arXiv:2302.00787v1 [cs.LG] 1 Feb 2023

FAVOR#：通过新型正随机特征类实现的尖锐注意力核近似

Valerii Likhosherstov [*1] **Krzysztof Choromanski** [*2 3] **Avinava Dubey** [*3]
Frederick Liu [3] **Tamas Sarlos** [3] **Adrian Weller** [1 4]

摘要

高效近似由高斯核或 softmax 核诱导的线性算子，通常采用随机特征（RFs）来实现，这些特征能够对算子的输出进行无偏近似。这类算子广泛应用于核方法和高效 Transformer 等重要领域。我们提出了参数化、正值且非三角函数的随机特征，用于近似高斯核和 softmax 核。与传统的 RF 近似不同，这些新方法的参数可以通过优化来降低近似的方差，且最优参数可以用闭式形式表示。我们证明，所提出的方法在实践中能够降低方差（方差缩小至原来的 e^{10} 倍甚至更多），并且在核回归任务中优于以往方法。利用所提出的机制，我们还提出了 FAVOR#，一种用于 Transformer 中自注意力近似的方法。我们证明，FAVOR# 在语音建模和自然语言处理中优于其他随机特征方法。

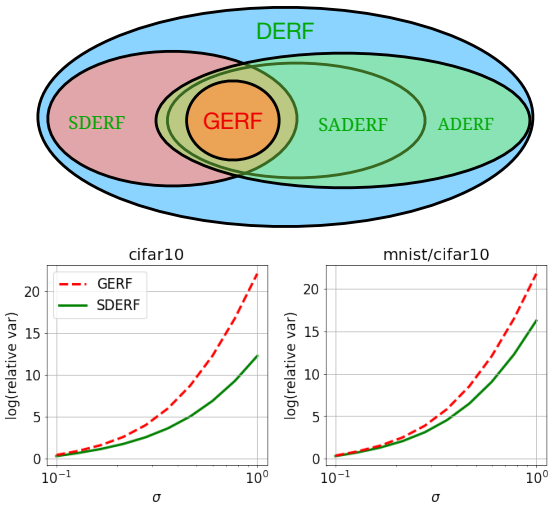

图 *1.*（上）我们提出的新型随机特征（绿色）的维恩图。（下）在从 CIFAR10 和 MNIST/CIFAR10 中采样的向量对上，不同随机特征映射相对方差的对数。新型随机特征映射 SDERF 相较于此前的最佳方法 GERF，始终能降低方差，最高可达 $\approx e^{10}$ 倍和 $\approx e^{5}$ 倍。图 2 扩展了此图，详情见第 5.1 节。

1. 引言

随机特征分解是将非线性核函数进行线性化的重要技术，并具有无偏性以及围绕真实核值的集中性等理论保证。线性化使得由核所诱导的算子规模的计算复杂度能够从二次复杂度显著降低至线性复杂度。该技术以 随机厨房水槽 （RKS）的名称出现，由（Rahimi & Recht, 2008a；2007；2008b）提出，并被应用于许多领域，例如核支持向量机（Sun et al., 2018；Le et al., 2013；Pennington et al., 2015；Avron et al., 2016）、降维（Dasgupta & Gupta，2003年；Ailon & Liberty，2013）、神经网络（Cho & Saul，2009；Gonon，2021年；Xie 等，2019；Han 等，2021年；Choromanski 等，2018）、函数到函数回归（Oliva 等，2015年）、核回归（Laparra 等，2015年；Avron 等，2017年）、非参数自适应控制（Boffi 等，2021年）、差分隐私机器学习算法（Chaudhuri 等，2011年）、算子值核（Minh，2016年）和半群核（Yang 等，2014年）。Li 等（2021年）、Yang 等（2012年）、Sutherland & Schneider（2015年）以及 Sriperumbudur & Szab{o}（2015年）对随机特征进行了深入的理论分析。

随机特征近期一个令人振奋的应用领域是可扩展的 Transformer 网络（Choromanski 等人，2021；2022；Chowdhury 等人，2022；Katharopoulos 等人，2020）。当序列较长时，该方法将自注意力矩阵近似为低秩矩阵。然而，随机厨房水槽（RKS）系列方法依赖于傅里叶变换，

[*]共同贡献 [1]剑桥大学 [2]哥伦比亚大学 [3]Google 研究院 [4]艾伦·图灵研究所。通信作者：Valerii Likhosherstov <vl304@cam.ac.uk>。

审稿中的预印本。

因此会产生正弦和余弦类型的随机特征。由于低秩矩阵中会出现负值，研究表明这类特征不适用于Transformer。Choromanski 等人（2021）提出了一种解决方案，即基于指数函数的正值随机特征（正随机特征，位置随机特征），并据此提出了用于自注意力近似的方法，称为 通过正交正随机特征实现快速注意力（*FAVOR+*）。Likhosherstov 等人（2022）通过谨慎选择指数中的线性组合参数，并采用所谓的齐次性启发式改进了这一方案；该启发式使得所有近似值都可以使用同一组参数。由此得到的随机特征称为广义指数随机特征（*GERFs*），相应的自注意力近似方法称为*FAVOR++*。

贡献： 本文通过提出稠密指数随机特征（*DERFs*），在正值随机特征的设计方面取得了突破。DERF 同时将位置随机特征和 GERF 纳入其特殊情形。不同于 GERF 中使用标量参数，DERF 依赖矩阵参数，并在指数内部采用稠密二次型。我们展示了如何高效地选择这些新型随机特征的参数，同时不损害整体的次二次复杂度。

更具体而言，我们的贡献如下：

1. 我们表明，Likhosherstov 等人（2022）提出的齐次性启发式实际上可以被视为平移对数方差目标函数的闭式最优解，而不仅仅是一种启发式方法。
2. 我们提出了 DERF 及其三种特殊实例：非对称 DERF（*ADERF*）、对称 DERF（*SDERF*）以及简化型 ADERF（*SADERF*）。所有这些实例都将 GERF 作为特殊情形（图 1，上方）。对于每种实例，我们都展示了如何高效地求得平移对数方差目标函数的闭式最优解。

3. 我们证明，实际应用中，我们提出的新变体相较于 GERF 和其他以往方法具有更低的方差（例如，如图 1 下方所示，方差最多改善 e^{10} 倍）。此外，我们还证明，在核回归和 Transformer 设置（语音建模和自然语言处理）中，DERF 的表现优于其他随机特征。我们将基于 DERF 的自注意力近似方法称为*FAVOR#*。

2. 预备知识

2.1. 缩放 Softmax 核与随机特征

通过缩放 *softmax* 核 $K^{(\alpha)}: \mathbb{R}^d \times \mathbb{R}^d \to (0, +\infty)$，其中 $\alpha \in \mathbb{R}$，我们将映射定义为 $K^{(\alpha)}(\mathbf{x}, \mathbf{y}) = \exp(\alpha\|\mathbf{x}\|^2 + \mathbf{x}^\top\mathbf{y} + \alpha\|\mathbf{y}\|^2)$，对所有 $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$，其中 $\|\cdot\|$ 是一个 L_2-范数。缩放 softmax 核的两个重要特例是：1）高斯核 $K^{(-1/2)}(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2/2)$；以及 2）softmax 核 $K^{(0)}(\mathbf{x}, \mathbf{y}) = \exp(\mathbf{x}^\top\mathbf{y})$。对于两组向量 $\mathcal{X} = \{\mathbf{x}^{(i)} \in \mathbb{R}^d\}_{i=1}^L$ 和 $\mathcal{Y} = \{\mathbf{y}^{(j)} \in \mathbb{R}^d\}_{j=1}^L$，我们用 $\mathcal{K}(\mathcal{X}, \mathcal{Y}) \in \mathbb{R}^{L\times L}$ 表示一个矩阵，其中$\mathcal{K}^{(\alpha)}(\mathcal{X}, \mathcal{Y})_{i,j} = K^{(\alpha)}(\mathbf{x}^{(i)}, \mathbf{y}^{(j)})$，对所有 $1 \le i, j \le L$均成立。

本文关注计算 $\mathcal{K}^{(\alpha)}(\mathcal{X}, \mathcal{Y})\mathbf{C}$ 的问题，其中 $\mathcal{X}$、$\mathcal{Y}$ 和矩阵 $\mathbf{C} \in \mathbb{R}^{L\times n}$ 作为输入给定。朴素方法需要 $O(L^2(d+n))$ 次运算来构造 $\mathcal{K}^{(\alpha)}(\mathcal{X}, \mathcal{Y})$ ($O(L^2 d)$)，并计算矩阵乘积 $\mathcal{K}^{(\alpha)}(\mathcal{X}, \mathcal{Y}) \times \mathbf{C}$($O(L^2 n)$)。作为替代，我们将使用蒙特卡洛方法高效近似 $\mathcal{K}^{(\alpha)}(\mathcal{X}, \mathcal{Y}) \times \mathbf{C}$，其依据是以下针对缩放 *softmax* 核的随机特征（*RFs*）的定义：

定义 2.1。对于缩放 softmax 核 $K^{(\alpha)}, \alpha \in \mathbb{R}$ 的随机特征，我们用三元组 $\mathcal{T} = \langle \nu, f^{(1)}, f^{(2)} \rangle$ 表示，其中 ν 是随机对象 $\boldsymbol{\omega} \in \Omega$ 上的概率分布，而 $f^{(i)}: \Omega \times \mathbb{R}^d \to \mathbb{R}$, $i \in \{1, 2\}$ 是满足以下条件的映射：对所有 $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$K^{(\alpha)}(\mathbf{x}, \mathbf{y}) = \mathbb{E}_\nu \left[f^{(1)}(\boldsymbol{\omega}, \mathbf{x}) f^{(2)}(\boldsymbol{\omega}, \mathbf{y}) \right]. \tag{1}$$

类型 (1) 的分解可用于对 $\mathcal{K}^{(\alpha)}(\mathcal{X}, \mathcal{Y})\mathbf{C}$ 进行高效的无偏近似。令 $\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(M)} \in \Omega$ 为从 ν 中独立同分布抽取的样本。定义矩阵 $\mathbf{P}, \mathbf{S} \in \mathbb{R}^{L\times M}$，其中对所有 $1 \le i, j \le L$,

$$\mathbf{P}_{i,:} = M^{-1/2}(f^{(1)}(\boldsymbol{\omega}^{(m)}, \mathbf{x}^{(i)}))_{m=1}^M, \tag{2}$$

$$\mathbf{S}_{j,:} = M^{-1/2}(f^{(2)}(\boldsymbol{\omega}^{(m)}, \mathbf{y}^{(j)}))_{m=1}^M, \tag{3}$$

其中 $\mathbf{P}_{i,:}$、$\mathbf{S}_{j,:} \in \mathbb{R}^d$ 是分别对应于 P、S 各行的列向量。根据 (1)，$\mathcal{K} = \mathbf{P}\mathbf{S}^\top$ 是基于 M 个样本对 $\mathcal{K}^{(\alpha)}(\mathcal{X}, \mathcal{Y})$ 的无偏蒙特卡洛 (MC) 近似。该近似的方差 $\mathrm{Var}\,\widehat{\mathcal{K}}_{i,j} = M^{-1} \; {}_\nu f^{(1)}(\boldsymbol{\omega}^{(i)}) f^{(2)}(\boldsymbol{\omega}^{(j)})$

Var (, x) (, y) 与 M 成反比，因此 M 是计算量与精度之间的权衡参数。现在，$\mathcal{K}\mathbf{C}$ 是 $\mathcal{K}^{(\alpha)}(\mathcal{X}, \mathcal{Y})\mathbf{C}$ 的无偏近似，但 $\widehat{\mathcal{K}} = \mathbf{P}\mathbf{S}^{\top\frown}$ 是一个秩为 M 的矩阵，因此计算 $\mathcal{K}\mathbf{C}$ 的复杂度为 $O(LMn)$。假设对每个 $\boldsymbol{\omega}^{(m)}$ 进行采样并计算 $f^{(\cdot)}(\cdot, \cdot)$ 都是 $O(d)$ 级操作（通常如此），则预计算 P 和 S 需要 $O(LMd)$ 次计算，总计算复杂度为 $O(LM(d+n))$。选择 $M \ll L$ 后，与精确版本相比，我们可以显著减少计算量：$O(LM(d+n)) \ll O(L^2(d+n))$。

类型为 $\mathcal{K}^{(\alpha)}(\mathcal{X}, \mathcal{Y})\mathbf{C}$ 的运算，尤其是高斯核 $\alpha = -1/2$，出现在核支持向量机（Rahimi & Recht，2007年）、核回归（Nadaraya，1964年；Watson，1964年）以及物理学中的高斯变换（Yang 等，2003年）中。另一个重要应用近来已

在高效 Transformer 领域出现，并将在下一节中讨论（Choromanski 等，2021年）。

2.2. 高效 Transformer 的随机特征

随机特征在高效长序列 Transformer 领域中有着重要应用（Choromanski等，2021年）。Transformer依靠自注意力模块在序列各元素之间传递信息。若序列长度为 L，输入矩阵记为$\mathbf{Q}$，$\mathbf{K}$，$\mathbf{V}$ $\in \mathbb{R}^{L\times d}$（查询、键 和值），则自注意力的输出矩阵如下:

$$\mathbf{Y} = \mathrm{diag}(\mathcal{K}^{(0)}(\mathcal{X},\mathcal{Y})\mathbf{1}_L)^{-1}\mathcal{K}^{(0)}(\mathcal{X},\mathcal{Y})\mathbf{V} \in \mathbb{R}^{L\times d}, \quad (4)$$

其中，$\mathbf{1}_L \in \mathbb{R}^L$ 是一个全 1 向量，diag(·)返回一个大小为（$L \times L$）的对角矩阵，其对角线元素为参数；$\mathcal{X} = \{\mathbf{x}^{(i)} = d^{-1/4}\mathbf{Q}_{i,:} \in \mathbb{R}^d\}$，且$\mathcal{Y} = \{\mathbf{y}^{(j)} = d^{-1/4}\mathbf{K}_{j,:} \in \mathbb{R}^d\}$。因此，在式（4）中用 $\widehat{\mathcal{K}}$ 代替 $\mathcal{K}^{(0)}$（$\mathcal{X}$，$\mathcal{Y}$），可将计算复杂度从 O（L^2d）降至 O（LMd）（$n = d+1$）。diag($\mathcal{K}^{(0)}$（$\mathcal{X}$，$\mathcal{Y}$）$\mathbf{1}_L$）$^{-1}\mathcal{K}^{(0)}$（$\mathcal{X}$，$\mathcal{Y}$）是对 $d^{-1/2}\mathbf{Q}\mathbf{K}^\top$的各行执行*soft-max* 运算所得的结果。

2.3. softmax 核的现有随机特征

表示形式 (1) 并不唯一，对于单个 $K^{(\alpha)}$ 可以提出不同的随机特征。注意，如果 $\langle \nu, f^{(1)}\ f^{(2)}\rangle$是 $K^{(0)}$ 的随机特征，那么 $\langle \nu, \widehat{f}^{(1)}, \widehat{f}^{(2)}\rangle$ 就是 $K^{(\alpha)}$ 在 $\alpha \in \mathbb{R}$ 下的随机特征，其中 $\widehat{f}_{(k)}(\boldsymbol{\omega},\mathbf{x}) = \exp(\alpha\|\mathbf{x}\|^2)f^{(k)}(\boldsymbol{\omega},\mathbf{x})$。因此，下文不失一般性地聚焦于 softmax 核 $K^{(0)}$。

Choromanski 等人（2020年）提出在高效 Transformer 应用中使用来自（Rahimi & Recht，2007）的三角随机特征（*TrigRFs*）：

$$\Omega_{\mathrm{trig}} = \mathbb{R}^{d+1},\ \nu_{\mathrm{trig}} = \mathrm{Unif}([0,2\pi]) \times \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)^d, \quad (5)$$

$$f^{(1)}_{\mathrm{trig}}((\theta,\widetilde{\boldsymbol{\omega}}),\mathbf{x}) = \sqrt{2}\exp(\|\mathbf{x}\|^2/2)\cos(\widetilde{\boldsymbol{\omega}}^\top\mathbf{x}+\theta), \quad (6)$$

$$f^{(2)}_{\mathrm{trig}}((\theta,\widetilde{\boldsymbol{\omega}}),\mathbf{y}) = \sqrt{2}\exp(\|\mathbf{y}\|^2/2)\cos(-\widetilde{\boldsymbol{\omega}}^\top\mathbf{y}+\theta), \quad (7)$$

其中，$\boldsymbol{\omega} = (\theta,\ \widetilde{\boldsymbol{\omega}})$，Unif(·) 表示定义在参数集合上的均匀分布，$\mathcal{N}(\mathbf{0}_d,\ \mathbf{I}_d)$ 表示均值为 $\mathbf{0}_d$（由 d 个零组成的向量）、协方差矩阵为 $\mathbf{I}_d$（大小为 $d \times d$ 的恒等矩阵）的多元高斯分布。

高效注意力近似方法的下一代（Choromanski等，2021年）发现了三角随机特征（5-7）存在一个问题。注意力矩阵diag($\mathcal{K}^{(0)}$($\mathcal{X}$，$\mathcal{Y}$)$\mathbf{1}_L$)$^{-1}\mathcal{K}^{(0)}$($\mathcal{X}$，$\mathcal{Y}$)（来自公式（4））是 右随机 的，这意味着其元素均为非负数，且每行之和为 1，这是由归一化项 diag($\mathcal{K}^{(0)}$($\mathcal{X}$，$\mathcal{Y}$)$\mathbf{1}_L$)$^{-1}$ 保证的。然而，由于 $f^{(\cdot)}$ 三角函数 可以是任意实数，$\mathbf{P}$，$\mathbf{S}$（2-3）以及因此的 $\mathcal{K}$ 都可能取负值。因此，diag($\widehat{\mathcal{K}}\mathbf{1}_L$)$^{-1}\widehat{\mathcal{K}}$通常不是右随机的，且 $\mathcal{K}\mathbf{1}_L$ 的元素可能非常小，和/或取负值，导致求逆 diag($\widehat{\mathcal{K}}\mathbf{1}_L$)$^{-1}$ 时行为不稳定。因此，Choromanski 等人（2021）提出了一种新型的正随机特征（位置随机特征），其形式如下:

$$\Omega_{\mathrm{pos}} = \mathbb{R}^d, \quad \nu_{\mathrm{pos}} = \mathcal{N}(0,1)^d, \quad (8)$$

$$f^{(1)}_{\mathrm{pos}}(\boldsymbol{\omega},\mathbf{x}) = f^{(2)}_{\mathrm{pos}}(\boldsymbol{\omega},\mathbf{x}) = \exp(\boldsymbol{\omega}^\top\mathbf{x} - \|\mathbf{x}\|^2/2). \quad (9)$$

显然，这类 $f^{(\cdot)}_{\mathrm{pos}}$ 只取严格为正的值，从而产生右随机的 diag($\widehat{\mathcal{K}}\mathbf{1}_L$)$^{-1}\widehat{\mathcal{K}}$，并使 Transformer 的训练过程保持稳定。

Likhosherstov 等人（2022）扩展了位置随机特征，提出了用于 $K^{(0)}$的广义指数随机特征（*GERF*）[1]:

$$\Omega_{\mathrm{GE}} = \mathbb{R}^d, \quad \nu_{\mathrm{GE}} = \mathcal{N}(0,1)^d, \quad f^{(1)}_{\mathrm{GE}}(\boldsymbol{\omega},\mathbf{x}) = \quad (10)$$

$$= f^{(2)}_{\mathrm{GE}}(\boldsymbol{\omega},\mathbf{x}) = D\exp(A\|\boldsymbol{\omega}\|^2 + B\boldsymbol{\omega}^\top\mathbf{x} + C\|\mathbf{x}\|^2/2), \quad (11)$$

