言模型（LLM）中，观察到了类似的效果：语义相关的词语似乎在向量空间中“吸引”彼此。然而这种吸引力并非根本性的——它是更深层次结构的结果：由神经网络建模的高维概率分布。嵌入空间反映了该分布的曲率，就像引力轨迹反映时空的曲率一样。

2.2 概率空间的几何与费舍尔信息

给定一个参数概率分布 $p(x; \theta)$ ，参数的空间 θ 形成一个统计流形。为了理解这个空间上的曲率，我们从欧几里得距离开始：

$$d(x, y) = \sqrt{(x^i - y^i)(x^j - y^j)}$$

相比之下，弯曲统计流形的几何结构由度量张量定义：

$$g_{ij}(\theta) = \mathbb{E} \left[\frac{\partial \log p(x; \theta)}{\partial \theta^i} \frac{\partial \log p(x; \theta)}{\partial \theta^j} \right]$$

这是费雪信息矩阵，它通过测量两个无限接近的分布之间的可区分性来捕捉局部曲率。

对于流形上的一个参数曲线 $\theta(t)$ ，其无穷小长度的平方由下式给出：

$$ds^2 = g_{ij}(\theta) d\theta^i d\theta^j$$

此构造与黎曼几何密切相关。正如广义相对论将引力描述为时空中的能量和质量引起的曲率，信息几何则将学习景观描述为模型的信息结构引起的曲率。在这个类比中，费雪信息矩阵的作用类似于广义相对论中的度量张量：它决定了如何测量距离、如何计算测地线，并最终决定了流形上的优化过程。

这种平行性进一步延伸。在广义相对论中，引力不是传统意义上的力，而是时空弯曲的可见效果——物体之所以以特定方式运动是因为它们所处的几何结构。在深度学习，特别是在大语言模型（LLM）中，可以在词嵌入的几何结构中观察到类似的效果：语义相关的词语似乎会在向量空间中“吸引”彼此。然而这种吸引力并非根本性的——它是更深层次结构的结果：由神经网络建模的高维概率分布。嵌入空间反映了该分布的曲率，就像引力轨迹反映时空的曲率一样。

如图 1 所示，统计流形的局部几何可以通过每个点处的切空间来描述，在该切空间中，切向量对应于参数的无穷小变化。度量张量——在我们的案例中由费雪信息导出——定义了该空间中的内积，从而可以计算距离和梯度。

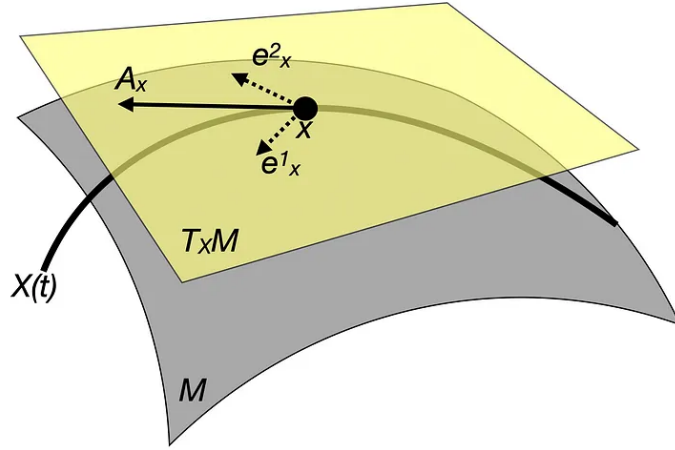


图 1: 流形上的曲线图示切空间和信息几何概念。一条曲线 $X(t)$ 在流形 M 上通过点 x 。在 x 处的切空间，记为 $T_x M$ ，被描绘成一个平面。切向量 A_x 被显示出来，并用基底 $\{e^1_x, e^2_x\}$ 表示。从对数似然函数 $\log p(x; \theta)$ 导出的切向量定义了信息几何的局部结构。

