

Variance-Reducing Couplings for Random Features via Optimal Transport

论文作者及其单位:

Isaac Reid¹, Stratis Markou¹, Krzysztof Choromanski^{2,3}, Richard E. Turner¹, Adrian Weller^{1,4}

¹ University of Cambridge, ² Google DeepMind, ³ Columbia, ⁴ Alan Turing Institute

阅读报告作者:

刘行 PB22000150 E-mail: lx_0318@mail.ustc.edu.cn

摘要

随机特征 (Random Features, RFs) 是一种用于扩展机器学习中的核方法的主流技术, 它通过随机的蒙特卡洛估计替代精确的核函数计算. 这一思想被广泛应用于包括高效 Transformer (通过近似注意力机制) 和稀疏谱高斯过程 (通过近似协方差函数) 在内的多种模型. 通过加速这些估计的收敛 (这是一个方差缩减问题), 可以进一步提升效率. 我们通过最优传输 (OT) 这一统一视角来解决此问题, 寻找能改进定义在欧几里得空间和离散输入空间上的随机特征的耦合方式. 这些改进方案具有理论保证, 有时能带来显著的下游任务性能提升, 包括在图结构上的可扩展推断任务中. 我们得出了关于方差缩减作为优化范式其益处与局限性的令人惊讶的结论, 并表明在高效 Transformer 的注意力估计中, 耦合的其他特性也应被优化.

背景与动机

核方法在机器学习中广泛应用, 具有理论优雅, 表达能力强的优点. 然而, 其最大瓶颈在于计算开销高: Gram 矩阵的构建与求逆通常为 $O(N^3)$. 为了提升核方法的可扩展性, Rahimi 和 Recht 提出 "随机特征" 方法, 通过采样构造低维特征空间, 从而将核函数近似表示为显式内积. 该方法的关键是样本数量 m 要足够小以保证效率, 同时要收敛快以保证精度.

当前主要使用独立同分布采样 (i.i.d.) 或正交采样来降低方差, 但这些方式或过于通用, 未考虑核函数结构, 或难以扩展. 本文提出使用 "最优传输" 理论来设计更优的耦合结构, 从而在同样样本数量下获得更低估计方差.

方法

作者将 RF 核估计器 $\hat{k}(x, y) = \phi(x)^T \phi(y)$ 中的频率集合 $\{\omega_i\}_{i=1}^m$ 的联合分布作为变量, 引入最优传输理论, 在边缘分布固定的条件下寻找能最小化方差的 "耦合分布".

对常见的 RFF (随机傅里叶特征) 与 RLF (随机拉普拉斯特征), 文章提出 "成对范数耦合 (PNC-OT)" 策略:

- 保持方向正交,
- 对范数进行 comonotonic 耦合 (最大相关),
- 对任意 m , 在每对频率之间按照分位数进行匹配, 构造带依赖结构的频率矩阵.

对图上的 GRF (Graph Random Features), 文章提出通过匹配从节点 x 和 y 出发的游走路径的分位数长度, 实现离散空间上的耦合. 这种 "单调二分匹配" 策略能在图结构数据上有效降低估计方差.

实验结果

作者在多个 UCI 数据集和图结构数据集上进行了测试. 在 RFF 和 RLF 上, PNC-OT 显著降低了方差, 尤其在小样本数量 ($m=16, 32$) 时提升更明显.

在图上, 耦合游走长度也能降低方差并改善高斯过程的推理性能. 总体来看, OT 耦合不仅带来理论上的方差缩减, 还能转化为实际任务中的性能提升.

但值得注意的是, 作者也观察到某些情况下方差更低并不代表下游误差更小, 如在 transformer 注意力估计中, 正值约束比方差更重要. 因此作者指出, "方差不是唯一目标", 应根据任务优化耦合方式.

优劣性与创新点分析

优点:

- 引入最优传输理论作为统一工具, 解决不同输入空间下的随机特征方差缩减问题;
- 提出成对范数耦合策略, 方法简单, 可扩展, 有效;
- 将 OT 首次应用于图结构的随机游走耦合, 具创新性.

不足:

- 方法依赖已有的频率分布 (如高斯), 并未探讨如何联合优化基函数结构;
- 虽然降低方差, 但在某些任务 (如 transformer 注意力) 中性能提升不明显, 说明优化目标需更贴合任务;
- 仍需更广泛的实验来证明方法在其他核, 输入空间与任务中的适用性.

总结与思考

本文以 OT 理论为桥梁, 统一地解决了随机特征估计方差缩减问题, 既有数学美感, 又有实用价值. 特别是对图数据的处理拓展了 RF 方法的边界.

更重要的是, 作者不仅给出了优化方法, 也指出了 "过度追求方差缩减" 的局限性, 提出应**面向具体任务进行耦合设计**. 这为未来研究指出了更实际的方向:

- 任务驱动的耦合优化, 如 attention, 分类边界等;
- 与结构学习, 神经网络融合, 例如 transformer 中的 kernel learning;
- 多模态输入空间下的 RF 构造方法.

总之, 这是一篇兼具理论深度与工程可行性的论文, 为随机特征方法注入了新生命, 也为深度学习中核方法的可扩展性提供了新路径.