其中，A，B，C，D均为实数[2]，且满足:

$$1-8A>0,\ B=\sqrt{1-4A},\ C=-\frac{1}{2},\ D=(1-4A)^{\frac{d}{4}}.$$

Likhosherstov 等人（2022）通过 A 表示 B，C，D，并推导出（1）的方差的闭式公式:

$$\mathrm{Var}_{\nu_{\mathrm{GE}}} f^{(1)}_{\mathrm{GE}}(\boldsymbol{\omega},\mathbf{x}) f^{(2)}_{\mathrm{GE}}(\boldsymbol{\omega},\mathbf{y}) = e^{\mathcal{L}_{\mathrm{GE}}(A,\mathbf{x},\mathbf{y})} - K^{(0)}(\mathbf{x},\mathbf{y})^2,$$

$$\mathcal{L}_{\mathrm{GE}}(A,\mathbf{x},\mathbf{y}) = d\log\left(\frac{1-4A}{\sqrt{1-8A}}\right) + \frac{2(1-4A)}{1-8A} \quad (12)$$

$$\times\|\mathbf{x}+\mathbf{y}\|^2 - \|\mathbf{x}\|^2 - \|\mathbf{y}\|^2. \quad (13)$$

最小方差对应于最小化$\mathcal{L}$(*A*, x, y)，因为 $K^{(0)}$(x, y)^2 不依赖于 A。由于$\mathcal{L}$(*A*, x, y)针对单个 x、y 对定义，而非针对集合 $\mathcal{X}$、$\mathcal{Y}$定义，Likhosherstov 等人（2022）提出了一种齐次性启发式方法：他们将（12-13）中的 $\|\mathbf{x}+\mathbf{y}\|^2$、$\|\mathbf{x}\|^2$、$\|\mathbf{y}\|^2$ 分别替换为对 $\mathcal{X}$、$\mathcal{Y}$的平均值：$L^{-2}\sum_{i,\ j}\|\mathbf{x}^{(i)}+\mathbf{y}^{(j)}\|^2$、$L^{-1}\sum_i\|\mathbf{x}^{(i)}\|^2$ 和$L^{-1}\sum_j\|\mathbf{y}^{(j)}\|^2$。该启发式方法基于以下假设：$\{\mathbf{x}^{(i)}\}$ 和 $\{\mathbf{y}^{(j)}\}$是齐次的，且其统计量紧密集中在均值附近。完成此替换后，可以以闭式形式求出（12-13）关于 A 的最小值。

3. 稠密指数随机特征（DERF）

我们证明，齐次性启发式对应于一个特定的最小化问题。随后，我们提出 DERF，它推广了 GERF，并为该问题提供了更紧致的解。

[1] Likhosherstov 等人（2022）为 $K^{(-1/2)}$ 定义了这些随机特征，但我们使用上文提到的技巧将其改用于 $K^{(0)}$。

[2] Likhosherstov 等人（2022年）考虑了一种更广义的形式，其中 A，B，C，D为复数，并引入了额外参数 $s=\pm1$；然而，他们仅提出了用于 Transformer 应用的子族（10–11），其中 $s=1$。

3.1. GERF 最小化的目标函数

我们的首要贡献是证明，GERF 中采用的齐次性启发式实际上是某个优化问题的解析解。定义

$$\overline{\mathcal{L}}(\boldsymbol{\theta};\mathcal{X},\mathcal{Y},\mathcal{T}) = L^{-2}\sum_{1\le i,j\le L}\log(\mathrm{Var}_{\nu}[f^{(1)}(\boldsymbol{\omega},\mathbf{x}^{(i)}) \times f^{(2)}(\boldsymbol{\omega},\mathbf{y}^{(j)})] + K^{(0)}(\mathbf{x}^{(i)},\mathbf{y}^{(j)})^2), \quad (14)$$

其中 $\mathcal{T} = \langle \nu, f_{(1)}\, f^{(2)}\rangle$ 是核 $K^{(0)}$的随机特征，$\boldsymbol{\theta}$ 是其参数。(14) 是加上 $K^{(0)}(\mathbf{x}^{(i)}, \mathbf{y}^{(j)})^2$后的平均对数方差。(14) 能达到的最佳值是 $\log K^{(0)}(\mathbf{x}^{(i)}, \mathbf{y}^{(j)})$，这对应于所有方差 $\mathrm{Var} f_{(1)}(\boldsymbol{\omega}, \mathbf{x}^{(i)}) f^{(2)}(\boldsymbol{\omega}, \mathbf{y}^{(j)})$均为零，也就是说，随机特征能够精确估计核。因此，最小化 (14) 可以使 $\mathcal{X}$，$\mathcal{Y}$上的估计量更加精确。我们将损失函数 $\mathcal{L}(\boldsymbol{\theta}; \mathcal{X}, \mathcal{Y}, \mathcal{T})$ 称为平移对数方差目标函数。

若在 (14) 中采用 $\mathcal{T}_{\mathrm{GE}} = \langle \nu_{\mathrm{GE}}, f^{(1)}{}_{\mathrm{GE}}\ f^{(2)}{}_{\mathrm{GE}}\rangle$，则$\boldsymbol{\theta}_{\mathrm{GE}} = \{A, B, C, D\}$ 且 $\overline{\mathcal{L}}(\boldsymbol{\theta}_{\mathrm{GE}}; \mathcal{X}, \mathcal{Y}, \mathcal{T}_{\mathrm{GE}}) = L^{-2}\sum_{i,\,j}\mathcal{L}_{\mathrm{GE}}(A; \mathbf{x}^{(i)}, \mathbf{y}^{(j)})$。利用 (12-13)，可得：

$$\overline{\mathcal{L}}(\boldsymbol{\theta}_{\mathrm{GE}}; \mathcal{X}, \mathcal{Y}, \mathcal{T}_{\mathrm{GE}}) = d\log\left(\frac{1-4A}{\sqrt{1-8A}}\right) + \frac{2-8A}{1-8A} \times \frac{1}{L^2}\sum_{i,j}\|\mathbf{x}^{(i)}+\mathbf{y}^{(j)}\|^2 - \frac{1}{L}\sum_i\|\mathbf{x}^{(i)}\|^2 - \frac{1}{L}\sum_j\|\mathbf{y}^{(j)}\|^2.$$

也就是说，当用在 $\mathcal{X}, \mathcal{Y}$ 上计算得到的平均统计量替换 $\|\mathbf{x}+\mathbf{y}\|^2$,$\|\mathbf{x}\|^2$, $\|\mathbf{y}\|^2$ 时，$\mathcal{L}(A; \mathcal{X}, \mathcal{Y})$ 与 (12-13) 一致。因此，齐次性启发式不过是对 (14) 的最小化。虽然一般而言，不清楚如何$\widehat{\quad}$ $\widehat{\quad}$求出 $\mathrm{Var}\mathcal{K}$ 或 $\mathrm{Var}(\mathcal{K}\mathbf{C})$ 的闭式最优解，但 (14) 的全局最小值是可行的，并且可以在$O(1)$ 时间内计算出来。此外，Likhosherstov 等人（2022）表明，对 (14) 进行优化，在高效 Transformer 的大规模应用中能取得很好的结果。在下一节中，我们将介绍 GERF 的若干扩展，它们的目标都是以闭式形式最小化 (14)。

4. 探索 DERF

稠密指数随机特征（DERF）是 GERF 的一种扩展，其中标量 A, B, C 被稠密矩阵取代。DERF 可视为一种推广形式，其中包含此前介绍的类别作为特例。我们将 DERF 定义如下：$\Omega_{\mathrm{DE}} = \mathbb{R}^d$,$\nu_{\mathrm{DE}} = \mathcal{N}(0,1)^d$，且对于 $k \in \{1,2\}$:

$$f^{(k)}_{\mathrm{DE}}(\boldsymbol{\omega},\mathbf{x}) = D\exp(\boldsymbol{\omega}^\top\mathbf{A}\boldsymbol{\omega} + \boldsymbol{\omega}^\top\mathbf{B}^{(k)}\mathbf{x} + \mathbf{x}^\top\mathbf{C}^{(k)}\mathbf{x}),$$

其中 $\mathbf{B}^{(k)}$,$\mathbf{C}^{(k)} \in \mathbb{R}^{d\times d}$, $D \in \mathbb{R}$, $\mathbf{A} \in \mathbb{S}_d$（一个由 $d\times d$实对称矩阵组成的集合）。显然，参数为A, B, C, D 的 GERF 可以通过参数为$\mathbf{A} = A\mathbf{I}_d$,$\mathbf{B}^{(1)} = \mathbf{B}^{(2)} = B\mathbf{I}_d$, $\mathbf{C}^{(1)} = \mathbf{C}^{(2)} = C\mathbf{I}_d$ 的 DERF 表示，

D 保持不变。我们的第一个理论结果给出了 $\mathcal{T}_{\mathrm{DE}} = \langle \nu_{\mathrm{DE}}, f^{(1)}{}_{\mathrm{DE}}\ f_{(2)\,\mathrm{DE}}\rangle$ 成为有效随机特征的条件：**定理 4.1.** 设以下条件成立：

$$8\mathbf{A} \prec \mathbf{I}_d, \quad (\mathbf{B}^{(1)})^\top(\mathbf{I}_d - 4\mathbf{A})^{-1}\mathbf{B}^{(2)} = \mathbf{I}_d, \quad \mathbf{C}^{(k)} = -\frac{1}{2}(\mathbf{B}^{(k)})^\top(\mathbf{I}_d - 4\mathbf{A})^{-1}\mathbf{B}^{(k)}, \ D = \det(\mathbf{I}_d - 4\mathbf{A})^{1/4}$$

其中 $k \in \{1,2\}$。那么，$\mathcal{T}_{\mathrm{DE}}$ 是用于 $K^{(0)}$ 的随机特征，并且对于所有$\mathbf{x}$、$\mathbf{y} \in \mathbb{R}^d$:

$$\mathrm{Var}_{\nu_{\mathrm{DE}}} f^{(1)}_{\mathrm{DE}}(\boldsymbol{\omega},\mathbf{x}) f^{(2)}_{\mathrm{DE}}(\boldsymbol{\omega},\mathbf{y}) = D^4\det(\mathbf{I}_d - 8\mathbf{A})^{-1/2} \times \exp\Bigg(2\mathbf{x}^\top\left(\mathbf{C}^{(1)} + (\mathbf{B}^{(1)})^\top(\mathbf{I}_d - 8\mathbf{A})^{-1}\mathbf{B}^{(1)}\right)\mathbf{x} + 2\mathbf{y}^\top\left(\mathbf{C}^{(2)} + (\mathbf{B}^{(2)})^\top(\mathbf{I}_d - 8\mathbf{A})^{-1}\mathbf{B}^{(2)}\right)\mathbf{y} + 4\mathbf{x}^\top(\mathbf{B}^{(1)})^\top(\mathbf{I}_d - 8\mathbf{A})^{-1}\mathbf{B}^{(2)}\mathbf{y}\Bigg) - K^{(0)}(\mathbf{x},\mathbf{y})^2. \quad (15)$$

我们的最终目标是找到最优参数 $\mathbf{A}$、$\mathbf{B}^{(k)}$、$\mathbf{C}^{(k)}$ 和 D，使给定集合 $\mathcal{X}$、$\mathcal{Y}$时，$\mathcal{K}^{(0)}(\mathcal{X},\mathcal{Y})$的低秩近似方差最小化。我们的第一个观察是，可以假设 $\mathbf{A} \in \mathbb{D}_d$（一组 $d\times d$ 实对角矩阵）。事实上，任何对称矩阵 $\mathbf{A}$ 都可以表示为 $\mathbf{Q}\widetilde{\ }\mathbf{A}\mathbf{Q}^\top$，其中 $\mathbf{Q} \in \mathbb{O}_d$（一组正交矩阵 $\{\mathbf{Z} \in \mathbb{R}^{d\times d} | \mathbf{Z}^\top\mathbf{Z} = \mathbf{I}_d\}$），且 $\widetilde{\ }\mathbf{A} \in \mathbb{D}_d$。令 $\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$。那么，对于任意 $\mathbf{x} \in \mathbb{R}^d$、$k \in \{1,2\}$,

$$f^{(k)}_{\mathrm{DE}}(\boldsymbol{\omega},\mathbf{x}) = D\exp(\boldsymbol{\omega}^\top\mathbf{Q}\widetilde{\mathbf{A}}\mathbf{Q}^\top\boldsymbol{\omega} + \boldsymbol{\omega}^\top\mathbf{B}^{(k)}\mathbf{x} + \mathbf{x}^\top\mathbf{C}^{(k)}\mathbf{x}) = \exp(\widetilde{\boldsymbol{\omega}}^\top\widetilde{\mathbf{A}}\widetilde{\boldsymbol{\omega}} + \widetilde{\boldsymbol{\omega}}^\top\widetilde{\mathbf{B}}^{(k)}\mathbf{x} + \mathbf{x}^\top\mathbf{C}^{(k)}\mathbf{x}) = \widetilde{f}^{(k)}_{\mathrm{DE}}(\widetilde{\boldsymbol{\omega}},\mathbf{x}),$$

其中 $\widetilde{\ }\mathbf{B}^{(k)} = \mathbf{Q}^\top\mathbf{B}^{(k)}$，$\widetilde{\boldsymbol{\omega}} = \mathbf{Q}^\top\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$，因为分布 $\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ 是等距的，即具有旋转不变性，并且 $\widetilde{f}^{(k)}$ DE 是参数为 $\widetilde{\ }\mathbf{A}$、$\widetilde{\ }\mathbf{B}^{(k)}$、$\mathbf{C}^{(k)}$、$D$的 DERF。我们得出结论，对于任意 $\mathbf{A}$，$f^{(k)}\mathrm{DE}(\boldsymbol{\omega}, \mathbf{x})$ 都可以表示为具有 $\widetilde{\ }\mathbf{A} \in \mathbb{D}_d$ 的 DERF $\widetilde{f}^{(k)}\,\mathrm{DE}$。因此，以下我们不失一般性地仅考虑 $\mathbf{A} \in \mathbb{D}_d$。

由于 $\mathbf{B}^{(k)}$、$\mathbf{C}^{(k)}$ 通常是稠密矩阵，计算 $f^{(k)}\,\mathrm{DE}(\boldsymbol{\omega}, \mathbf{x})$ 需要 $O(d^2)$ 时间，这高于 TrigRF、位置随机特征和 GERF 的 $O(d)$ 复杂度。不过，P 和 S 矩阵（2-3）仍可在关于 L 的次二次时间内计算。为此，先对所有 $k \in \{1,2\}$、$1 \le m \le M$、$1 \le i, j \le L$ 预计算 $(\mathbf{B}^{(k)})^\top\boldsymbol{\omega}^{(m)}$、$\mathbf{C}^{(1)}\mathbf{x}^{(i)}$ 和 $\mathbf{C}^{(2)}\mathbf{y}^{(j)}$，耗时为 $O((M+L)d^2)$。随后，对所有 $1 \le i, j \le L$、$1 \le m \le M$ 计算 $f^{(1)}\,\mathrm{DE}(\boldsymbol{\omega}^{(m)}, \mathbf{x}^{(i)})$ 和 $f^{(2)}\,\mathrm{DE}(\boldsymbol{\omega}^{(m)}, \mathbf{y}^{(j)})$，需要 $O(LMd)$ 次运算。因此，构造（2-3）的复杂度为 $O(L(Md + d^2) + Md^2)$，对于 L 仍是次二次的。

我们的目标是针对 $\boldsymbol{\theta}_{\mathrm{DE}} = \{\mathbf{A}, \mathbf{B}^{(1)}, \mathbf{B}^{(2)}, \mathbf{C}^{(1)}, \mathbf{C}^{(2)}, D\}$，最小化 $\mathcal{L}(\boldsymbol{\theta}_{\mathrm{DE}}; \mathcal{X}, \mathcal{Y}, \mathcal{T}_{\mathrm{DE}})$。然而，我们发现

即使只有一对 $\mathbf{x}$，$\mathbf{y}$，也不清楚如何以闭式形式最小化方差 (15)。因此，下面我们考虑解析解可行的特殊情况。

4.1. 非对称稠密指数随机特征 atures

以与 $\mathcal{T}_{\mathrm{DE}}$ 相同的方式定义随机特征 $\mathcal{T}_{\mathrm{ADE}} = \langle \nu_{\mathrm{ADE}}, f^{(1)}$ ADE f(2) ADE$\rangle$，唯一的区别在于A $= A\mathrm{I}_d$，其中 $\lambda \in \mathbb{R}$。我们将这些随机特征称为非对称稠密指数随机特征（ADERF），因为一般而言 $f^{(1)}$ ADE $\neq f^{(2)}$ ADE。与DERF相比，ADERF唯一额外的限制是A $\in \mathbb{D}_d$ 的所有对角元素都相同。$\mathcal{T}_{\mathrm{ADE}}$ 的参数为 $\boldsymbol{\theta}_{\mathrm{ADE}} = \{A, \mathrm{B}^{(1)}$、$\mathrm{B}^{(2)}$、$\mathrm{C}^{(1)}$、$\mathrm{C}^{(2)}$、$D\}$。用 Θ_{ADE} 表示所有能够得到核 $K^{(0)}$的正确随机特征的 $\boldsymbol{\theta}_{\mathrm{ADE}}$集合，即满足A $= A\mathrm{I}_d$时定理 4.1中的条件。以下结果给出了$\mathcal{L}(\boldsymbol{\theta}_{\mathrm{ADE}}; \mathcal{X}, \mathcal{Y}, \mathcal{T}_{\mathrm{ADE}})$全局最小值的解析公式。定理中，我们使用对称矩阵的奇异值分解和特征值分解的概念（Trefethen & Bau，1997）（所有证明见附录）。

定理 *4.2.* 令$\mathcal{X}$ $= \{x^{(i)} \in \mathbb{R}^d\}_{i=1}^L$,$\mathcal{Y} = \{y^{(j)} \in \mathbb{R}^d\}_{j=1}^L$。令 $M^{(1)} = \frac{1}{L}\sum_{i=1}^L x^{(i)}(x^{(i)})^\top$, $M^{(2)} = \frac{1}{L}\sum_{j=1}^L y^{(j)}(y^{(j)})^\top$。假设 $M^{(1)}, M^{(2)} \in \mathbb{S}_d$ 均可逆。定义$\mu^{(3)} = d^{-1}L^{-2}\left(\sum_{i=1}^L x^{(i)}\right)^\top \left(\sum_{j=1}^L y^{(j)}\right) \in \mathbb{R}$。对于$k \in \{1,2\}$，令 $M^{(k)} = Q^{(k)}\mathbf{\Lambda}^{(k)}(Q^{(k)})^\top$ 为对称 $M^{(k)}$ 的特征值分解，其中 $Q^{(k)} \in \mathbb{O}_d$。根据定义，$\mathbf{\Lambda}^{(k)} \in \mathbb{D}_d$具有严格为正的对角值，因为 $M^{(k)} \succeq 0$ 且 $M^{(k)}$ 可逆。令 $U\mathbf{\Sigma}V^\top$ 为 $(\mathbf{\Lambda}^{(1)})^{\frac{1}{2}}(Q^{(1)})^\top Q^{(2)}(\mathbf{\Lambda}^{(2)})^{\frac{1}{2}}$ 的奇异值分解，其中 U、$V \in \mathbb{O}^d$, $\mathbf{\Sigma} \in \mathbb{D}_d$ 具有非负对角元素。