2.3 基于梯度的优化和曲率意识

在欧几里得空间中，梯度下降遵循标量损失 $L(\theta)$ 的负梯度：

$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla L$$

然而，在一个曲流形上，梯度必须考虑度量。 L 的梯度变为：

$$\nabla L = g^{ij} \frac{\partial L}{\partial \theta^j}$$

这导致了自然梯度下降，它使用费雪信息的逆来重新缩放更新方向：

$$\theta^{(t+1)} = \theta^{(t)} - \eta F^{-1} \nabla_{\theta} L$$

虽然这种方法在理论上很有吸引力，但对于大规模模型而言，计算和求逆 F 通常是不切实际的——这是我们之后将重新讨论的一个限制。

2.4 量子几何一瞥

量子力学描述物理系统不是通过点粒子，而是通过复希尔伯特空间中的状态向量。与经典系统不同，在经典系统中概率被赋予确定的结果，在量子系统中存在叠加态，并且它们的动力学受线性、么正变换的支配。这个空间的几何结构不是欧几里得的，而是射影的：全局相位在物理上是无关紧要的，因此真正的配置空间是希尔伯特空间中的光线集。这自然导致了 Fubini-Study 度量，这是一种定义在纯量子态空间上的黎曼度量。它衡量的是附近量子态的区别性，考虑到了复结构和相位不变性。至关重要是，正是这种度量诱导出了量子费舍尔信息矩阵 (QFI)，它是经典费舍尔矩阵的量子对应物 [3]。QFI 决定了量子系统对参数变化的敏感程度——这正好是也影响深度学习动态的那种曲率测量类型。而经典费舍尔信息捕获了似然函数对参数变异的灵敏度，QFI 则捕捉到了在无穷小参数变动下量子态变化的速度，嵌入了一种更丰富的几何结构。从这个意义上说，量子流形是“更尖锐”的——它的局部几何更加弯曲——这意味着优化这种流形可能遵循更陡峭、更有方向性的梯度。

2.5 缩放定律与经典几何的极限

近年来，在深度学习领域最引人入胜的发现之一是缩放定律的出现：模型大小、数据集大小、计算预算和性能之间可预测的关系 [4,5]。这些规律表明，达到一定限度之前，更大的模型在更多的数据上使用更多计算进行训练会系统性地降低损失。然而，它们也表现出递减收益——每次计算翻倍带来的性能提升越来越小。这引发了一个推测：这些限制是根本性的还是仅仅经典架构和训练方法的产物？如果由经典费舍尔信息诱导的几何结构对优化路径施加了隐含约束，那么一种更丰富的几何结构——例如量子流形——可能会突破这些缩放上限。受量子启发的训练原则上可以实现参数空间中更有效的探索、局部极小值之间更大的分离以及更陡峭的下降路径。

3 相关工作与理论背景

理解大型神经网络的训练动力学一直是多个学科的研究热点，涵盖统计物理、微分几何到信息论和复杂性科学等领域。同样地，量子机器学习 (QML) 领域的探索也迅速增长，研究者们要么试图在经典模型中模拟量子行为，要么利用量子硬件加速学习过程。本节我们简要回顾与讨论相交的主要线索。

3.1 几何与深度学习中的曲率问题等等

几项基础性工作探讨了深度神经网络的几何性质。Amari 的信息几何 [1] 工作引入了 Fisher 信息矩阵作为参数空间上的自然黎曼度量，为自然梯度下降提供了基础——这是一种考虑统计模型流形结构的优化方法。最近，Martens [10] 及其他人在可扩展深度学习背景下重新审视了这一思想，展示了如何利用曲率感知的方法改善在高维、病态损失景观中的收敛性。Pennington 和 Bahri [?] 利用随机矩阵理论工具对广泛神经网络的损失面有了重要贡献。他们的工作揭示了曲率——通过 Hessian 或 Fisher 矩阵捕获——不仅指导优化，还与泛化能力和模式连通性密切相关。这些研究表明，优化流形的几何结构至关重要，Fisher 信息是塑造这种结构的关键因素。

3.2 量子信息几何与量子费舍尔信息

在量子理论中，量子费舍尔信息矩阵 (QFI) 扮演着类似但更为丰富的角色。它量化了邻近量子态的可区分性，并通过量子克拉美- Rao 不等式 [3, 11] 对参数估计提供了界限。在量子背景下，克拉美- Rao 界限通过量子费舍尔信息进行了推广，为无偏估计量的方差提供了一个下界。与它的经典对应物不同，QFI 源自射影希尔伯特空间上的富比尼-施泰德度量，反映了量子态的真实几何结构。