其中一个解 $\boldsymbol{\theta}^*_{\mathrm{ADE}} = \{A, \mathbf{B}^{(1)}, \mathbf{B}^{(2)}, \mathbf{C}^{(1)}, \mathbf{C}^{(2)}, D\}$ 如下。令 $\phi = 2d^{-1}\sum_{l=1}^d \mathbf{\Sigma}_{l,l} + 2\mu^{(3)}$ 并且，对于 $k \in \{1,2\}$,

$$
\begin{aligned}
A &= \frac{1}{16}\left(1 - 2\phi - \sqrt{(2\phi+1)^2 + 8\phi}\right),\\
\mathbf{B}^{(1)} &= \sqrt{1-4A}\,\mathbf{\Sigma}^{1/2}\mathbf{U}^\top(\mathbf{\Lambda}^{(1)})^{-1/2}(\mathbf{Q}^{(1)})^\top,\\
\mathbf{B}^{(2)} &= \sqrt{1-4A}\,\mathbf{\Sigma}^{-1/2}\mathbf{U}^\top(\mathbf{\Lambda}^{(1)})^{1/2}(\mathbf{Q}^{(1)})^\top,\\
\mathbf{C}^{(k)} &= -\frac{1}{2(1-4A)}(\mathbf{B}^{(k)})^\top\mathbf{B}^{(k)}, \quad D = (1-4A)^{d/4}.
\end{aligned}
$$

此外，我们有：

$$
\begin{aligned}
\overline{\mathcal{L}}(\boldsymbol{\theta}^*_{\mathrm{ADE}}; \mathcal{X}, \mathcal{Y}, \mathcal{T}_{\mathrm{ADE}}) = d\Bigg(&\log(1-4A) - \frac{1}{2}\log(1-8A)\\
&+2(1-8A)^{-1}\left(d^{-1}\sum_{l=1}^d \mathbf{\Sigma}_{l,l} + \mu^{(3)}\right) + 2\mu^{(3)}\Bigg). \qquad (16)
\end{aligned}
$$

定理 4.2 蕴含了一种高效求解 $\boldsymbol{\theta}^*_{\mathrm{ADE}}$ 的算法。具体而言，在 $\mathbf{M}^{(k)}, k \in \{1,2\}$ ($O(Ld^2)$ 时间内计算并在 $\mu^{(3)}$ ($O(Ld)$ 时间内完成。然后，进行矩阵分解，以在 $O(d^3)$ 时间内得到 $\mathbf{Q}^{(k)}, \mathbf{\Lambda}^{(k)}$, $k \in \{1,2\}$ 和 $\mathbf{U}$ $\mathbf{\Sigma}$, $\mathbf{V}$。此后，利用定理 4.2 中的公式，可在 $O(d^3)$ 时间内计算出全部 $A, \mathbf{B}^{(1)}, \mathbf{B}^{(2)}, \mathbf{C}^{(1)}, \mathbf{C}^{(2)}, D$。因此，该近似方案的总时间复杂度为 $O(L(Md + d^2) + Md^2 + d^3)$, 相对于 L 是次二次的，符合要求。

4.2. 对称稠密指数随机特征

以相同的方式定义 $\mathcal{T}_{\mathrm{SDE}} = \langle \nu_{\mathrm{SDE}}, f^{(1)}$ SDE $f^{(2)}$ SDE$\rangle$，唯一的区别是 $\mathrm{B}^{(1)} = \mathrm{B}^{(2)} = \mathrm{B}$。由定理 4.1 中的条件可立即得出，$\mathrm{C}^{(1)} = \mathrm{C}^{(2)} = \mathrm{C}$ 也成立。因此，$f^{(1)}$SDE $= f$(2) SDE，我们将这些随机特征称为对称稠密指数随机特征（SDERF）。$\mathcal{T}_{\mathrm{SDE}}$的参数为 $\boldsymbol{\theta}_{\mathrm{SDE}} = \{\mathrm{A}, \mathrm{B}, \mathrm{C}\ D\}$。用$\Theta_{\mathrm{SDE}}$ 表示所有能得到正确核随机特征的 $\boldsymbol{\theta}_{\mathrm{SDE}}$ 的集合，即满足定理 4.1 中条件且 $\mathrm{B}^{(k)} = \mathrm{B}$、$\mathrm{C}^{(k)} = \mathrm{C}$ 的 $k \in \{1,2\}$。下面的定理给出了 $\mathcal{L}(\boldsymbol{\theta}_{\mathrm{SDE}}; \mathcal{X}, \mathcal{Y}, \mathcal{T}_{\mathrm{SDE}})$ 全局最小值的解析解。

定理 4 3 $\mathcal{X} = \{^{(i)}_{(1)} \in \mathbb{R}^d\}_{i=1}^L$ $\mathcal{Y} = \{^{(j)}_{(2)} \in \mathbb{R}^d\}_{j=1}^L$。令 $\mathbf{x}$、$\mathbf{y}$ 以及 $\mathbf{M}$、$\mathbf{M}$ 如定理 *4.2* 中所定义，并定义 $\boldsymbol{\mu}^{(4)} = \frac{1}{L}\sum_{i=1}^L \mathbf{x}^{(i)} \in \mathbb{R}^d$、$\boldsymbol{\mu}^{(5)} = \frac{1}{L}\sum_{j=1}^L \mathbf{y}^{(j)} \in \mathbb{R}^d$。此外，设$\mathbf{Q}^{(3)}\mathbf{\Lambda}^{(3)}(\mathbf{Q}^{(3)})^\top$ 是对称半正定矩阵$\mathbf{M}^{(1)} + \boldsymbol{\mu}^{(4)}(\boldsymbol{\mu}^{(5)})^\top + \boldsymbol{\mu}^{(5)}(\boldsymbol{\mu}^{(4)})^\top + \mathbf{M}^{(2)}$ 的特征值分解，其中$\mathbf{Q}^{(3)} \in \mathbb{O}_d$ 且 $\mathbf{\Lambda}^{(3)} \in \mathbb{D}_d$ 的对角元素均非负。此外，我们假设$\mathbf{\Lambda}^{(3)}$ 的对角元素按非升序排列。

min $\boldsymbol{\theta}^*_{\mathrm{SDE}} = \{\mathbf{A}, \mathbf{B}, \mathbf{C}\ D\}$ 的$\min_{\boldsymbol{\theta}_{\mathrm{SDE}} \in \Theta_{\mathrm{SDE}}} \mathcal{L}(\boldsymbol{\theta}_{\mathrm{SDE}}; \mathcal{X}, \mathcal{Y}, \mathcal{T}_{\mathrm{SDE}})$ 解之一如下。$\mathbf{A} \in \mathbb{D}_d$，对所有 $1 \le l \le d$:

$$
\mathbf{A}_{l,l} = \frac{1}{16}\left(1 - 2\mathbf{\Lambda}^{(3)}_{l,l} - \sqrt{\left(2\mathbf{\Lambda}^{(3)}_{l,l} + 1\right)^2 + 8\mathbf{\Lambda}^{(3)}_{l,l}}\right),
$$

$\mathbf{B} = (\mathbf{I}_d - 4\mathbf{A})^{1/2}(\mathbf{Q}^{(3)})^\top$，$\mathbf{C} = -\frac{1}{2}\mathbf{I}_d$, $D = \det(\mathbf{I}_d - 4\mathbf{A})^{1/4}$。此外，我们有：

$$
\begin{aligned}
\overline{\mathcal{L}}(\boldsymbol{\theta}_{\mathrm{SDE}}; \mathcal{X}, \mathcal{Y}, \mathcal{T}_{\mathrm{SDE}}) = \sum_{l=1}^d \Bigg(&\log(1 - 4\mathbf{A}_{l,l})\\
&-\frac{1}{2}\log(1 - 8\mathbf{A}_{l,l}) + \left(1 + (1 - 8\mathbf{A}_{l,l})^{-1}\right)\mathbf{\Lambda}^{(3)}_{l,l}\Bigg)\\
&-L^{-1}\sum_{i=1}^L \|\mathbf{x}^{(i)}\|^2 - L^{-1}\sum_{j=1}^L \|\mathbf{y}^{(j)}\|^2 \qquad (17)
\end{aligned}
$$

再次，定理 4.3 表明，可以用低于 L 次方的时间找到 $\boldsymbol{\theta}^*_{\mathrm{SDE}}$。也就是说，我们可以在总时间 $O(Ld^2)$ 内计算 $\mathbf{M}^{(1)}$,$\mathbf{M}^{(2)}$, $\mu^{(3)}$, $\boldsymbol{\mu}^{(4)}$, $\boldsymbol{\mu}^{(5)}$。然后进行特征值分解，以在 $O(d^3)$ 时间内得到 $\mathbf{Q}^{(3)}, \mathbf{\Lambda}^{(3)}$。之后，根据定理 4.3 中的公式，可以在 $O(d^3)$ 时间内计算 $\mathbf{A}, \mathbf{B}$,$\mathbf{C}$ D。该近似方案的总时间复杂度

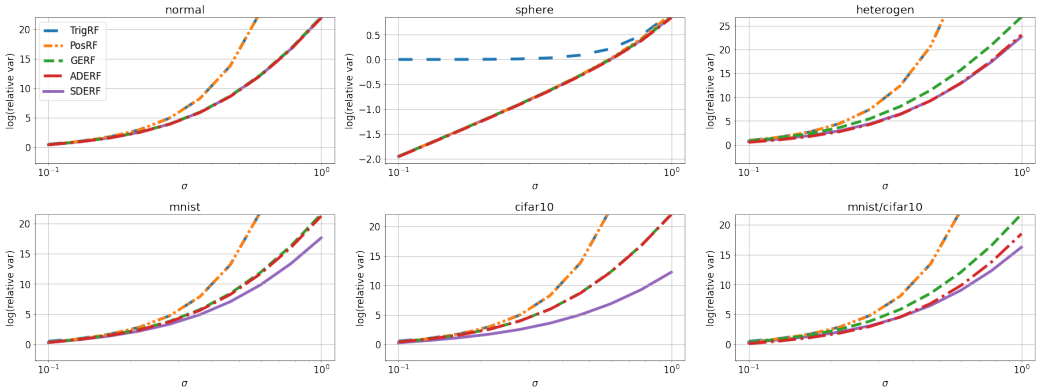

图 2.新旧随机特征机制相对方差的对数，多次采样的均值。 $0.1 \leq \sigma \leq 1$。

.

与 ADERF 相同：$O(L(Md + d^2) + Md^2 + d^3)$；如果假设 $L \geq d$，则为 $O(L(Md + d^2) + Md^2)$。

4.3. 简化 ADERF

尽管 ADERF 和 SDERF 都具有紧致的闭式表达式，但它们都依赖特征值分解和奇异值分解：在支持 GPU 和 TPU 的主流深度学习库中，这些操作的实现尚未成熟。因此，我们提出简化 *ADERFs*（*SADERFs*） $\mathcal{T}_{\mathrm{SADE}} = \langle \nu_{\mathrm{SADE}}, f^{(1)}{}_{\mathrm{SADE}}\ f_{(2)\,\mathrm{SADE}} \rangle$，它们扩展了 GERF，但只需要基本的一元运算。SADERF 通过 GERF 定义如下：$\Omega_{\mathrm{SADE}} = \mathbb{R}^d$，$\mathcal{N}\ 0,1\ ^d\ f^{\ 1}\quad \boldsymbol{\omega} = f^{(1)}\ \boldsymbol{\omega}\ \boldsymbol{\Psi}\quad f^{(2)}\ \boldsymbol{\omega} = (),$ $()_{\mathrm{SADE}}(,\mathbf{x})\ _{\mathrm{GE}}(,\mathbf{x})$，$_{\mathrm{SADE}}(,\mathbf{y})\ f^{(2)}{}_{\mathrm{GE}}(\boldsymbol{\omega}, \boldsymbol{\Psi}^{-1}\mathbf{y})$，其中 $\boldsymbol{\Psi} \in \mathbb{D}_d$ 是一个对角元素非零的对角矩阵。首先，$\mathcal{T}_{\mathrm{SADE}}$ 是 softmax 核的有效随机特征 $K^{(0)}$，因为：

$$\mathbb{E}_{\nu_{\mathrm{SADE}}}[f^{(1)}_{\mathrm{SADE}}(\boldsymbol{\omega}, \mathbf{x}) f^{(2)}_{\mathrm{SADE}}(\boldsymbol{\omega}, \mathbf{y})] = \mathbb{E}_{\nu_{\mathrm{GE}}}[f^{(1)}_{\mathrm{GE}}(\boldsymbol{\omega}, \boldsymbol{\Psi}\mathbf{x}) \times f^{(2)}_{\mathrm{GE}}(\boldsymbol{\omega}, \boldsymbol{\Psi}^{-1}\mathbf{y})] = K^{(0)}(\boldsymbol{\Psi}\mathbf{x}, \boldsymbol{\Psi}^{-1}\mathbf{y}) = K^{(0)}(\mathbf{x}, \mathbf{y}),$$

根据定义，我们使用 $K^{(0)}(\mathbf{x},\ \mathbf{y}) = \exp(\mathbf{x}^\top \mathbf{y})$。

我们通过对 $\mathcal{T}_{\mathrm{SADE}}$ 优化目标函数 (14) 得到 $\boldsymbol{\Psi}$，其形式可由 $\mathcal{L}(\boldsymbol{\theta}_{\mathrm{GE}};\ \mathcal{X},\ \mathcal{Y},\ \mathcal{T}_{\mathrm{GE}})$ 直接推导得出：

$$\overline{\mathcal{L}}(\boldsymbol{\theta}_{\mathrm{SADE}}; \mathcal{X}, \mathcal{Y}, \mathcal{T}_{\mathrm{SADE}}) - d \log\left(\frac{1-4A}{\sqrt{1-8A}}\right) = \frac{2-8A}{1-8A} \times \frac{1}{L^2}\sum_{i,j} \|\boldsymbol{\Psi}\mathbf{x}^{(i)} + \boldsymbol{\Psi}^{-1}\mathbf{y}^{(j)}\|^2 - \frac{1}{L}\sum_i (\|\boldsymbol{\Psi}\mathbf{x}^{(i)}\|^2 + \|\boldsymbol{\Psi}^{-1}\mathbf{y}^{(j)}\|^2) = \frac{1}{L^2(1-8A)}\sum_{i,j}\|\boldsymbol{\Psi}\mathbf{x}^{(i)} + \boldsymbol{\Psi}^{-1}\mathbf{y}^{(j)}\|^2 + \frac{2}{L^2}\sum_{i,j}(\mathbf{x}^{(i)})^\top \mathbf{y}^{(j)}, \tag{18}$$

其中，我们将一项与 $\boldsymbol{\Psi}$ 无关的项移至等式左侧。由于 $1 - 8A > 0$，我们得出，最小化 (18) 等价于最小化

$$\sum_{i,j}\|\boldsymbol{\Psi}\mathbf{x}^{(i)} + \boldsymbol{\Psi}^{-1}\mathbf{y}^{(j)}\|^2 = \sum_l \sum_{i,j} (\boldsymbol{\Psi}_{l,l}\mathbf{x}_l^{(i)} + \boldsymbol{\Psi}_{l,l}^{-1}\mathbf{y}_l^{(j)})^2 = \sum_l \sum_{i,j} (\boldsymbol{\Psi}_{l,l}^2 (\mathbf{x}_l^{(i)})^2 + 2\mathbf{x}_l^{(i)}\mathbf{y}_l^{(j)} + \boldsymbol{\Psi}_{l,l}^{-2}(\mathbf{y}_l^{(j)})^2). \tag{19}$$

优化（19）可归结为分别针对 $\boldsymbol{\Psi}_{l,\ l}$, $1 \leq l \leq d$ 的独立优化问题。每个问题都是凸问题，只需令导数为零即可轻松求得解：

$$\forall 1 \leq l \leq d: \boldsymbol{\Psi}^*_{l,l} = (\sum_j (\mathbf{y}_l^{(j)})^2 / \sum_i (\mathbf{x}_l^{(i)})^2)^{1/4}. \tag{20}$$

（20）可在 $O(dL)$ 时间内计算，随后可按照第 2.3 节所述，高效地找到 $f_{(1)\,\mathrm{GE}}\ f^{(2)}{}_{\mathrm{GE}}$ 的参数。

不难看出，$\mathcal{T}_{\mathrm{SADE}}$ 是 ADERF（第 4.1 节）的特例，这也解释了它们名称的由来。此外，情形 $\boldsymbol{\Psi} = \mathbf{I}_d$ 可化简为 $\mathcal{T}_{\mathrm{GE}}$，因此后者是前者的特例。图 1（上）以维恩图的形式展示了所有新型随机特征。

5. 实验

本节通过实验评估 DERF 在各种机器学习应用中的表现。附录 B 中提供了各项实验的更多详情。

5.1. 方差比较

我们遵循 (Likhosherstov 等人，2022) 中的方差比较设置：对向量$\mathbf{x}$、$\mathbf{y}$进行采样，并计算近似值 Var $\widehat{K}^{(0)}(\mathbf{x},\mathbf{y})/K^{(0)}(\mathbf{x},\mathbf{y})$ 的相对方差，其中 $\widehat{K}^{(0)}$ 表示 RF 近似，而 Var $\widehat{K}^{(0)}(\mathbf{x},\mathbf{y})$ 根据式 (15) 计算。我们

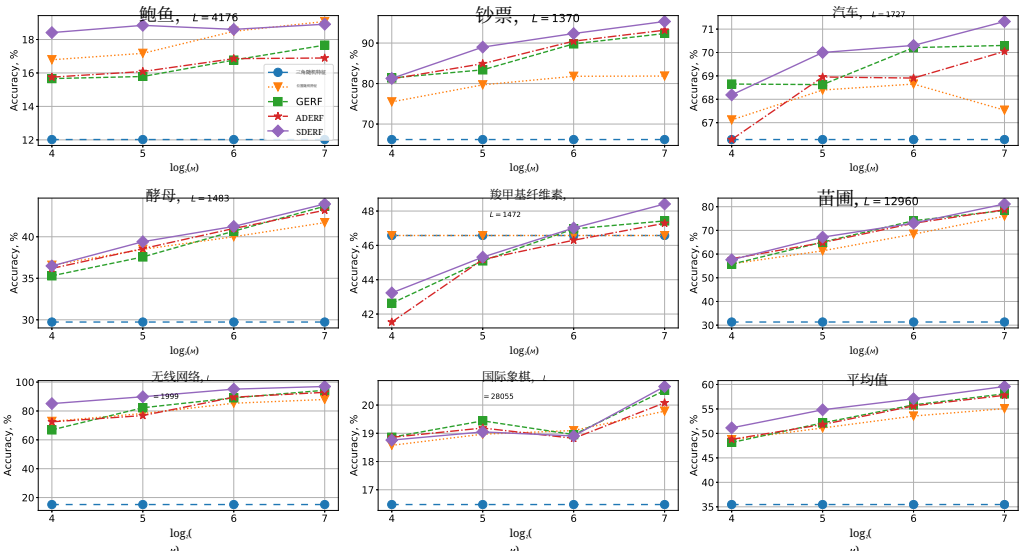

图 *3.* 核分类，测试准确率（%）。最后一张图展示了 8 个基准测试上的平均曲线。我们观察到，所提出的 SDERF 方法在大多数（基准测试， M）组合中准确率最高，且平均表现也最佳。

按 (Likhosherstov 等人，2022) 中的方式设置 $d = 64$，并采用 6 种不同的**x**、**y**采样方案：normal 表示从 $\mathcal{N}(\mathbf{0}_d, \sigma^2\mathbf{I}_d)$ 中抽取 **x** 、 **y**；sphere 表示在球面 $\sigma\mathcal{S}^{d-1}$ 上均匀抽取 **x** 、 **y**；heterogen 表示从不同分布 $\mathcal{N}(\mathbf{0}_d, \sigma^2\mathbf{I}_d)$ 和 $\mathcal{N}(\sigma\mathbf{1}_d, \sigma^2\mathbf{I}_d)$ 中抽取 **x** 、 **y**。mnist 和 cifar10 表示从 MNIST (Deng, 2012) 或 CIFAR10 (Krizhevsky 等人) 中随机抽取图像，将其调整为 8×8，按 $\sigma > 0$ 缩放后展平。最后，mnist/cifar10 表示一种采样方案，其中**x**按照 mnist 方案抽取，而**y** 按照 cifar10 方案抽取。

我们未报告 SADERF，因为它们是 ADERF 的一种特殊情况（图 2）。在所有设置中，SDERF 的表现均优于或不逊于其他方法：在 heterogen、mnist 和 mnist/cifar10 中，性能约为 GERF 的 e^5 倍；在 cifar10 中则约为 e^{10} 倍。此外，在 mnist/cifar10 中，当**x** 和**y** 以“非对称”方式抽取时，ADERF 的表现约为 GERF 的 e^3 倍。

5.2. 核分类

在这项实验中，我们按照 Likhosherstov 等人（2022）的设置，在来自加州大学欧文分校（Dua 和 Graff，2017）的 8 个基准测试上比较不同随机特征方法在核分类中的准确率。我们采用核回归（Nadaraya，1964；Watson，1964）来预测类别概率。训练对象记为 $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(L)} \in \mathbb{R}^d$，其独热标签记为 $\mathbf{r}^{(1)}, \ldots, \mathbf{r}^{(L)} \in \mathbb{R}^n$。测试时，目标是将新对象 $\mathbf{u}^*$ 预测为类别 $\mathrm{argmax}_{1 \leq l \leq n} \mathbf{r}^*$，其中 $\mathbf{r}^* = \sum_{i=1}^{L} K^{(-0.5)}(\sigma\mathbf{u}^*, \sigma\mathbf{u}^{(i)})\mathbf{r}^{(i)}$，且 $\sigma > 0$ 在验证集上调优。采用 $O(nLM)$ 预处理后，我们使用随机特征来寻找 $\mathbf{r}^*$ 在 $O(nM)$ 中的无偏近似，而非精确计算 $O(nL)$。

对于每个基准测试，我们将 M 的取值范围设为从 2^4 到 2^7 （图 3 ）。我们观察到，本文提出的 SDERF 在大多数（基准测试， M）组合中准确率最高，平均表现也最佳。

5.3. 面向长序列 Transformer 的 DERF

本节中，我们在多种 Performer-Transformer 训练设置下评估 DERF 的自注意力近似效果（Choromanski等，2021年）。我们将基于 DERF 的自注意力近似方法称为*FAVOR#*。

5.3.1. 语音建模

在第一组实验中，我们专注于语音模型。我们训练 Performer编码器，并在*LibriSpeech* 语料库（Panayotov 等，2015年）上测试，这一语料库常用于语音模型的基准测试。我们考虑了两种 Transformer 架构/训练设置：**(a)**以常规方式训练的 Conformer-Transducer（Gulati等，2020 年）（TRANS），以及：**(b)** Park等（2020年）提出的噪声学生训练（NST）变体。

我们比较了这些架构的“Performer化”变体，应用带有 SDERF 的 FAVOR#（因为它在先前的设置中效果最佳），以及 FAVOR++ （Likhosherstov 等人，2022）。

在第一种设置中，我们发现，当 m 较小时，FAVOR# 始终优于 FAVOR++（此时方差降低

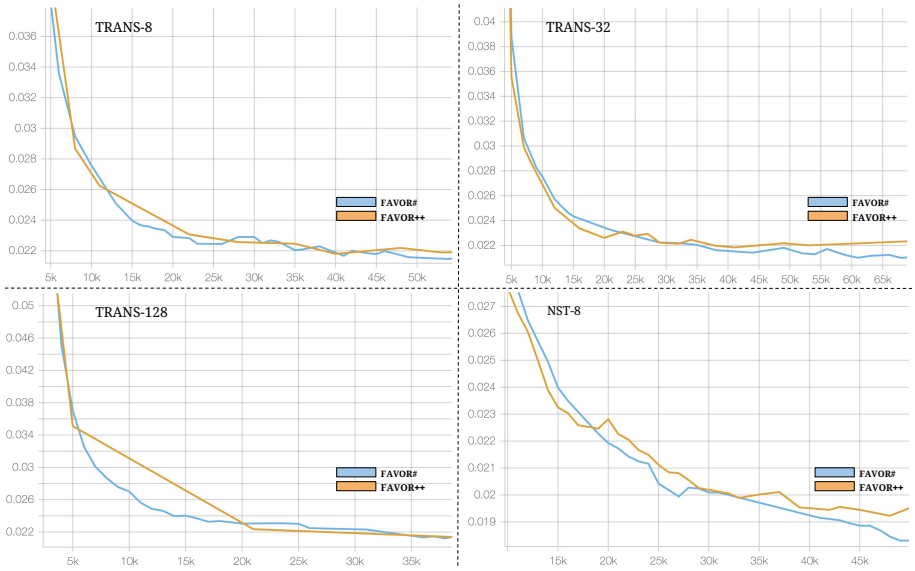

图*4.* 在 *LibriSpeech* 语料库上，比较使用 SDRF 的 FAVOR# 与采用 m 个随机特征（TRANS-m）进行常规 Conformer-Transducer 训练的 FAVOR++ Performer，以及采用 m 个随机特征（NST-m）的噪声学生训练变体。我们报告常用的归一化词错误率（NWER）指标。

表1. 基座规模模型在 GLUE 开发集上的结果。每项任务下方列出了训练样本数。CoLA 报告 MCC 分数，MRPC 报告 F1 分数，STS-B 报告斯皮尔曼相关系数，其余任务报告准确率分数。**最佳**结果、次佳结果。

系统	MNLI(m) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k
ELU（Katharopoulos 等人，2020）	82.58	90.05	89.81	92.43	58.63	87.91	87.50	67.15
RELU（Choromanski等，2021年）	82.49	**90.71**	89.68	92.32	57.57	**88.15**	87.25	68.95
FAVOR+（Choromanski等，2021年）	77.69	86.69	89.41	91.80	54.87	83.78	80.73	66.19
FAVOR++ （Likhosherstov 等人，2022）	82.29	90.43	89.73	92.20	58.85	85.90	**88.73**	67.63
FAVOR#	**82.69**	90.68	**90.01**	**92.53**	**59.33**	85.48	87.99	**69.68**

对 softmax 核估计更为关键）；而当 m 较大时，两者的得分相近。在 NST 实验中，我们重点关注较小的 m 变体，此时 FAVOR# 再次优于 FAVOR++。所有结果均展示于图 4。

5.3.2. 自然语言处理

通用语言理解评估（GLUE）基准测试（Wang 等人，2018）包含 8 项不同的自然语言理解任务，序列长度范围为 32 至 128。我们使用该基准测试来评估不同低秩注意力方法在自然语言处理任务上的性能。我们采用了（Devlin 等人，2018）中提到的相同训练参数（详情请参见附录 B.3.2）。我们使用预训练的 BERT-base 模型检查点（Devlin 等人，2018）对所有低秩 Transformer 进行热启动，从而比较低秩方法对 softmax 核的近似效果。

我们将 FAVOR++ (Likhosherstov 等人，2022年)、FA- VOR + (Choromanski 等，2021年)、ELU (Katharopoulos 等人，2020) 和 ReLU (Choromanski 等，2021年) 这几种 Performers (Choromanski 等，2021年) 变体与 FAVOR# 变体进行了比较，并在表1中报告结果。在此设置中，我们无法使用 SDERF，因为与语音建模实验相比，实现方式有所不同，特征分解在 TPU 上导致了错误。因此，我们使用了无需进行任何矩阵分解的 SADERF。在大多数任务中，我们发现 FAVOR# 是表现最佳的变体，彰显了其在为 Transformer 建模 softmax 核方面的有效性。

6. 结论

我们提出了广义指数随机特征（GERF）在高斯核和 softmax 核上的扩展：稠密指数随机特征（DERF）。DERF 使用矩阵参数，比 GERF 更灵活。我们在核回归和两种 Transformer 训练设置中评估了 DERF，结果表明其具有显著优势。

7. 致谢

V. L. 感谢剑桥信托和 DeepMind 提供支持。V. L. 在攻读博士学位期间曾兼职受雇于 Google。A. W. 感谢图灵人工智能奖学金（英国工程与自然科学研究委员会资助项目 EP/V025279/1）、艾伦·图灵研究所，以及莱弗休姆信托通过 CFI 提供的支持。

参考文献

Ailon, N. 和 Liberty, E. 一种近乎最优的无约束快速 Johnson-Lindenstrauss变换。《*ACM*算法汇刊》，9(3):21:1–21:12，2013年。doi: 10. 1145/2483699.2483701。URL https://doi.org/ 10.1145/2483699.2483701。

Avron, H.、Sindhwani, V.、Yang, J. 和 Mahoney, M.W. 用于平移不变核的准蒙特卡罗特征映射。《机器学习研究杂志》，17:120:1–120:38，2016年。URL http://jmlr.org/papers/v17/14-538. html。

Avron, H.、Kapralov, M.、Musco, C.、Musco, C.、Velingker, A. 和 Zandieh, A. 用于核岭回归的随机傅里叶特征：近似界与统计保证。载于 Precup, D. 和 Teh, Y. W.（编），第*34*届国际机器学习会议论文集，*ICML 2017*，悉尼，新南威尔士州，澳大利亚，*2017*年*8*月*6*日至*11*日，第70卷《机器学习研究论文集》，第253–262页。PMLR，2017年。URL http://proceedings.mlr.press/ v70/avron17a.html。

Boffi, N. M., Tu, S. 和 Slotine, J. E. 非参数自适应控制与预测：理论与随机化算法。载于第*60*届 *IEEE* 决策与控制会议（*CDC 2021*），奥斯汀，得克萨斯州，美国，*2021*年*12*月*14*日至*17*日，第2935–2942页。IEEE，2021年。doi: 10.1109/CDC45484.2021.9682907。网址: https://doi.org/10.1109/CDC45484.2021.9682907。

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S. 和 Zhang, Q. JAX: Python+NumPy 程序的可组合变换，2018年。网址: http://github.com/google/jax。

Brockett, R. 用于对列表排序、对矩阵进行对角化以及求解线性规划问题的动力系统。线性代数及其应用，146:79–91，1991年。ISSN 0024-3795。doi: https://doi.org/10.1016/0024-3795(91)90021-N。URL https://www.sciencedirect.com/science/article/pii/002437959190021N。

Chaudhuri, K.、Monteleoni, C. 和 Sarwate, A. D. 差分隐私经验风险最小化。机器学习研究期刊学习研究，12:1069–1109，2011 年。URL http://dl. acm.org/citation.cfm?id=2021036。

Cho, Y. 和 Saul, L. K. 深度学习的核方法。载于 Bengio, Y.、Schuurmans, D.、Lafferty, J. D.、Williams, C. K. I. 和 Culotta, A.（编），神经信息处理系统进展 *22*：第*23*届神经信息处理系统年会（*2009*）。会议论文集，会议于*2009*年*12*月*7*日至*10*日在加拿大不列颠哥伦比亚省温哥华举行，第342–350页。Curran Associates, Inc.，2009 年。

Choromanski, K.、Downey, C. 和 Boots, B. 初始化至关重要：正交预测状态循环神经网络。在第六届国际学习表征会议，*ICLR 2018*，温哥华（加拿大不列颠哥伦比亚省），*2018*年*4*月*30*日至*5*月*3*日，会议论文集。OpenReview.net，2018年。URL https://openreview. net/forum?id=HJJ23bW0b。

Choromanski, K.、Likhosherstov, V.、Dohan, D.、Song, X.、Davis, J.、Sarl´os, T.、Belanger, D.、Colwell, L. J. 和 Weller, A. 通过线性可扩展长上下文变换器进行蛋白质掩码语言建模。*CoRR*，abs/2006.03555，2020年。URL https://arxiv.org/ abs/2006.03555。

Choromanski, K., Chen, H., Lin, H., Ma, Y., Sehanobish, A., Jain, D., Ryoo, M. S., Varley, J., Zeng, A., Likhosherstov, V., Kalashnikov, D., Sindhwani, V. 和 Weller, A. 混合随机特征。在国际学习表征会议（*ICLR*），2022年。

Choromanski, K. M., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J. Q., Mo- hiuddin, A., Kaiser, L., Belanger, D. B., Colwell, L. J. 和Weller,A. 重新思考使用Performers的注意力机制。在国际学习表征会议，2021年。网址 https://openreview.net/forum? id=Ua6zuk0WRH。

Chowdhury, S. P.、Solomou, A.、Dubey, A. 和 Sachan, M. 关于 Transformer 内核的学习。《机器学习研究汇刊》，2022年。URL https: //arxiv.org/abs/2110.08323。

Dasgupta, S. 和 Gupta, A. Johnson–Lindenstrauss 定理的初等证明。《随机结构与算法》，22(1):60–65，2003年。doi: 10.1002/rsa.10073。URL https://doi.org/10.1002/rsa.10073。

Deng, L. 面向机器学习研究的手写数字图像 MNIST 数据库。《*IEEE* 信号处理杂志》，29(6):141–142，2012年。

Devlin, J., Chang, M.-W., Lee, K. 和 Toutanova, K. BERT：用于语言理解的深度双向变换器预训练。*arXiv* 预印本 *arXiv:1810.04805*，2018年。

Dua, D. 和 Graff, C. UCI机器学习资料库，2017年。URLhttp://archive.ics.uci.edu/ml。

Gonon, L. 随机特征神经网络无需遭受维数灾难即可学习 Black-Scholes 型偏微分方程。*CoRR*，abs/2106.08900，2021年。URL https://arxiv. org/abs/2106.08900。

Gulati, A., Qin, J., Chiu, C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y. 和 Pang, R. Conformer：用于语音识别的卷积增强型变换器。*CoRR*，abs/2005.08100，2020年。URL https://arxiv.org/abs/2005.08100。

Han, I., Avron, H., Shoham, N., Kim, C. 和 Shin, J. 用于神经切线核的随机特征。*CoRR*，abs/2104.01351，2021年。URL https://arxiv.org/ abs/2104.01351。

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del R´ıo, J. F., Wiebe, M., Peterson, P., G´erard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. 使用 NumPy 进行数组编程。《自然》，585(7825):357–362, 2020年9月。doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2. Katharopoulos, A., Vyas, A., Pappas, N., and Fleuret, F. Transformer 是循环神经网络：具有线性注意力的快速自回归变换器。发表于第*37*届国际机器学习会议论文集，2020年。Krizhevsky, A., Nair, V., and Hinton, G. CIFAR-10（加拿大高级研究所）。URL http://www. cs.toronto.edu/~kriz/cifar.html. Laparra, V., Gonzalez, D. M., Tuia, D., and Camps-Valls, G. 用于核回归的大规模随机特征。发表于*2015*年*IEEE*国际地球科学与遥感研讨会（*IGARSS*），第17–20页，2015年。doi: 10.1109/IGARSS.2015.7325686. Le, Q. V., Sarl´os, T., and Smola, A. J. Fastfood——以对数线性时间计算希尔伯特空间展开。发表于第*30*届国际机器学习大会（*ICML 2013*），美国佐治亚州亚特兰大，*2013*年*6*月*16*日至*21*日，第28卷*JMLR*研讨会与会议论文集，第244–252页。JMLR.org，2013年。URL http: // proceedings.mlr.press/v28/le13.html。

Li, Z., Ton, J., Oglic, D., and Sejdinovic, D. 关于随机傅里叶特征统一分析的探索。《机器学习研究杂志》，22:108:1–108:51，2021年。URL http://jmlr.org/papers/v22/20-1369.html。

Likhosherstov, V., Choromanski, K., Dubey, A., Liu, F., Sarlos, T., 和 Weller, A. 厨师随机表：非三角随机特征。载于神经信息处理系统进展。Curran Associates, Inc., 2022年。

Minh, H. Q. 算子值 Bochner 定理、算子值核的傅里叶特征映射与向量值学习。*CoRR*，abs/1608.05639，2016年。URL http://arxiv.org/abs/1608.05639。

Nadaraya, E. A. 关于回归估计。概率论及其应用，9(1):141–142，1964年。doi: 10.1137/1109020。URLhttps://doi.org/10. 1137/1109020。

Oliva, J. B., Neiswanger, W., P´oczos, B., Xing, E. P., Trac, H., Ho, S., 和 Schneider, J. G. 快速的函数到函数回归。载于 Lebanon, G. 和 Vishwanathan, S. V. N.（编），第十八届人工智能与统计国际会议论文集，*AISTATS 2015*，美国加利福尼亚州圣迭戈，*2015*年*5*月*9*日至*12*日，第38卷，*JMLR* 研讨会与会议论文集。JMLR.org，2015年。网址：http://proceedings.mlr.press/v38/oliva15.html。Panayotov, V., Chen, G., Povey, D., 和 Khudanpur, S. LibriSpeech：基于公有领域有声书的自动语音识别语料库。载于*2015*年 *IEEE* 声学、语音与信号处理国际会议，*ICASSP 2015*，澳大利亚昆士兰州南布里斯班，*2015*年*4*月*19*日至*24*日，第5206–5210页。IEEE，2015年。doi：10.1109/ICASSP.2015.7178964。网址：https://doi.org/10.1109/ICASSP.2015.7178964。Park, D. S., Zhang, Y., Jia, Y., Han, W., Chiu, C., Li, B., Wu, Y., 和 Le, Q. V. 改进的自动语音识别噪声学生训练。载于 Meng, H., Xu, B., 和 Zheng, T. F.（编），*Interspeech 2020*，国际语音通信协会第*21*届年会，线上活动，中国上海，*2020*年*10*月*25*日至*29*日，第2817–2821页。国际语音通信协会，2020年。doi: 10.21437/Interspeech.2020-1470。网址：https://doi.org/10.21437/Interspeech.2020-1470。Pennington, J., Yu, F. X., 和 Kumar, S. 用于多项式核的球面随机特征。载于 Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., 和 Garnett, R.（编），神经信息处理系统进展第*28*卷：*2015*年神经信息处理系统年会，*12*月

*2015*年*12*月*7*日至*12*日，加拿大魁北克省蒙特利尔，第1846–1854页，2015年。网址：https://proceedings.neurips.cc/paper/2015/hash/f7f580e11d00a75814d2ded41fe8e8fe-Abstract.html。

Rahimi, A. 和 Recht, B. 面向大规模核机器的随机特征。载于 Platt, J.、Koller, D.、Singer, Y. 和 Roweis, S.（编），神经信息处理系统进展，第20卷。Curran Associates, Inc.，2007年。

Rahimi, A. 和 Recht, B. 利用随机基对函数进行一致逼近。载于*2008*年第*46*届*Allerton*通信、控制与计算会议，美国加利福尼亚州洛斯阿拉米托斯，2008年9月a。IEEE计算机学会。doi：10.1109/ALLERTON.2008.4797607。URL https://doi.ieeecomputersociety.org/10.1109/ALLERTON.2008.4797607。Rahimi, A. 和 Recht, B. 随机厨房水槽的加权和：在学习中以随机化取代最小化。载于 Koller, D.、Schuurmans, D.、Bengio, Y. 和 Bottou, L.（编），神经信息处理系统进展第*21*卷：第二十二届神经信息处理系统年会论文集，*2008*年*12*月*8*日至*11*日，加拿大不列颠哥伦比亚省温哥华，第1313–1320页。Curran Associates, Inc.，2008年b。Sriperumbudur, B. K. 和 Szab´o, Z. 随机傅里叶特征的最优速率。载于 Cortes, C.、Lawrence, N. D.、Lee, D. D.、Sugiyama, M. 和 Garnett, R.（编），神经信息处理系统进展第*28*卷：*2015*年神经信息处理系统年会，*2015*年*12*月*7*日至*12*日，加拿大魁北克省蒙特利尔，第1144–1152页，2015年。Sun, Y.、Gilbert, A. C. 和 Tewari, A. 但理论上它究竟如何运作？带随机特征的线性支持向量机。载于 Bengio, S.、Wallach, H. M.、Larochelle, H.、Grauman, K.、Cesa-Bianchi, N. 和 Garnett, R.（编），神经信息处理系统进展第*31*卷：*2018*年神经信息处理系统年会，神经信息处理系统大会*2018*，*2018*年*12*月*3*日至*8*日，蒙特´利尔，加拿大，第3383–3392页，2018年。Sutherland, D. J. 和 Schneider, J. G. 随机傅里叶特征误差。载于 Meila, M. 和 Heskes, T.（编），第三十一届人工智能不确定性会议论文集，*UAI 2015*，*2015*年*7*月*12*日至*16*日，荷兰阿姆斯特丹，第862–871页。AUAI出版社，2015年。URL http://auai.org/uai2015/proceedings/papers/168.pdf。Trefethen, L. N. 和 Bau, D. 数值线性代数。工业与应用数学学会，1997年。ISBN 0898713617。