在量子信息几何中，Fubini-Study 度量定义了纯态的射影希尔伯特空间上的黎曼结构。给定由 θ 参数化的归一化量子态 $|\psi(\theta)\rangle$ ，Fubini-Study 线元素为：

$$ds^2 = \langle d\psi | d\psi \rangle - |\langle \psi | d\psi \rangle|^2$$

该度量捕获了附近量子态的可区分性，并在经典信息几何中扮演着类似于费舍尔信息度量的角色。

量子费舍尔信息 (QFI) 将这一概念扩展到由密度矩阵 $\rho(\theta)$ 表示的混合态。QFI 矩阵定义为：

$$F_{ij} = \text{Tr} [\rho(\theta) L_i L_j]$$

其中 L_i 是对称对数导数 (SLD)，由以下方式隐式定义：

$$\frac{\partial \rho(\theta)}{\partial \theta^i} = \frac{1}{2} (\rho L_i + L_i \rho)$$

这种表述自然地推广了经典的费舍尔度量，并构成了量子统计模型几何的基础。在我们的推测性类比中，我们建议那些根据富比尼-施泰德几何进行优化的系统可能能够访问参数空间中的更有效路径，实际上体现了一种没有近似的自然梯度下降形式。

量子 Fisher 信息已被用于研究量子相变、纠缠结构以及变分量子算法——在这些领域中，对参数变化的敏感性起着关键作用 [3, 11]。最近的研究探讨了 QFI 在优化上下文中的影响，表明其更陡峭的曲率可能有助于变分量子电路 (VQCs) 中的收敛 [?, ?]。这些见解为“曲率丰富的度量可以改进优化”的想法提供了支持，并为本文中所作的类比提供了理论基础。

3.3 量子启发式和混合模型

一项不断增长的研究探讨了使用量子启发的经典模型或混合量子-经典架构的应用。例如，量子核 [?] 和张量网络模型 [?] 已被提出作为在经典硬件上模拟量子纠缠的方法。其他人则研究了参数化量子电路的信息内容及其相对于深度神经网络的表达能力 [?]。更广泛地说，量子机器学习 (QML) 领域探索了量子系统是否能在训练、表示或推理方面提供计算优势。虽然实际的量子优势仍然难以捉摸，但该领域仍在不断演变，并且一些基准测试表明，在理想条件下，量子模型可能以更为紧凑的方式表示某些函数，或者用较少的查询来学习它们。这些发展支持这样的主张：量子几何不仅仅是比喻性的，而是有可能重塑学习理论的概念基础——即使是在经典环境中。

4 讨论与未来方向

本文中探讨的类比不仅仅是美学上的；它们对如何思考深度学习优化、模型扩展和几何感知学习提出了实际和理论上的意义。虽然量子力学与机器学习之间的联系目前仍主要停留在概念层面，但它为未来的研究开辟了几条有希望的道路。

4.1 重新考虑几何在优化中的作用

由经典费舍尔信息矩阵诱导的几何长期以来被认为是理解模型行为的关键。算法如自然梯度下降利用这一洞察，在参数流形上遵循测地线，使优化步骤适应局部曲率。然而，在实践中，这些方法很少大规模使用：计算或求逆费舍尔矩阵对于现代架构来说成本高昂。相比之下，量子系统本质上在弯曲的流形上演化，其几何结构由富比尼-斯图迪度量控制。该度量定义了投影希尔伯特空间中量子态之间的最短路径，并且与其相关的量子费舍尔信息矩阵 (QFI) 自然地系统的结构中浮现出来。在这个意义上，量子系统“内置”了优化的几何——仅仅作为其演化的结果而表示局部曲率，几乎没有或没有近似。这种对比令人印象深刻：经典深度学习必须通过昂贵的计算来估计或近似信息几何，而量子系统则内在地实现了它。这意味着在量子流形上的优化可能本质上遵循更知情、适应曲率的轨迹——无需额外的算法机制。这提出了一个有趣的可能性，即我们在嵌入空间中观察到的语义相似性类似于物理学中的重力：不是一种原始力量，而是底层曲率的结果。在这种观点下，神经网络不是“学习意义”，而是在参数空间中诱导几何结构；我们与智能或一致性相关的效果作为这个弯曲流形内的测地线对齐效果出现。

4.2 缩放律的影响

如第 2 节所述，当前的缩放定律显示出随着模型规模或计算量的增加，收益递减。这引发了一个问题：这些幂律趋势是模型空间经典几何的内在属性，还是在不同的假设下可能会发生改变？量子几何视角为替代的缩放机制打开了大门。如果类似于 QFI 的曲率能够更有效地区分函数类别，或者通过锐化梯度轨迹来实现更快的收敛，那么量子增强模型可能偏离经典缩放趋势，尤其是在数据稀疏或对噪声敏感的机制中。检验这个假设需要对量子优化几何进行模拟，或者使用近似 QFI 行为的混合模型进行实验——这两个方向都是可行的研究方向。