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. GLUE：用于自然语言理解的多任务基准测试与分析平台。*arXiv*预印本 *arXiv:1804.07461*，2018年。Watson, G. S. 平滑回归分析。*Sankhy¯a*：印度统计学杂志，*A*辑（*1961-2002*），26(4)：359–372，1964年。ISSN 0581572X。网址：http://www.jstor.org/stable/25049340。Xie, J., Liu, F., Wang, K., and Huang, X. 基于随机傅里叶特征的深度核学习。*CoRR*，abs/1910.02660，2019年。网址：http://arxiv.org/ abs/1910.02660。Yang, Duraiswami, Gumerov, and Davis. 改进的快速高斯变换与高效核密度估计。载于第九届*IEEE*国际计算机视觉会议论文集，第664–671页，第1卷，2003年。doi：10.1109/ ICCV.2003.1238383。Yang, J., Sindhwani, V., Fan, Q., Avron, H., and Mahoney, M. W. 用于直方图半群核的随机拉普拉斯特征映射。载于*2014*年*IEEE*计算机视觉与模式识别会议，*CVPR 2014*，美国俄亥俄州哥伦布，*2014*年*6*月*23*日至*28*日，第971–978页。IEEE计算机学会，2014年。doi：10.1109/CVPR.2014. 129。网址：https://doi.org/10.1109/CVPR. 2014.129。Yang, T., Li, Y., Mahdavi, M., Jin, R., and Zhou, Z. Nystr¨om方法与随机傅里叶特征：理论与实证比较。载于 Bartlett, P. L., Pereira, F. C. N., Burges, C. J. C., Bottou, L., and Weinberger, K. Q.（编），神经信息处理系统进展*25*：*2012*年第*26*届神经信息处理系统年会。会议论文集，会议于*2012*年*12*月*3*日至*6*日在美国内华达州太浩湖举行，第485–493页，2012年。Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urta-sun, R., Torralba, A., and Fidler, S. 对齐书籍与电影：通过观看电影和阅读书籍实现故事化视觉解释。载于*IEEE*国际计算机视觉会议，第19–27页，2015年。

A. 证明

A.1. 定理 4.1 的证明

证明。根据 $\langle \nu_{\mathrm{DE}}, f^{(1)}{}_{\mathrm{DE}}\ f^{(2)}{}_{\mathrm{DE}}\rangle$的定义，我们有：

$$\mathbb{E}_{\nu_{\mathrm{DE}}} f^{(1)}_{\mathrm{DE}}(\boldsymbol{\omega},\mathbf{x}) f^{(2)}_{\mathrm{DE}}(\boldsymbol{\omega},\mathbf{y}) = (2\pi)^{-\frac{d}{2}} D^2 \int_{\mathbb{R}^d} \exp\Big(-\frac{1}{2}\|\boldsymbol{\omega}\|^2 + 2\boldsymbol{\omega}^\top \mathbf{A}\boldsymbol{\omega} + \boldsymbol{\omega}^\top(\mathbf{B}^{(1)}\mathbf{x} + \mathbf{B}^{(2)}\mathbf{y}) + \mathbf{x}^\top \mathbf{C}^{(1)}\mathbf{x}$$
$$+ \mathbf{y}^\top \mathbf{C}^{(2)}\mathbf{y}\Big) d\boldsymbol{\omega} = (2\pi)^{-\frac{d}{2}} D^2 \exp\left(\mathbf{x}^\top \mathbf{C}^{(1)}\mathbf{x} + \mathbf{y}^\top \mathbf{C}^{(2)}\mathbf{y}\right) \int_{\mathbb{R}^d} \exp\Big(-\frac{1}{2}\boldsymbol{\omega}^\top(\mathbf{I}_d - 4\mathbf{A})\boldsymbol{\omega} + \boldsymbol{\omega}^\top(\mathbf{B}^{(1)}\mathbf{x} + \mathbf{B}^{(2)}\mathbf{y})\Big) d\boldsymbol{\omega}.$$

由于 $8\mathbf{A} \prec \mathbf{I}_d$，我们有 $4\mathbf{A} \prec 0.5\mathbf{I}_d \prec \mathbf{I}_d$，这意味着 $\mathbf{I}_d - 4\mathbf{A}$ 是正定且可逆的。以下恒等式很容易验证：

$$-\frac{1}{2}\boldsymbol{\omega}^\top(\mathbf{I}_d - 4\mathbf{A})\boldsymbol{\omega} + \boldsymbol{\omega}^\top(\mathbf{B}^{(1)}\mathbf{x} + \mathbf{B}^{(2)}\mathbf{y}) = -\frac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\omega} - \boldsymbol{\mu}) + \frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu},$$
$$\boldsymbol{\Sigma} = (\mathbf{I}_d - 4\mathbf{A})^{-1}, \quad \boldsymbol{\mu} = \boldsymbol{\Sigma}(\mathbf{B}^{(1)}\mathbf{x} + \mathbf{B}^{(2)}\mathbf{y}).$$

因此，我们有：

$$\mathbb{E}_{\nu_{\mathrm{DE}}} f^{(1)}_{\mathrm{DE}}(\boldsymbol{\omega},\mathbf{x}) f^{(2)}_{\mathrm{DE}}(\boldsymbol{\omega},\mathbf{y}) = (2\pi)^{-d/2} D^2 \exp\left(\mathbf{x}^\top \mathbf{C}^{(1)}\mathbf{x} + \mathbf{y}^\top \mathbf{C}^{(2)}\mathbf{y} + \frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)$$
$$\times \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\omega} - \boldsymbol{\mu})\right) d\boldsymbol{\omega}.$$

接下来，我们利用这样一个事实：均值为 $\boldsymbol{\mu}$、方差为 $\boldsymbol{\Sigma}$ 的多元高斯分布的积分为 1：

$$(2\pi)^{-d/2} \det(\boldsymbol{\Sigma})^{-1/2} \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\omega} - \boldsymbol{\mu})\right) d\boldsymbol{\omega} = 1.$$

由此我们得出：

$$\mathbb{E}_{\nu_{\mathrm{DE}}} f^{(1)}_{\mathrm{DE}}(\boldsymbol{\omega},\mathbf{x}) f^{(2)}_{\mathrm{DE}}(\boldsymbol{\omega},\mathbf{y}) = D^2 \det(\boldsymbol{\Sigma})^{1/2} \exp\left(\mathbf{x}^\top \mathbf{C}^{(1)}\mathbf{x} + \mathbf{y}^\top \mathbf{C}^{(2)}\mathbf{y} + \frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)$$
$$= D^2 \det(\mathbf{I}_d - 4\mathbf{A})^{-1/2} \exp\left(\mathbf{x}^\top \mathbf{C}^{(1)}\mathbf{x} + \mathbf{y}^\top \mathbf{C}^{(2)}\mathbf{y} + \frac{1}{2}(\mathbf{B}^{(1)}\mathbf{x} + \mathbf{B}^{(2)}\mathbf{y})^\top (\mathbf{I}_d - 4\mathbf{A})^{-1}(\mathbf{B}^{(1)}\mathbf{x} + \mathbf{B}^{(2)}\mathbf{y})\right)$$
$$= D^2 \det(\mathbf{I}_d - 4\mathbf{A})^{-1/2} \exp\Big(\mathbf{x}^\top \left(\mathbf{C}^{(1)} + \frac{1}{2}(\mathbf{B}^{(1)})^\top (\mathbf{I}_d - 4\mathbf{A})^{-1}\mathbf{B}^{(1)}\right)\mathbf{x}$$
$$+ \mathbf{y}^\top \left(\mathbf{C}^{(2)} + \frac{1}{2}(\mathbf{B}^{(2)})^\top (\mathbf{I}_d - 4\mathbf{A})^{-1}\mathbf{B}^{(2)}\right)\mathbf{y} + \mathbf{x}^\top (\mathbf{B}^{(1)})^\top (\mathbf{I}_d - 4\mathbf{A})^{-1}\mathbf{B}^{(2)}\mathbf{y}\Big).$$

根据这个表达式，我们确实可以得出：对于所有满足定理陈述中条件的 $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$，$\mathbb{E}_{\nu_{\mathrm{DE}}} f^{(1)}{}_{\mathrm{DE}}(\boldsymbol{\omega},\mathbf{x}) f^{(2)}{}_{\mathrm{DE}}(\boldsymbol{\omega},\mathbf{y}) = K^{(0)}(\mathbf{x},\mathbf{y})$。

接下来，我们计算方差的表达式。对于任意随机变量 Z，$\operatorname{Var} Z = \mathbb{E}Z^2 - (\mathbb{E}Z)^2$。特别地，如果$Z = f^{(1)}{}_{\mathrm{DE}}(\boldsymbol{\omega},\mathbf{x}) f^{(2)}{}_{\mathrm{DE}}(\boldsymbol{\omega},\mathbf{y})$，$\boldsymbol{\omega} \sim \nu_{\mathrm{DE}}$，则有：

$$\operatorname{Var}_{\nu_{\mathrm{DE}}} f^{(1)}_{\mathrm{DE}}(\boldsymbol{\omega},\mathbf{x}) f^{(2)}_{\mathrm{DE}}(\boldsymbol{\omega},\mathbf{y}) = \mathbb{E}_{\nu_{\mathrm{DE}}} f^{(1)}_{\mathrm{DE}}(\boldsymbol{\omega},\mathbf{x})^2 f^{(2)}_{\mathrm{DE}}(\boldsymbol{\omega},\mathbf{y})^2 - \left(\mathbb{E}_{\nu_{\mathrm{DE}}} f^{(1)}_{\mathrm{DE}}(\boldsymbol{\omega},\mathbf{x}) f^{(2)}_{\mathrm{DE}}(\boldsymbol{\omega},\mathbf{y})\right)^2$$
$$= \mathbb{E}_{\nu_{\mathrm{DE}}} f^{(1)}_{\mathrm{DE}}(\boldsymbol{\omega},\mathbf{x})^2 f^{(2)}_{\mathrm{DE}}(\boldsymbol{\omega},\mathbf{y})^2 - K^{(0)}(\mathbf{x},\mathbf{y})^2.$$

我们有：

$$\mathbb{E}_{\nu_{\mathrm{DE}}} f^{(1)}_{\mathrm{DE}}(\boldsymbol{\omega},\mathbf{x})^2 f^{(2)}_{\mathrm{DE}}(\boldsymbol{\omega},\mathbf{y})^2 = (2\pi)^{\frac{d}{2}} D^4 \int_{\mathbb{R}^d} \exp\Big(-\frac{1}{2}\|\boldsymbol{\omega}\|^2 + 4\boldsymbol{\omega}^\top \mathbf{A}\boldsymbol{\omega} + 2\boldsymbol{\omega}^\top(\mathbf{B}^{(1)}\mathbf{x} + \mathbf{B}^{(2)}\mathbf{y}) + 2\mathbf{x}^\top \mathbf{C}^{(1)}\mathbf{x}$$
$$+ 2\mathbf{y}^\top \mathbf{C}^{(2)}\mathbf{y}\Big) d\boldsymbol{\omega} = (2\pi)^{\frac{d}{2}} D^4 \exp\left(2\mathbf{x}^\top \mathbf{C}^{(1)}\mathbf{x} + 2\mathbf{y}^\top \mathbf{C}^{(2)}\mathbf{y}\right)$$

$$\times \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\boldsymbol{\omega}^\top(\mathbf{I}_d - 8\mathbf{A})\boldsymbol{\omega} + 2\boldsymbol{\omega}^\top(\mathbf{B}^{(1)}\mathbf{x} + \mathbf{B}^{(2)}\mathbf{y})\right) d\boldsymbol{\omega}.$$

上述积分的计算方法与计算 $\mathbb{E}_{\nu_{\mathrm{DE}}} f^{(1)}{}_{\mathrm{DE}}(\boldsymbol{\omega},\ \mathbf{x}) f^{(2)}{}_{\mathrm{DE}}(\boldsymbol{\omega},\ \mathbf{y})$ 相同，注意到$\mathbf{I}_d - 8\mathbf{A}$ 是正定且可逆的。结果如下:

$$\mathbb{E}_{\nu_{\mathrm{DE}}} f^{(1)}_{\mathrm{DE}}(\boldsymbol{\omega},\mathbf{x})^2 f^{(2)}_{\mathrm{DE}}(\boldsymbol{\omega},\mathbf{y})^2 = D^4 \det(\mathbf{I}_d - 8\mathbf{A})^{-1/2} \exp\Bigg(2\mathbf{x}^\top \left(\mathbf{C}^{(1)} + (\mathbf{B}^{(1)})^\top (\mathbf{I}_d - 8\mathbf{A})^{-1}\mathbf{B}^{(1)}\right)\mathbf{x}$$
$$+2\mathbf{y}^\top \left(\mathbf{C}^{(2)} + (\mathbf{B}^{(2)})^\top (\mathbf{I}_d - 8\mathbf{A})^{-1}\mathbf{B}^{(2)}\right)\mathbf{y} + 4\mathbf{x}^\top (\mathbf{B}^{(1)})^\top (\mathbf{I}_d - 8\mathbf{A})^{-1}\mathbf{B}^{(2)}\mathbf{y}\Bigg).$$

我们得出结论，定理陈述中给出的方差表达式是正确的。 □

A.2. 重要引理

下面我们证明一个将在后续证明中用到的重要引理:

引理 A.1. 考虑一个函数 $f\colon (-\infty, 1/8)$，定义为

$$f(A) = \log(1 - 4A) - \frac{1}{2}\log(1 - 8A) + \frac{\phi}{1 - 8A} \tag{21}$$

其中 $\phi \geq 0$。那么，f 在 $(-\infty, 1/8)$ 上的最小值在以下位置取得

$$A^* = \frac{1}{16}\left(1 - 2\phi - \sqrt{(2\phi+1)^2 + 8\phi}\right). \tag{22}$$

证明。令 $\gamma = (1 - 8A)^{-1} \in (0,\ +\infty)$。注意到 $\gamma \in (0,\ +\infty)$ 与$A \in (-\infty,\ 1/8)$之间存在一一对应关系。因此，我们可以在（21）中代入 $\gamma^{-1} = 1 - 8A$ 和 $1 - 4A = ((1 - 8A) + 1)/2 = (\gamma^{-1} + 1)/2 = \frac{1+\gamma}{2\gamma}$，并等价地对 γ 进行最小化:

$$\min_{\gamma \in (0,+\infty)} \quad h(\gamma) = \log\left(\frac{\gamma+1}{2\gamma}\right) + \frac{1}{2}\log\gamma + \phi\gamma = \log(\gamma+1) - \frac{1}{2}\log\gamma - \log 2 + \phi\gamma.$$

对于 $h(\cdot)$ 的导数，我们有:

$$h'(\gamma) = \phi + \frac{1}{\gamma+1} - \frac{1}{2\gamma} = \phi + \frac{\gamma - 1}{2\gamma(\gamma+1)} \tag{23}$$
$$= \frac{2\phi\gamma(\gamma+1) + \gamma - 1}{2\gamma(\gamma+1)} = \frac{2\phi\gamma^2 + (2\phi+1)\gamma - 1}{2\gamma(\gamma+1)}. \tag{24}$$

根据（23），我们看到，当 $\gamma \to 0$ 时，$h'(\gamma) \to -\infty$，并且对于所有 $\gamma > 1$，$h'(\gamma) > \phi \geq 0$。因此，我们得出 $h(\cdot)$ 在 $(0,\ +\infty)$ 上有下界，并且其全局最小值 γ^* 在 $(0,\ +\infty)$ 上存在；它是满足 $h'(\gamma^*) = 0$ 的点之一。因此，它是（24）分子中多项式的正根之一。

如果 $\phi = 0$，则（24）的分子中的多项式只有一个根 $\gamma^* = 1$，因此它是 $h(\cdot)$ 的全局最小值。如果 $\phi > 0$，则（24）的分子中的多项式有两个根:

$$\gamma^*_- = \frac{-(2\phi+1) - \sqrt{(2\phi+1)^2 + 8\phi}}{4\phi},$$
$$\gamma^*_+ = \frac{-(2\phi+1) + \sqrt{(2\phi+1)^2 + 8\phi}}{4\phi}. \tag{25}$$

注意，若 $\phi > 0$，则 $2\phi + 1 > 0$ 且 $(2\phi+1)^2 + 8\phi \geq (2\phi+1)^2$。因此，$\gamma^*_- < 0$ 且 $\gamma^*_+ > 0$。由此可知，$\gamma^* = \gamma^*_+$ 是 $h(\cdot)$ 在 $(0, +\infty)$ 上的最小值。我们将 (25) 右侧的分子和分母同乘 $(2\phi+1) + \sqrt{(2\phi+1)^2 + 8\phi} > 0$:

$$\gamma^* = \gamma^*_+ = \frac{((2\phi+1)^2 + 8\phi) - (2\phi+1)^2}{4\phi\left((2\phi+1) + \sqrt{(2\phi+1)^2 + 8\phi}\right)} = \frac{2}{2\phi + 1 + \sqrt{(2\phi+1)^2 + 8\phi}}. \tag{26}$$

注意，当 $\phi > 0$ 时，(26) 的右侧与 (25) 等价；但在 $\phi = 0$ 的情况下（即当 $\gamma^* = 1$ 时）也成立。由此可知，由于 $\gamma^* = (1-8A^*)^{-1}$，$f(\cdot)$ 在 $\mathbf{A}^* = \frac{1}{8}(1-(\gamma^*)^{-1})$ 处取最小值。显然，(22) 可直接由 (26) 得出。

□

A.3. 定理 4.2 的证明

证明。 当$\mathbf{A} = A\mathbf{I}_d$时，定理 4.1 中的条件为

$$8A < 1, \quad \frac{1}{1-4A}(\mathbf{B}^{(1)})^\top \mathbf{B}^{(2)} = \mathbf{I}_d, \quad \mathbf{C}^{(k)} = -\frac{1}{2(1-4A)}(\mathbf{B}^{(k)})^\top \mathbf{B}^{(k)}, \; D = (1-4A)^{d/4} \tag{27}$$

对于 $k \in \{1,2\}$。对于所有$\mathbf{x}$，$\mathbf{y} \in \mathbb{R}^d$，方差表达式 (15) 转化为

$$\mathrm{Var}_{\nu_{\mathrm{ADE}}} f^{(1)}_{\mathrm{ADE}}(\boldsymbol{\omega}, \mathbf{x}) f^{(2)}_{\mathrm{ADE}}(\boldsymbol{\omega}, \mathbf{y}) = D^4 (1-8A)^{-d/2} \exp\Bigg(2\mathbf{x}^\top \left(\mathbf{C}^{(1)} + \frac{1}{1-8A}(\mathbf{B}^{(1)})^\top \mathbf{B}^{(1)} \right) \mathbf{x}$$
$$+ 2\mathbf{y}^\top \left(\mathbf{C}^{(2)} + \frac{1}{1-8A}(\mathbf{B}^{(2)})^\top \mathbf{B}^{(2)} \right) \mathbf{y} + \frac{4}{1-8A} \mathbf{x}^\top (\mathbf{B}^{(1)})^\top \mathbf{B}^{(2)} \mathbf{y} \Bigg) - K^{(0)}(\mathbf{x}, \mathbf{y})^2.$$