4.3 算法设计的灵感来源

从这一观察中得出的实际结论是，受量子几何启发的优化算法可能具有结构上的优势。在经典设置中，基于费雪信息矩阵的曲率感知方法常常因为计算成本而被放弃。但如果量子系统自然编码了菲布尼-斯特迪几何——实质上将二阶信息嵌入到它们的演化过程中——那么训练量子模型就等同于默认执行信息几何优化。这一见解促使在经典模型中开发量子启发式的近似方法：架构或参数化方案模仿遵循菲布尼-斯特迪度量系统演化的行为。或者，可以设计混合量子-经典系统让量子组件处理优化景观中的几何复杂部分，而经典组件则进行更新规则或评估。在这两种情况下，核心理念依然不变：经典模型难以近似曲率的地方，量子模型则生活在那里。

4.4 限制和注意事项

当然，这种类比是有局限性的。神经网络不是量子系统。它们的参数遵循不同的物理定律演化，尽管量子理论的数学可能提供强大的概念工具，但必须谨慎应用。这里的目标不是声称等同，而是提出一个富有成效的比喻——一种可以增强我们对难以分析的系统理解的视角。此外，量子机器学习的实际挑战——硬件脆弱性、有限的量子比特数量和贫瘠高原——仍然没有解决。任何来自量子几何的好处都必须与这些限制进行权衡。

4.5 未来工作

几个研究方向由此产生：比较经典和量子损失景观几何形式的推导。基于 QFI 优化的经典模型模拟。用于理解参数空间中曲率差异的可视化工具。探索具有量子启发模块或初始化方案的混合 LLM。在 QFI 动机训练方案下的缩放行为的经验测试。将 LLM 优化视为在弯曲信息流形上的类似量子坍缩形式，我们获得了新的语言和工具来推理泛化、曲率和收敛性。无论这一路径是否能带来更好的模型，它都提供了一个独特跨学科的视角——一个融合了机器学习与物理学历史上最成功的抽象概念的视角。

5 结论

在本文中，我们探讨了大型语言模型 (LLMs) 训练动态与量子系统几何学之间的概念类比。通过将梯度下降与波函数坍缩、损失景观与能量势能以及经典和量子费雪信息进行关联，我们提出了一种视角，在这种视角下，深度学习优化被视为穿过高维信息流形的轨迹。这一观点的核心是几何的作用。理论上，经典学习算法可以从基于费雪信息矩阵的曲率感知方法中受益。然而，由于计算成本和不稳定性，这些方法很少在大规模使用。相比之下，量子系统根据 Fubini-Study 度量演变，这是一种自然黎曼几何，直接诱导出量子费雪信息矩阵。因此，量子系统内在地编码了二阶信息，允许它们以几乎或完全没有近似的方式表示并穿越曲率优化景观。这种区别暗示了一条可能的前进道路：量子增强或受量子启发的系统可能会提供一种根本不同的学习方式——不是通过增加复杂性来逼近曲率，而是存在于已经存在曲率的几何中。这是否会导致更好的泛化、更快的收敛或新

的缩放行为仍是一个开放问题。但这种类比不仅仅是美学上的；它促使我们重新思考如何理解学习本身。

参考文献

- [1] Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.
- [2] Shun-ichi Amari. *Information Geometry and Its Applications*. Springer, 2016.
- [3] Ingemar Bengtsson and Karol Życzkowski. *Geometry of Quantum States: An Introduction to Quantum Entanglement*. Cambridge University Press, 2nd edition, 2017.
- [4] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.
- [5] Carl W Helstrom. *Quantum Detection and Estimation Theory*. Academic Press, 1976.
- [6] Jordan Hoffmann et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- [7] Jared Kaplan, Sam McCandlish, Tom Henighan, et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- [8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [9] Jing Liu, Haidong Yuan, Xue-Mei Lu, and Xiaoguang Wang. Quantum fisher information matrix and multiparameter estimation. *Journal of Physics A: Mathematical and Theoretical*, 53(2):023001, 2020.
- [10] James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pages 2408–2417. PMLR, 2015.
- [11] Dénes Petz. *Introduction to Quantum Information Theory*, volume 795 of *Springer Lecture Notes in Physics*. Springer, 2011.