我们在上面的方程中使用 (27)，通过 A，$\mathbf{B}^{(k)}$ 表示$\mathbf{C}^{(k)}$，并通过 A 表示 D：

$$\mathrm{Var}_{\nu_{\mathrm{ADE}}} f^{(1)}_{\mathrm{ADE}}(\boldsymbol{\omega}, \mathbf{x}) f^{(2)}_{\mathrm{ADE}}(\boldsymbol{\omega}, \mathbf{y}) = \left(\frac{1-4A}{\sqrt{1-8A}} \right)^d \exp\Bigg(\left(\frac{2}{1-8A} - \frac{1}{1-4A} \right) \mathbf{x}^\top (\mathbf{B}^{(1)})^\top \mathbf{B}^{(1)} \mathbf{x}$$
$$+ \left(\frac{2}{1-8A} - \frac{1}{1-4A} \right) \mathbf{y}^\top (\mathbf{B}^{(2)})^\top \mathbf{B}^{(2)} \mathbf{y} + \frac{4}{1-8A} \mathbf{x}^\top (\mathbf{B}^{(1)})^\top \mathbf{B}^{(2)} \mathbf{y} \Bigg) - K^{(0)}(\mathbf{x}, \mathbf{y})^2.$$

由于 $\frac{1}{1-4A}(\mathbf{B}^{(1)})^\top \mathbf{B}^{(2)}$ 是一个满秩矩阵$\mathbf{I}_d$ (27)，$\mathbf{B}^{(1)}$ 和$\mathbf{B}^{(2)}$ 均为满秩。因此，我们可以表示$\mathbf{B}^{(2)} = (1-4A)(\mathbf{B}^{(1)})^{-\top}$。另外，请注意

$$\frac{2}{1-8A} - \frac{1}{1-4A} = \frac{2-8A-1+8A}{(1-8A)(1-4A)} = (1-8A)^{-1}(1-4A)^{-1}.$$

我们利用上述恒等式和$\mathbf{B}^{(2)}$的公式，重写方差表达式：

$$\mathrm{Var}_{\nu_{\mathrm{ADE}}} f^{(1)}_{\mathrm{ADE}}(\boldsymbol{\omega}, \mathbf{x}) f^{(2)}_{\mathrm{ADE}}(\boldsymbol{\omega}, \mathbf{y}) = \left(\frac{1-4A}{\sqrt{1-8A}} \right)^d \exp\Bigg((1-8A)^{-1}(1-4A)^{-1} \mathbf{x}^\top (\mathbf{B}^{(1)})^\top \mathbf{B}^{(1)} \mathbf{x}$$
$$+ (1-8A)^{-1}(1-4A) \mathbf{y}^\top ((\mathbf{B}^{(1)})^\top \mathbf{B}^{(1)})^{-1} \mathbf{y} + 4(1-8A)^{-1}(1-4A) \mathbf{x}^\top \mathbf{y} \Bigg) - K^{(0)}(\mathbf{x}, \mathbf{y})^2.$$

我们利用上述表达式，重写 $\langle \nu, f^{(1)}, f_{(2)} \rangle = \langle \nu_{\mathrm{ADE}}, f_{(1)\,\mathrm{ADE}} \; f_{(2)\,\mathrm{ADE}} \rangle$ 中的 (14)，如下 s:

$$\overline{\mathcal{L}}(\boldsymbol{\theta}_{\mathrm{ADE}}; \mathcal{X}, \mathcal{Y}, \mathcal{T}_{\mathrm{ADE}}) = L^{-2} \sum_{1 \le i,j \le L} \log(\mathrm{Var}_{\nu_{\mathrm{ADE}}} f^{(1)}_{\mathrm{ADE}}(\boldsymbol{\omega}, \mathbf{x}^{(i)}) f^{(2)}_{\mathrm{ADE}}(\boldsymbol{\omega}, \mathbf{y}^{(j)}) + K^{(0)}(\mathbf{x}^{(i)}, \mathbf{y}^{(j)}))$$
$$= d \log(1-4A) - \frac{d}{2} \log(1-8A) + (1-8A)^{-1}(1-4A)^{-1} L^{-1} \sum_{i=1}^{L} (\mathbf{x}^{(i)})^\top (\mathbf{B}^{(1)})^\top \mathbf{B}^{(1)} \mathbf{x}^{(i)}$$
$$+ (1-8A)^{-1}(1-4A) L^{-1} \sum_{j=1}^{L} (\mathbf{y}^{(j)})^\top (\mathbf{B}^{(1)})^{-1} (\mathbf{B}^{(1)})^{-\top} \mathbf{y}^{(j)} + 4(1-8A)^{-1}(1-4A) L^{-2} \sum_{1 \le i,j \le L} (\mathbf{x}^{(i)})^\top \mathbf{y}^{(j)}. \tag{28}$$

记作$\mathbf{E} = (\mathbf{B}^{(1)})^\top \mathbf{B}^{(1)}$。那么 (28) 变为：

$$\overline{\mathcal{L}}(\boldsymbol{\theta}_{\mathrm{ADE}}; \mathcal{X}, \mathcal{Y}, \mathcal{T}_{\mathrm{ADE}}) = d \log(1-4A) - \frac{d}{2} \log(1-8A) + (1-8A)^{-1}(1-4A)^{-1} L^{-1} \sum_{i=1}^{L} (\mathbf{x}^{(i)})^\top \mathbf{E} \mathbf{x}^{(i)}$$
$$+ (1-8A)^{-1}(1-4A) L^{-1} \sum_{j=1}^{L} (\mathbf{y}^{(j)})^\top \mathbf{E}^{-1} \mathbf{y}^{(j)} + 4(1-8A)^{-1}(1-4A) L^{-2} \sum_{1 \le i,j \le L} (\mathbf{x}^{(i)})^\top \mathbf{y}^{(j)}. \tag{29}$$

接下来我们证明以下引理：

引理 A.2. 令$\mathbf{B}^{(1)*}=\sqrt{1-4A}\boldsymbol{\Sigma}^{1/2}\mathbf{U}^\top(\boldsymbol{\Lambda}^{(1)})^{-1/2}(\mathbf{Q}^{(1)})^\top$。当 A (8$A<1$) 固定时，$\mathbf{E}=\mathbf{E}^*=(\mathbf{B}^{(1)*})^\top\mathbf{B}^{(1)*}$ 关于$\mathbf{E}$最小化式（29）的右侧。

证明。我们有：

$$L^{-1}\sum_{i=1}^{L}(\mathbf{x}^{(i)})^\top\mathbf{E}\mathbf{x}^{(i)}=L^{-1}\sum_{i=1}^{L}\mathrm{Trace}((\mathbf{x}^{(i)})^\top\mathbf{E}\mathbf{x}^{(i)})=L^{-1}\sum_{i=1}^{L}\mathrm{Trace}(\mathbf{E}\mathbf{x}^{(i)}(\mathbf{x}^{(i)})^\top)$$
$$=\mathrm{Trace}\left(\mathbf{E}\left(L^{-1}\sum_{i=1}^{L}\mathbf{x}^{(i)}(\mathbf{x}^{(i)})^\top\right)\right)=\mathrm{Trace}(\mathbf{E}\mathbf{M}^{(1)})$$

其中，我们用到了迹的循环性质 Trace(·) 和迹的线性性。类似地，我们得到$L^{-1}\sum_{j=1}^{L}(\mathbf{y}^{(j)})^\top\mathbf{E}^{-1}\mathbf{y}^{(j)}=$ Trace$(\mathbf{E}^{-1}\mathbf{M}^{(2)})$。假设 A 固定，则关于$\mathbf{E}$对式（29）进行优化，可归结为以下最小化问题：

$$\min_{\mathbf{E}\in\mathbb{S}_d,\mathbf{E}\succ 0}\quad\mathcal{F}(\mathbf{E})=\beta_1\mathrm{Trace}(\mathbf{E}\mathbf{M}^{(1)})+\beta_2\mathrm{Trace}(\mathbf{E}^{-1}\mathbf{M}^{(2)})\tag{30}$$

其中，$\beta_1=(1-8A)^{-1}(1-4A)^{-1},\beta_2=(1-8A)^{-1}(1-4A)$，且约束$\mathbf{E}\in\mathbb{S}_d,\mathbf{E}\succ 0$ 由$\mathbf{E}=(\mathbf{B}^{(1)})^\top\mathbf{B}^{(1)}$ 以及$\mathbf{E}$ 可逆这一事实得出。我们有 $1-8A>0$ 且 $1-4A=(1-8A)/2+1/2>0$。因此，$\beta_1,\beta_2>0$。对于任意$\mathbf{E}\succ 0$ 和任意 $\boldsymbol{\Delta}\in\mathbb{S}_d$，存在足够小的 $t\in\mathbb{R}$，使得$\mathbf{E}+t\mathbf{B}$ 可逆，并且以下纽曼级数收敛：

$$(\mathbf{E}+t\boldsymbol{\Delta})^{-1}=\mathbf{E}^{-1}(\mathbf{I}_d+t\boldsymbol{\Delta}\mathbf{E}^{-1})^{-1}=\sum_{l=0}^{\infty}(-t)^l\mathbf{E}^{-1}(\boldsymbol{\Delta}\mathbf{E}^{-1})^l$$

我们进一步推导出：

$$\mathrm{Trace}((\mathbf{E}+t\boldsymbol{\Delta})^{-1}\mathbf{M}^{(2)})=\mathrm{Trace}\left(\left(\sum_{l=0}^{\infty}(-t)^l\mathbf{E}^{-1}(\boldsymbol{\Delta}\mathbf{E}^{-1})^l\right)\mathbf{M}^{(2)}\right)=\sum_{l=0}^{\infty}(-t)^l\mathrm{Trace}\left(\mathbf{E}^{-1}(\boldsymbol{\Delta}\mathbf{E}^{-1})^l\mathbf{M}^{(2)}\right)$$

因此，

$$\mathcal{F}(\mathbf{E}+t\boldsymbol{\Delta})=\beta_1\mathrm{Trace}((\mathbf{E}+t\boldsymbol{\Delta})\mathbf{M}^{(1)})+\beta_2\sum_{l=0}^{\infty}(-t)^l\mathrm{Trace}\left(\mathbf{E}^{-1}(\boldsymbol{\Delta}\mathbf{E}^{-1})^l\mathbf{M}^{(2)}\right)$$
$$=\beta_1\mathrm{Trace}(\mathbf{E}\mathbf{M}^{(1)})+t\beta_1\mathrm{Trace}(\boldsymbol{\Delta}\mathbf{M}^{(1)})+\beta_2\sum_{l=0}^{\infty}(-t)^l\mathrm{Trace}\left(\mathbf{E}^{-1}(\boldsymbol{\Delta}\mathbf{E}^{-1})^l\mathbf{M}^{(2)}\right).\tag{31}$$

此外，我们有：

$$\frac{\partial}{\partial t}\mathcal{F}(\mathbf{E}+t\boldsymbol{\Delta})=\beta_1\mathrm{Trace}(\boldsymbol{\Delta}\mathbf{M}^{(1)})+\beta_2\sum_{l=1}^{\infty}(-1)^l lt^{l-1}\mathrm{Trace}\left(\mathbf{E}^{-1}(\boldsymbol{\Delta}\mathbf{E}^{-1})^l\mathbf{M}^{(2)}\right),$$
$$\frac{\partial^2}{(\partial t)^2}\mathcal{F}(\mathbf{E}+t\boldsymbol{\Delta})=\beta_2\sum_{l=2}^{\infty}(-1)^l l(l-1)t^{l-2}\mathrm{Trace}\left(\mathbf{E}^{-1}(\boldsymbol{\Delta}\mathbf{E}^{-1})^l\mathbf{M}^{(2)}\right),$$
$$\frac{\partial^2}{(\partial t)^2}\mathcal{F}(\mathbf{E}+t\boldsymbol{\Delta})\bigg|_{t=0}=2\beta_2\mathrm{Trace}\left(\mathbf{E}^{-1}\boldsymbol{\Delta}\mathbf{E}^{-1}\boldsymbol{\Delta}\mathbf{E}^{-1}\mathbf{M}^{(2)}\right).\tag{32}$$

我们重新置$\mathbf{M}^{(2)}=\mathbf{Q}^{(2)}(\boldsymbol{\Lambda}^{(2)})^{1/2}(\boldsymbol{\Lambda}^{(2)})^{1/2}(\mathbf{Q}^{(2)})^\top$，并在式（32）中应用迹的循环性质：

$$\frac{\partial^2}{(\partial t)^2}\mathcal{F}(\mathbf{E}+t\boldsymbol{\Delta})\bigg|_{t=0}=2\beta_2\mathrm{Trace}\left((\boldsymbol{\Lambda}^{(2)})^{1/2}(\mathbf{Q}^{(2)})^\top\mathbf{E}^{-1}\boldsymbol{\Delta}\mathbf{E}^{-1}\boldsymbol{\Delta}\mathbf{E}^{-1}\mathbf{Q}^{(2)}(\boldsymbol{\Lambda}^{(2)})^{1/2}\right)=2\beta_2\mathrm{Trace}\left(\mathbf{T}\mathbf{E}^{-1}\mathbf{T}^\top\right)$$

其中 $\mathbf{T}=(\boldsymbol{\Lambda}^{(2)})^{1/2}(\mathbf{Q}^{(2)})^\top\mathbf{E}^{-1}\boldsymbol{\Delta}$。由于$\mathbf{E}$ 是正定的，$\mathbf{E}^{-1}$ 也是正定的，而$\mathbf{T}\mathbf{E}^{-1}\mathbf{T}^\top$ 至少是半正定的。因此，Trace $\left(\mathbf{T}\mathbf{E}^{-1}\mathbf{T}^\top\right)\geq 0$，并且 $\frac{\partial^2}{(\partial t)_2}\mathcal{F}(\mathbf{E}+t\boldsymbol{\Delta})|_{t=0}\geq 0$。我们得出，$\mathcal{F}(\mathbf{E})$ 是一个

凸函数，定义在 $\{\mathbf{E} \in \mathbb{S}_d \,|\mathbf{E} \succ 0\}$ 上。由于 $\{\mathbf{E} \in \mathbb{S}_d \,|\, \mathbf{E} \succ 0\}$ 是一个开集，（每个）全局最小值 $\mathbf{E}$ 满足（30）中的两个条件

$$1)\ \mathbf{E} \succ 0, \quad \text{and} \quad 2)\ \nabla \mathcal{F}(\mathbf{E}) = \mathbf{0}_{d\times d} \tag{33}$$

令 $t = 1$，并假设 $\mathbf{\Delta} \in \mathbb{S}_d$ 的范数足够小，使得 $\mathbf{E} + \mathbf{\Delta}$ 可逆，且 $(\mathbf{I}_d + \mathbf{\Delta}\mathbf{E}^{-1})^{-1}$ 的纽曼级数收敛。那么，对于 $t = 1$，式（31）成立：

$$\begin{aligned}
\mathcal{F}(\mathbf{E}+\mathbf{\Delta}) &= \beta_1 \mathrm{Trace}(\mathbf{E}\mathbf{M}^{(1)}) + \beta_1 \mathrm{Trace}(\mathbf{\Delta}\mathbf{M}^{(1)}) + \beta_2 \sum_{l=0}^{\infty} (-1)^l \mathrm{Trace}\left(\mathbf{E}^{-1}(\mathbf{\Delta}\mathbf{E}^{-1})^l \mathbf{M}^{(2)}\right) \\
&= \mathcal{F}(\mathbf{E}) + \beta_1 \mathrm{Trace}(\mathbf{\Delta}\mathbf{M}^{(1)}) + \beta_2 \sum_{l=1}^{\infty} (-1)^l \mathrm{Trace}\left(\mathbf{E}^{-1}(\mathbf{\Delta}\mathbf{E}^{-1})^l \mathbf{M}^{(2)}\right) \\
&= \mathcal{F}(\mathbf{E}) + \beta_1 \mathrm{Trace}(\mathbf{\Delta}\mathbf{M}^{(1)}) - \beta_2 \mathrm{Trace}\left(\mathbf{E}^{-1}\mathbf{\Delta}\mathbf{E}^{-1}\mathbf{M}^{(2)}\right) + \beta_2 \sum_{l=2}^{\infty} (-1)^l \mathrm{Trace}\left(\mathbf{E}^{-1}(\mathbf{\Delta}\mathbf{E}^{-1})^l \mathbf{M}^{(2)}\right).
\end{aligned}$$

显然，$\beta_2 \sum_{l=2}^{\infty} (-1)^l$迹 $\left(\mathbf{E}^{-1}(\mathbf{\Delta}\mathbf{E}^{-1})^l \mathbf{M}^{(2)}\right) = o(\|\mathbf{\Delta}\|)$，其中 $\|\cdot\|$ 是一个 L_2-范数。此外，利用迹的循环性质，可得：

$$\mathrm{Trace}\left(\mathbf{E}^{-1}\mathbf{\Delta}\mathbf{E}^{-1}\mathbf{M}^{(2)}\right) = \mathrm{Trace}\left(\mathbf{\Delta}\mathbf{E}^{-1}\mathbf{M}^{(2)}\mathbf{E}^{-1}\right).$$

因此，我们有：

$$\mathcal{F}(\mathbf{E}+\mathbf{\Delta}) = \mathrm{Trace}\left(\mathbf{\Delta}\left(\beta_1 \mathbf{M}^{(1)} - \beta_2 \mathbf{E}^{-1}\mathbf{M}^{(2)}\mathbf{E}^{-1}\right)\right) + o(\|\mathbf{\Delta}\|). \tag{34}$$

由于 $\mathbf{\Delta}$，$\mathbf{E}^{-1}$，$\mathbf{M}^{(1)}$，$\mathbf{M}^{(2)} \in \mathbb{S}_d$，由式（34）可知

$$\nabla \mathcal{F}(\mathbf{E}) = \beta_1 \mathbf{M}^{(1)} - \beta_2 \mathbf{E}^{-1}\mathbf{M}^{(2)}\mathbf{E}^{-1}. \tag{35}$$

令 $\mathbf{E}^* = (\mathbf{B}^{(1)*})^\top \mathbf{B}^{(1)*} \succeq 0$。注意

$$\sqrt[4]{\frac{\beta_2}{\beta_1}} = \sqrt[4]{\frac{(1-8A)^{-1}(1-4A)}{(1-8A)^{-1}(1-4A)^{-1}}} = \sqrt{1-4A}.$$

由于 $\sqrt{\beta_2/\beta_1} \neq 0$, $\mathbf{\Sigma}$, $\mathbf{U}$，$\mathbf{\Lambda}^{-1/2}$, $\mathbf{Q}^{(1)}$ 均为满秩矩阵，因此 $\mathbf{E}^*$ 也为满秩矩阵，所以 $\mathbf{E}^* \succ 0$满足（33）中的条件 1。注意到

$$\begin{aligned}
\mathbf{E}^*\mathbf{Q}^{(1)}(\mathbf{\Lambda}^{(1)})^{1/2} &= \sqrt{\beta_2/\beta_1}\mathbf{Q}^{(1)}(\mathbf{\Lambda}^{(1)})^{-1/2}\mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top(\mathbf{\Lambda}^{(1)})^{-1/2}(\mathbf{Q}^{(1)})^\top\mathbf{Q}^{(1)}(\mathbf{\Lambda}^{(1)})^{1/2} \\
&= \sqrt{\beta_2/\beta_1}\mathbf{Q}^{(1)}(\mathbf{\Lambda}^{(1)})^{-1/2}\mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top \\
&= \sqrt{\beta_2/\beta_1}\mathbf{Q}^{(1)}(\mathbf{\Lambda}^{(1)})^{-1/2}(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)\mathbf{V}\mathbf{U}^\top \\
&= \sqrt{\beta_2/\beta_1}\mathbf{Q}^{(1)}(\mathbf{\Lambda}^{(1)})^{-1/2}((\mathbf{\Lambda}^{(1)})^{\frac{1}{2}}(\mathbf{Q}^{(1)})^\top\mathbf{Q}^{(2)}(\mathbf{\Lambda}^{(2)})^{\frac{1}{2}})\mathbf{V}\mathbf{U}^\top \\
&= \sqrt{\beta_2/\beta_1}\mathbf{Q}^{(2)}(\mathbf{\Lambda}^{(2)})^{1/2}\mathbf{V}\mathbf{U}^\top
\end{aligned}$$

其中我们使用了 $\mathbf{E}^*$, $\mathbf{U}$，$\mathbf{\Sigma}$, $\mathbf{V}$ 的定义，以及 $\mathbf{Q}^{(1)}$,$\mathbf{Q}^{(2)}$, $\mathbf{U}$, $\mathbf{V}$的正交性。因此，我们推得

$$\begin{aligned}
\beta_1\mathbf{E}^*\mathbf{M}^{(1)}\mathbf{E}^* &= \beta_1\mathbf{E}^*\mathbf{Q}^{(1)}(\mathbf{\Lambda}^{(1)})^{1/2}\left((\mathbf{\Lambda}^{(1)})^{1/2}(\mathbf{Q}^{(1)})^\top\mathbf{E}^*\right) = \beta_1\frac{\beta_2}{\beta_1}\mathbf{Q}^{(2)}(\mathbf{\Lambda}^{(2)})^{1/2}\left((\mathbf{\Lambda}^{(2)})^{1/2}(\mathbf{Q}^{(2)})^\top\right) \\
&= \beta_2\mathbf{M}^{(2)}
\end{aligned} \tag{36}$$

根据 $\mathbf{Q}^{(1)}$, $\mathbf{\Lambda}^{(1)}$, $\mathbf{Q}^{(2)}$, $\mathbf{\Lambda}^{(2)}$ 的定义，并利用 $\mathbf{V}$和 $\mathbf{U}$的正交性。将（36）左乘和右乘（$\mathbf{E}^*$）$^{-1}$，我们推得

$$\beta_1\mathbf{M}^{(1)} = \beta_2(\mathbf{E}^*)^{-1}\mathbf{M}^{(2)}(\mathbf{E}^*)^{-1}$$

或者换句话说，$\nabla\mathcal{F}(\mathbf{E}^*) = \mathbf{0}_{d\times d}$ 且式 (33) 中的条件 2 也得到满足。我们得出结论，式 (30) 的全局最小值在$\mathbf{E}^*$处取得。 □

根据引理 A.2，在固定 A 时，$\mathbf{B}^{(1)}=\mathbf{B}^{(1)*}$ 是式 (28) 右侧的全局最小值。事实上，若存在$\mathbf{B}^{(1)}$ 使式 (28) 的值更小，则 $\mathbf{E}=(\mathbf{B}^{(1)})^\top\mathbf{B}^{(1)}$ 会使式 (29) 右侧的值更小。此外，根据定义，这个$\mathbf{E}$ 是正定的（注意$\mathbf{B}^{(1)}$ 是非奇异的），这与引理 A.2 矛盾。

在 (29) 中用$\mathbf{E}^*$ 替代$\mathbf{E}$，对应于固定 A 时 $\mathcal{L}(\boldsymbol{\theta}_{\mathrm{AGE}};\alpha,\mathcal{X},\mathcal{Y},\mathcal{T}_{\mathrm{AGE}})$ 的最小值。接下来，我们要对 A 最小化这个表达式。记 $\mathbf{F}=\mathbf{Q}^{(1)}(\mathbf{\Lambda}^{(1)})^{-1/2}\mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top(\mathbf{\Lambda}^{(1)})^{-1/2}(\mathbf{Q}^{(1)})^\top$。那么$\mathbf{E}^*=(1-4A)\mathbf{F}$，其中$\mathbf{F}$ 不依赖于 A。我们将$\mathbf{E}^*$ 代入 (29)，得到：

$$\begin{aligned}
&d\log(1-4A)-\frac{d}{2}\log(1-8A)+(1-8A)^{-1}(1-4A)^{-1}\mathrm{Trace}((1-4A)\mathbf{F}\mathbf{M}^{(1)})\\
&+(1-8A)^{-1}(1-4A)\mathrm{Trace}((1-4A)^{-1}\mathbf{F}^{-1}\mathbf{M}^{(2)})+4(1-8A)^{-1}(1-4A)L^{-2}\sum_{1\le i,j\le L}(\mathbf{x}^{(i)})^\top\mathbf{y}^{(j)}\\
&=d\log(1-4A)-\frac{d}{2}\log(1-8A)+(1-8A)^{-1}\mathrm{Trace}(\mathbf{F}\mathbf{M}^{(1)})+(1-8A)^{-1}\mathrm{Trace}(\mathbf{F}^{-1}\mathbf{M}^{(2)})\\
&+2\left(1+(1-8A)^{-1}\right)d\mu^{(3)}
\end{aligned}\tag{37}$$

这里我们还将替换

$$L^{-2}\sum_{1\le i,j\le L}(\mathbf{x}^{(i)})^\top\mathbf{y}^{(j)}=L^{-2}\left(\sum_{i=1}^{L}\mathbf{x}^{(i)}\right)^\top\left(\sum_{j=1}^{L}\mathbf{y}^{(j)}\right)=d\mu^{(3)}$$

and

$$(1-8A)^{-1}(1-4A)=\frac{(1-8A)+1}{2(1-8A)}=\frac{1}{2}\left(1+(1-8A)^{-1}\right)$$

根据 (36)，且由于$\mathbf{F}=\sqrt{\beta_1/\beta_2\mathbf{E}^*}$，我们得出$\mathbf{F}\mathbf{M}^{(1)}\mathbf{F}=\mathbf{M}^{(2)}$，或者$\mathbf{M}^{(1)}\mathbf{F}=\mathbf{F}^{-1}\mathbf{M}^{(2)}$。利用迹的循环性质，我们得到：

$$\mathrm{Trace}(\mathbf{F}\mathbf{M}^{(1)})=\mathrm{Trace}(\mathbf{M}^{(1)}\mathbf{F})=\mathrm{Trace}(\mathbf{F}^{-1}\mathbf{M}^{(2)}).$$

根据 定义 $\mathbf{F}$ $\mathbf{\Lambda}^{(1)}$、$\mathbf{Q}^{(1)}$ 并利用$\mathbf{Q}^{(1)}$、$\mathbf{U}$的循环性质和正交性，我们有 ：

$$\begin{aligned}
\mathrm{Trace}(\mathbf{F}\mathbf{M}^{(1)})&=\mathrm{Trace}\left(\mathbf{Q}^{(1)}(\mathbf{\Lambda}^{(1)})^{-1/2}\mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top(\mathbf{\Lambda}^{(1)})^{-1/2}\mathbf{Q}^{(1)})^\top\left(\mathbf{Q}^{(1)}\mathbf{\Lambda}^{(1)}(\mathbf{Q}^{(1)})^\top\right)\right)\\
&=\mathrm{Trace}\left(\mathbf{Q}^{(1)}(\mathbf{\Lambda}^{(1)})^{-1/2}\mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top(\mathbf{\Lambda}^{(1)})^{1/2}(\mathbf{Q}^{(1)})^\top\right)\\
&=\mathrm{Trace}\left(\mathbf{\Sigma}\mathbf{U}^\top(\mathbf{\Lambda}^{(1)})^{1/2}(\mathbf{Q}^{(1)})^\top\mathbf{Q}^{(1)}(\mathbf{\Lambda}^{(1)})^{-1/2}\mathbf{U}\right)\\
&=\mathrm{Trace}(\mathbf{\Sigma})=\sum_{l=1}^{d}\mathbf{\Sigma}_{l,l}.
\end{aligned}$$

因此，(37) 最终变为：

$$\begin{aligned}
&d\log(1-4A)-\frac{d}{2}\log(1-8A)+2(1-8A)^{-1}\sum_{l=1}^{d}\mathbf{\Sigma}_{l,l}+2\left(1+(1-8A)^{-1}\right)d\mu^{(3)}\\
&=d\left(\log(1-4A)-\frac{1}{2}\log(1-8A)+2(1-8A)^{-1}\left(d^{-1}\sum_{l=1}^{d}\mathbf{\Sigma}_{l,l}+\mu^{(3)}\right)+2\mu^{(3)}\right).
\end{aligned}\tag{38}$$

接下来，我们使用引理 A.1 ($\phi=d^{-1}\sum_{l=1}^{d}\mathbf{\Sigma}_{l,\ l}+\mu^{(3)}\ge 0$) 推导出使 (38) 最小化的 A 的表达式。该表达式与定理陈述中的表达式一致。$\mathbf{B}^{(2)}$、$\mathbf{C}^{(1)}$、$\mathbf{C}^{(2)}$ 的表达式可直接由 (27)、最优$\mathbf{B}^{(1)}=\mathbf{B}^{(1)*}$ 和 A 得出。(16) 可由 (38) 得出。证明至此完成。 □

A.4. 定理 4.3 的证明

证明。 结合$\mathbf{B}^{(1)}=\mathbf{B}^{(2)}=\mathbf{B}$和$\mathbf{C}^{(1)}=\mathbf{C}^{(2)}=\mathbf{C}$，定理 4.1 的条件可写为

$$8\mathbf{A}\prec\mathbf{I}_d,\quad \mathbf{B}^\top(\mathbf{I}_d-4\mathbf{A})^{-1}\mathbf{B}=\mathbf{I}_d,\quad \mathbf{C}=-\frac{1}{2}\mathbf{B}^\top(\mathbf{I}_d-4\mathbf{A})^{-1}\mathbf{B}=-\frac{1}{2}\mathbf{I}_d,\quad D=\det(\mathbf{I}_d-4\mathbf{A})^{1/4}.\tag{39}$$

记$\mathbf{Q} = (\mathbf{I}_d - 4\mathbf{A})^{-1/2}\mathbf{B} \in \mathbb{R}^{d\times d}$。根据式 (39)，$\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_d$，即$\mathbf{Q} \in \mathbb{O}_d$。我们利用式 (39) 重写式 (15)，然后代入$\mathbf{B} = (\mathbf{I}_d - 4\mathbf{A})^{1/2}\mathbf{Q}$:

$$\begin{aligned}\mathrm{Var}_{\nu_{\mathrm{SDE}}} f^{(1)}_{\mathrm{SDE}}(\boldsymbol{\omega},\mathbf{x}) f^{(2)}_{\mathrm{SDE}}(\boldsymbol{\omega},\mathbf{y}) &= \det(\mathbf{I}_d - 4\mathbf{A})\det(\mathbf{I}_d - 8\mathbf{A})^{-1/2}\exp\Big(-\|\mathbf{x}\|^2 + 2\mathbf{x}^\top\mathbf{B}^\top(\mathbf{I}_d - 8\mathbf{A})^{-1}\mathbf{B}\mathbf{x}\\ &\quad -\|\mathbf{y}\|^2 + 2\mathbf{y}^\top\mathbf{B}^\top(\mathbf{I}_d - 8\mathbf{A})^{-1}\mathbf{B}\mathbf{y} + 4\mathbf{x}^\top\mathbf{B}^\top(\mathbf{I}_d - 8\mathbf{A})^{-1}\mathbf{B}\mathbf{y}\Big) - K^{(0)}(\mathbf{x},\mathbf{y})^2\\ &= \det(\mathbf{I}_d - 4\mathbf{A})^{1/4}\det(\mathbf{I}_d - 8\mathbf{A})^{-1/2}\exp\Big(-\|\mathbf{x}\|^2 - 2\mathbf{x}^\top\mathbf{Q}^\top\mathbf{E}\mathbf{Q}\mathbf{x} - \|\mathbf{y}\|^2 - 2\mathbf{y}^\top\mathbf{Q}^\top\mathbf{E}\mathbf{Q}\mathbf{y} - 4\mathbf{x}^\top\mathbf{Q}^\top\mathbf{E}\mathbf{Q}\mathbf{y}\Big)\\ &\quad -K^{(0)}(\mathbf{x},\mathbf{y})^2 \end{aligned} \tag{40}$$

其中我们记:

$$\begin{aligned}\mathbf{E} &= -(\mathbf{I}_d - 4\mathbf{A})^{1/2}(\mathbf{I}_d - 8\mathbf{A})^{-1}(\mathbf{I}_d - 4\mathbf{A})^{1/2} = -(\mathbf{I}_d - 4\mathbf{A})(\mathbf{I}_d - 8\mathbf{A})^{-1}\\ &= -\frac{1}{2}\left((\mathbf{I}_d - 8\mathbf{A}) + \mathbf{I}_d\right)(\mathbf{I}_d - 8\mathbf{A})^{-1} = -\frac{1}{2}\mathbf{I}_d - \frac{1}{2}(\mathbf{I}_d - 8\mathbf{A})^{-1}\end{aligned} \tag{41}$$

由于其属于$\mathbb{D}_d$，且$\mathbf{A} \in \mathbb{D}_d$。接下来，我们观察到:

$$2\mathbf{x}^\top\mathbf{Q}^\top\mathbf{E}\mathbf{Q}\mathbf{x} + 2\mathbf{y}^\top\mathbf{Q}^\top\mathbf{E}\mathbf{Q}\mathbf{y} + 4\mathbf{x}^\top\mathbf{Q}^\top\mathbf{E}\mathbf{Q}\mathbf{y} = 2(\mathbf{x}+\mathbf{y})^\top\mathbf{Q}^\top\mathbf{E}\mathbf{Q}(\mathbf{x}+\mathbf{y})$$

我们将其代入 (40)，并利用所得表达式重写 (14)，得到$\langle \nu, f_{(1)}\ f^{(2)}\rangle = \langle \nu_{\mathrm{SDE}}, f^{(1)}{}_{\mathrm{SDE}}\ f^{(2)}{}_{\mathrm{SDE}}\rangle$如下:

$$\begin{aligned}\overline{\mathcal{L}}(\boldsymbol{\theta}_{\mathrm{SDE}};\mathcal{X},\mathcal{Y},\mathcal{T}_{\mathrm{SDE}}) &= L^{-2}\sum_{1\le i,j\le L}\log(\mathrm{Var}_{\nu_{\mathrm{SDE}}} f^{(1)}_{\mathrm{SDE}}(\boldsymbol{\omega},\mathbf{x}^{(i)}) f^{(2)}_{\mathrm{SDE}}(\boldsymbol{\omega},\mathbf{y}^{(j)}) + K^{(0)}(\mathbf{x}^{(i)},\mathbf{y}^{(j)}))\\ &= \log\det(\mathbf{I}_d - 4\mathbf{A}) - \frac{1}{2}\log\det(\mathbf{I}_d - 8\mathbf{A}) - L^{-1}\sum_{i=1}^{L}\|\mathbf{x}^{(i)}\|^2 - L^{-1}\sum_{j=1}^{L}\|\mathbf{y}^{(j)}\|^2\\ &\quad -2L^{-2}\sum_{1\le i,j\le L}(\mathbf{x}^{(i)}+\mathbf{y}^{(j)})^\top\mathbf{Q}^\top\mathbf{E}\mathbf{Q}(\mathbf{x}^{(i)}+\mathbf{y}^{(j)}).\end{aligned} \tag{42}$$

利用线性性和迹的循环性质，我们可推出

$$\begin{aligned}L^{-2}\sum_{1\le i,j\le L}(\mathbf{x}^{(i)}+\mathbf{y}^{(j)})^\top\mathbf{Q}^\top\mathbf{E}\mathbf{Q}(\mathbf{x}^{(i)}+\mathbf{y}^{(j)}) &= L^{-2}\sum_{1\le i,j\le L}\mathrm{Trace}\left((\mathbf{x}^{(i)}+\mathbf{y}^{(j)})^\top\mathbf{Q}^\top\mathbf{E}\mathbf{Q}(\mathbf{x}^{(i)}+\mathbf{y}^{(j)})\right)\\ &= L^{-2}\sum_{1\le i,j\le L}\mathrm{Trace}\left(\mathbf{Q}^\top\mathbf{E}\mathbf{Q}(\mathbf{x}^{(i)}+\mathbf{y}^{(j)})(\mathbf{x}^{(i)}+\mathbf{y}^{(j)})^\top\right)\\ &= \mathrm{Trace}\left(\mathbf{Q}^\top\mathbf{E}\mathbf{Q}\left(L^{-2}\sum_{1\le i,j\le L}(\mathbf{x}^{(i)}+\mathbf{y}^{(j)})(\mathbf{x}^{(i)}+\mathbf{y}^{(j)})^\top\right)\right)\end{aligned}$$

注意到

$$\begin{aligned}L^{-2}\sum_{1\le i,j\le L}(\mathbf{x}^{(i)}+\mathbf{y}^{(j)})(\mathbf{x}^{(i)}+\mathbf{y}^{(j)})^\top &= L^{-2}\sum_{1\le i,j\le L}\left(\mathbf{x}^{(i)}(\mathbf{x}^{(i)})^\top + \mathbf{x}^{(i)}(\mathbf{y}^{(j)})^\top + \mathbf{y}^{(j)}(\mathbf{x}^{(i)})^\top + \mathbf{y}^{(j)}(\mathbf{x}^{(j)})^\top\right)\\ &= L^{-1}\sum_{i=1}^{L}\mathbf{x}^{(i)}(\mathbf{x}^{(i)})^\top + \left(L^{-1}\sum_{i=1}^{L}\mathbf{x}^{(i)}\right)\left(L^{-1}\sum_{j=1}^{L}\mathbf{y}^{(j)}\right)^\top + \left(L^{-1}\sum_{j=1}^{L}\mathbf{y}^{(j)}\right)\left(L^{-1}\sum_{i=1}^{L}\mathbf{x}^{(i)}\right)^\top\\ &\quad + L^{-1}\sum_{j=1}^{L}\mathbf{y}^{(j)}(\mathbf{y}^{(j)})^\top = \mathbf{M}^{(1)} + \boldsymbol{\mu}^{(4)}(\boldsymbol{\mu}^{(5)})^\top + \boldsymbol{\mu}^{(5)}(\boldsymbol{\mu}^{(4)})^\top + \mathbf{M}^{(2)}.\end{aligned}$$

记$\mathbf{N} = \mathbf{M}^{(1)} + \boldsymbol{\mu}^{(4)}(\boldsymbol{\mu}^{(5)})^\top + \boldsymbol{\mu}^{(5)}(\boldsymbol{\mu}^{(4)})^\top + \mathbf{M}^{(2)}$。由此可得

$$\overline{\mathcal{L}}(\boldsymbol{\theta}_{\mathrm{SDE}};\mathcal{X},\mathcal{Y},\mathcal{T}_{\mathrm{SDE}}) = \log\det(\mathbf{I}_d - 4\mathbf{A}) - \frac{1}{2}\log\det(\mathbf{I}_d - 8\mathbf{A}) - L^{-1}\sum_{i=1}^{L}\|\mathbf{x}^{(i)}\|^2 - L^{-1}\sum_{j=1}^{L}\|\mathbf{y}^{(j)}\|^2$$

$$-2\mathrm{Trace}\left(\mathbf{Q}^\top\mathbf{EQN}\right). \tag{43}$$

固定$\mathbf{A}$ 后，我们对$\mathbf{Q}$ 最小化（43）式的右侧，这等价于在固定$\mathbf{A}$的情况下对$\mathbf{B}$ 最小化$\mathcal{L}(\boldsymbol{\theta}_{\mathrm{SDE}};\ \mathcal{X},\mathcal{Y},\mathcal{T}_{\mathrm{SDE}})$，因为$\mathbf{B}$ 与$\mathbf{Q}$之间存在一一对应关系。再次利用迹的循环性质，这等价于最大化：

$$\mathrm{Trace}\left(\mathbf{Q}^\top\mathbf{EQN}\right)=\mathrm{Trace}\left(\mathbf{EQNQ}^\top\right) \tag{44}$$

关于$\mathbf{Q}$。我们先证明以下引理：

引理 A.3. 假设$\mathbf{E}$ 的对角元素各不相同， $\mathbf{\Lambda}^{(3)}$也满足同样条件。令 $\mathbf{\Pi}\in\{0,1\}^{d\times d}$ 为一个置换矩阵，用于按降序排列$\mathbf{E}$ 的对角元素（即通过应用 $\mathbf{\Pi E\Pi}^\top$），对应的置换为 $\boldsymbol{\pi}\in\mathbb{N}^d$。令$\mathbf{Q}^*=\mathbf{\Pi}^\top(\mathbf{Q}^{(3)})^\top\in\mathbb{O}_d$。则有：

$$\mathrm{Trace}\left(\mathbf{EQ}^*\mathbf{N}(\mathbf{Q}^*)^\top\right)=\sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\mathbf{\Lambda}^{(3)}_{l,l} \tag{45}$$

$$=\sup_{\mathbf{Q}\in\mathbb{O}_d}\mathrm{Trace}\left(\mathbf{EQNQ}^\top\right) \tag{46}$$

证明。 首先，我们有：

$$\mathrm{Trace}\left(\mathbf{EQ}^*\mathbf{N}(\mathbf{Q}^*)^\top\right)=\mathrm{Trace}\left(\mathbf{E\Pi}^\top(\mathbf{Q}^{(3)})^\top\mathbf{NQ}^{(3)}\mathbf{\Pi}\right)=\mathrm{Trace}\left(\mathbf{E\Pi}^\top\mathbf{\Lambda}^{(3)}\mathbf{\Pi}\right)=\mathrm{Trace}\left(\mathbf{\Pi E\Pi}^\top\mathbf{\Lambda}^{(3)}\right) \tag{47}$$

$$=\sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\mathbf{\Lambda}^{(3)}_{l,l}, \tag{48}$$

即满足式 (45)。
求 $\sup_{\mathbf{Q}\in\mathbb{O}_d}\mathrm{Trace}\left(\mathbf{EQNQ}^\top\right)$ 的优化问题已得到充分研究（Brockett，1991）。根据定义， $\mathbf{\Lambda}^{(3)}$的特征值位于主对角线上，而$\mathbf{E}\in\mathbb{D}_d$ 因此其特征值也位于主对角线上。随后，正如（Brockett，1991）中所证明的，当$\mathbf{E}$ 和$\mathbf{N}$的特征值各不相同时，$\mathbf{Q}^*$ 确实是该问题的全局最大值。由此证明了式 (46)。

□

接下来，我们证明一个 引理 A.3 的推广，其中$\mathbf{E}$ 和 $\mathbf{\Lambda}^{(3)}$ 的对角元素不一定互不相同 t:

引理 A.4. 设 $\mathbf{\Pi}\in\{0,1\}^{d\times d}$ 是一个置换矩阵，用于将$\mathbf{E}$ 的对角元素按任意非升序 排列，对应于一个置换 $\boldsymbol{\pi}\in\mathbb{N}^d$（即通过应用 $\mathbf{\Pi E\Pi}^\top$）。令 $\mathbf{Q}^*=\mathbf{\Pi}^\top(\mathbf{Q}^{(3)})^\top\in\mathbb{O}_d$。则有：

$$\mathrm{Trace}\left(\mathbf{EQ}^*\mathbf{N}(\mathbf{Q}^*)^\top\right)=\sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\mathbf{\Lambda}^{(3)}_{l,l} \tag{49}$$

$$=\sup_{\mathbf{Q}\in\mathbb{O}_d}\mathrm{Trace}\left(\mathbf{EQNQ}^\top\right) \tag{50}$$

证明。 与 (47-48) 同理，我们证明 $\mathrm{Trace}\left(\mathbf{EQ}^*\mathbf{N}(\mathbf{Q}^*)^\top\right)=\sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\ \boldsymbol{\pi}_l}\mathbf{\Lambda}^{(3)}_{l\ l\ ,\ ,}$，即 (49) 成立。接下来，我们证明对任意$\mathbf{Q}\in\mathbb{O}_d$,

$$\mathrm{Trace}\left(\mathbf{EQNQ}^\top\right)\le\sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\mathbf{\Lambda}^{(3)}_{l,l}. \tag{51}$$

这将推出 (50)。
我们的证明采用反证法。首先，我们可以假设$\mathbf{E}$ $\mathbf{\Lambda}^{(3)}$ 是非零矩阵，否则（50）式显然成立。由于 Trace $($ $\mathbf{EQNQ}^\top)$ 是关于$\mathbf{Q}$ 的连续函数，且 $\mathbb{O}_d$ 是紧致的，因此 $\sup_{\mathbf{Q}\in\mathbb{O}_d}$Trace $\left(\mathbf{EQNQ}^\top\right)$ 有限。假设存在 $\delta>0$ 使得

$$\delta=\sup_{\mathbf{Q}\in\mathbb{O}_d}\mathrm{Trace}\left(\mathbf{EQNQ}^\top\right)-\sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\mathbf{\Lambda}^{(3)}_{l,l}. \tag{52}$$

令 $\widetilde{\mathbf{E}}$ $\widetilde{\mathbf{\Lambda}}^{(3)} \in \mathbb{D}_d$ 为对角线上各值互不相同的矩阵，使得

$$\|\widetilde{\mathbf{E}}-\mathbf{E}\|_{\mathrm{F}} \leq \min\left(\|\mathbf{E}\|_{\mathrm{F}}, \frac{\delta}{12\|\mathbf{\Lambda}^{(3)}\|_{\mathrm{F}}}\right), \quad \|\widetilde{\mathbf{\Lambda}}^{(3)}-\mathbf{\Lambda}^{(3)}\|_{\mathrm{F}} \leq \frac{\delta}{12\|\mathbf{E}\|_{\mathrm{F}}} \tag{53}$$

其中 $\|\cdot\|_{\mathrm{F}}$ 表示弗罗贝尼乌斯范数，$\|\mathbf{E}\|_{\mathrm{F}}, \|\mathbf{\Lambda}^{(3)}\|_{\mathrm{F}} \neq 0$，因为这些是非零矩阵。此外，我们假设 $\widetilde{\mathbf{\Lambda}}^{(3)}$ 的对角元素按降序排列；除 $\mathbf{\Lambda}^{(3)}$外，$\boldsymbol{\pi}$ 还将 $\widetilde{\mathbf{E}}$ 的元素按非升序（降序）排列。显然，这样的 $\widetilde{\mathbf{E}}$、$\widetilde{\mathbf{\Lambda}}^{(3)}$ 可通过对 $\mathbf{E}$、$\mathbf{\Lambda}^{(3)}$进行微小扰动得到。此外，记$\widetilde{\mathbf{N}} = \mathbf{Q}^{(3)}\widetilde{\mathbf{\Lambda}}^{(3)}(\mathbf{Q}^{(3)})^\top$。由于 $\mathbb{O}_d$是一个紧致闭集，且 $\mathrm{Trace}\left(\mathbf{EQNQ}^\top\right)$ 是$\mathbf{Q}$的连续函数，因此存在$\mathbf{Q}^{**} \in \mathbb{O}_d$ 使得

$$\mathrm{Trace}\left(\mathbf{E}\mathbf{Q}^{**}\mathbf{N}(\mathbf{Q}^{**})^\top\right) = \sup_{\mathbf{Q}\in\mathbb{O}_d} \mathrm{Trace}\left(\mathbf{EQNQ}^\top\right). \tag{54}$$

根据 $\widetilde{\mathbf{E}}$，$\widetilde{\mathbf{\Lambda}}^{(3)}$，$\widetilde{\mathbf{N}}$ 的定义，我们有:

$$\begin{aligned}
&\mathrm{Trace}\left(\mathbf{E}\mathbf{Q}^{**}\mathbf{N}(\mathbf{Q}^{**})^\top\right) - \mathrm{Trace}\left(\widetilde{\mathbf{E}}\mathbf{Q}^{**}\widetilde{\mathbf{N}}(\mathbf{Q}^{**})^\top\right) \\
=& \left(\mathrm{Trace}\left(\mathbf{E}\mathbf{Q}^{**}\mathbf{N}(\mathbf{Q}^{**})^\top\right) - \mathrm{Trace}\left(\mathbf{E}\mathbf{Q}^{**}\widetilde{\mathbf{N}}(\mathbf{Q}^{**})^\top\right)\right) + \left(\mathrm{Trace}\left(\mathbf{E}\mathbf{Q}^{**}\widetilde{\mathbf{N}}(\mathbf{Q}^{**})^\top\right) - \mathrm{Trace}\left(\widetilde{\mathbf{E}}\mathbf{Q}^{**}\widetilde{\mathbf{N}}(\mathbf{Q}^{**})^\top\right)\right) \\
=& \mathrm{Trace}\left(\mathbf{E}\mathbf{Q}^{**}\left(\mathbf{N}-\widetilde{\mathbf{N}}\right)(\mathbf{Q}^{**})^\top\right) + \mathrm{Trace}\left(\left(\mathbf{E}-\widetilde{\mathbf{E}}\right)\mathbf{Q}^{**}\widetilde{\mathbf{N}}(\mathbf{Q}^{**})^\top\right).
\end{aligned}$$

接下来，我们对两项分别应用柯西-施瓦茨不等式:

$$\begin{aligned}
\mathrm{Trace}\left(\mathbf{E}\mathbf{Q}^{**}\left(\mathbf{N}-\widetilde{\mathbf{N}}\right)(\mathbf{Q}^{**})^\top\right) &\leq \|(\mathbf{Q}^{**})^\top\mathbf{E}\|_{\mathrm{F}}\|(\mathbf{N}-\widetilde{\mathbf{N}})(\mathbf{Q}^{**})^\top\|_{\mathrm{F}} = \|\mathbf{E}\|_{\mathrm{F}}\|\mathbf{N}-\widetilde{\mathbf{N}}\|_{\mathrm{F}}, \\
\mathrm{Trace}\left(\left(\mathbf{E}-\widetilde{\mathbf{E}}\right)\mathbf{Q}^{**}\widetilde{\mathbf{N}}(\mathbf{Q}^{**})^\top\right) &\leq \|(\mathbf{Q}^{**})^\top(\mathbf{E}-\widetilde{\mathbf{E}})\|_{\mathrm{F}}\|\widetilde{\mathbf{N}}(\mathbf{Q}^{**})^\top\|_{\mathrm{F}} = \|\mathbf{E}-\widetilde{\mathbf{E}}\|_{\mathrm{F}}\|\widetilde{\mathbf{N}}\|_{\mathrm{F}}
\end{aligned}$$

这里我们利用了弗罗贝尼乌斯范数在正交矩阵乘法下的不变性。再次利用这一不变性，我们推得

$$\begin{aligned}
\|\mathbf{N}-\widetilde{\mathbf{N}}\|_{\mathrm{F}} &= \|\mathbf{Q}^{(3)}(\mathbf{\Lambda}^{(3)}-\widetilde{\mathbf{\Lambda}}^{(3)})(\mathbf{Q}^{(3)})^\top\|_{\mathrm{F}} = \|\mathbf{\Lambda}^{(3)}-\widetilde{\mathbf{\Lambda}}^{(3)}\|_{\mathrm{F}}, \\
\|\widetilde{\mathbf{N}}\|_{\mathrm{F}} &= \|\mathbf{Q}^{(3)}\widetilde{\mathbf{\Lambda}}^{(3)}(\mathbf{Q}^{(3)})^\top\|_{\mathrm{F}} = \|\widetilde{\mathbf{\Lambda}}^{(3)}\|_{\mathrm{F}}.
\end{aligned}$$

由此可得

$$\text{迹}\left(\mathbf{E}\mathbf{Q}^{**}\mathbf{N}(\mathbf{Q}^{**})^\top\right) \leq \mathrm{Trace}\left(\widetilde{\mathbf{E}}\mathbf{Q}^{**}\widetilde{\mathbf{N}}(\mathbf{Q}^{**})^\top\right) + \|\mathbf{E}\|_{\mathrm{F}}\|\mathbf{\Lambda}^{(3)}-\widetilde{\mathbf{\Lambda}}^{(3)}\|_{\mathrm{F}} + \|\mathbf{E}-\widetilde{\mathbf{E}}\|_{\mathrm{F}}\|\widetilde{\mathbf{\Lambda}}^{(3)}\|_{\mathrm{F}} \tag{55}$$

接下来，我们对$\mathbf{E} = \widehat{\mathbf{E}}$，$\mathbf{\Lambda}^{(3)} = \widehat{\mathbf{\Lambda}}^{(3)}$ 应用引理 A.3，并推得

$$\begin{aligned}
\mathrm{Trace}\left(\widetilde{\mathbf{E}}\mathbf{Q}^{**}\widetilde{\mathbf{N}}(\mathbf{Q}^{**})^\top\right) &\leq \sum_{l=1}^{d}\widetilde{\mathbf{E}}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\widetilde{\mathbf{\Lambda}}^{(3)}_{l,l} = \sum_{l=1}^{d}\left(\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\mathbf{\Lambda}^{(3)}_{l,l} + \left(\widetilde{\mathbf{E}}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}-\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\right)\mathbf{\Lambda}^{(3)}_{l,l} + \widetilde{\mathbf{E}}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\left(\widetilde{\mathbf{\Lambda}}^{(3)}_{l,l}-\mathbf{\Lambda}^{(3)}_{l,l}\right)\right) \\
&= \sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\mathbf{\Lambda}^{(3)}_{l,l} + \sum_{l=1}^{d}\left(\widetilde{\mathbf{E}}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}-\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\right)\mathbf{\Lambda}^{(3)}_{l,l} + \sum_{l=1}^{d}\widetilde{\mathbf{E}}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\left(\widetilde{\mathbf{\Lambda}}^{(3)}_{l,l}-\mathbf{\Lambda}^{(3)}_{l,l}\right) \\
&= \sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\mathbf{\Lambda}^{(3)}_{l,l} + \mathrm{Trace}\left(\mathbf{\Pi}(\widetilde{\mathbf{E}}-\mathbf{E})\mathbf{\Pi}^\top\widetilde{\mathbf{\Lambda}}^{(3)}\right) + \mathrm{Trace}\left(\mathbf{\Pi}\widetilde{\mathbf{E}}\mathbf{\Pi}^\top(\widetilde{\mathbf{\Lambda}}^{(3)}-\mathbf{\Lambda}^{(3)})\right).
\end{aligned}$$

我们再次对第二项和第三项应用柯西-施瓦茨不等式:

$$\begin{aligned}
\mathrm{Trace}\left(\mathbf{\Pi}(\widetilde{\mathbf{E}}-\mathbf{E})\mathbf{\Pi}^\top\widetilde{\mathbf{\Lambda}}^{(3)}\right) &\leq \|(\widetilde{\mathbf{E}}-\mathbf{E})\mathbf{\Pi}^\top\|_{\mathrm{F}}\|\mathbf{\Pi}^\top\widetilde{\mathbf{\Lambda}}^{(3)}\|_{\mathrm{F}} = \|\widetilde{\mathbf{E}}-\mathbf{E}\|_{\mathrm{F}}\|\widetilde{\mathbf{\Lambda}}^{(3)}\|_{\mathrm{F}}, \\
\mathrm{Trace}\left(\mathbf{\Pi}\widetilde{\mathbf{E}}\mathbf{\Pi}^\top(\widetilde{\mathbf{\Lambda}}^{(3)}-\mathbf{\Lambda}^{(3)})\right) &\leq \|\widetilde{\mathbf{E}}\mathbf{\Pi}^\top\|_{\mathrm{F}}\|\mathbf{\Pi}^\top(\widetilde{\mathbf{\Lambda}}^{(3)}-\mathbf{\Lambda}^{(3)})\|_{\mathrm{F}} = \|\widetilde{\mathbf{E}}\|_{\mathrm{F}}\|\widetilde{\mathbf{\Lambda}}^{(3)}-\mathbf{\Lambda}^{(3)}\|_{\mathrm{F}}
\end{aligned}$$

其中 我们利用弗罗贝尼乌斯范数在列置换和行置换下的不变性。由此可得 t

$$\mathrm{Trace}\left(\widetilde{\mathbf{E}}\mathbf{Q}^{**}\widetilde{\mathbf{N}}(\mathbf{Q}^{**})^\top\right) \le \sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\boldsymbol{\Lambda}^{(3)}_{l,l} + \|\widetilde{\mathbf{E}}-\mathbf{E}\|_{\mathrm{F}}\|\widetilde{\boldsymbol{\Lambda}}^{(3)}\|_{\mathrm{F}} + \|\widetilde{\mathbf{E}}\|_{\mathrm{F}}\|\widetilde{\boldsymbol{\Lambda}}^{(3)}-\boldsymbol{\Lambda}^{(3)}\|_{\mathrm{F}}.$$

将此不等式与 (55) 结合，可得：

$$\mathrm{Trace}\left(\mathbf{E}\mathbf{Q}^{**}\mathbf{N}(\mathbf{Q}^{**})^\top\right) \le \sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\boldsymbol{\Lambda}^{(3)}_{l,l} + 2\|\mathbf{E}\|_{\mathrm{F}}\|\boldsymbol{\Lambda}^{(3)}-\widetilde{\boldsymbol{\Lambda}}^{(3)}\|_{\mathrm{F}} + 2\|\mathbf{E}-\widetilde{\mathbf{E}}\|_{\mathrm{F}}\|\widetilde{\boldsymbol{\Lambda}}^{(3)}\|_{\mathrm{F}}. \tag{56}$$

接下来，利用三角不等式可推出

$$\|\widetilde{\boldsymbol{\Lambda}}^{(3)}\|_{\mathrm{F}} \le \|\boldsymbol{\Lambda}^{(3)}\|_{\mathrm{F}} + \|\widetilde{\boldsymbol{\Lambda}}^{(3)}-\boldsymbol{\Lambda}^{(3)}\|_{\mathrm{F}}.$$

因此，我们继续推导 (56)：

$$\begin{aligned}
\mathrm{Trace}\left(\mathbf{E}\mathbf{Q}^{**}\mathbf{N}(\mathbf{Q}^{**})^\top\right) &\le \sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\boldsymbol{\Lambda}^{(3)}_{l,l} + 2\|\mathbf{E}\|_{\mathrm{F}}\|\boldsymbol{\Lambda}^{(3)}-\widetilde{\boldsymbol{\Lambda}}^{(3)}\|_{\mathrm{F}} + 2\|\mathbf{E}-\widetilde{\mathbf{E}}\|_{\mathrm{F}}\left(\|\boldsymbol{\Lambda}^{(3)}\|_{\mathrm{F}} + \|\widetilde{\boldsymbol{\Lambda}}^{(3)}-\boldsymbol{\Lambda}^{(3)}\|_{\mathrm{F}}\right)\\
&= \sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\boldsymbol{\Lambda}^{(3)}_{l,l} + 2\|\mathbf{E}\|_{\mathrm{F}}\|\boldsymbol{\Lambda}^{(3)}-\widetilde{\boldsymbol{\Lambda}}^{(3)}\|_{\mathrm{F}} + 2\|\boldsymbol{\Lambda}^{(3)}\|_{\mathrm{F}}\|\mathbf{E}-\widetilde{\mathbf{E}}\|_{\mathrm{F}} + 2\|\mathbf{E}-\widetilde{\mathbf{E}}\|_{\mathrm{F}}\|\widetilde{\boldsymbol{\Lambda}}^{(3)}-\boldsymbol{\Lambda}^{(3)}\|_{\mathrm{F}}\\
&\le \sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\boldsymbol{\Lambda}^{(3)}_{l,l} + 2\|\mathbf{E}\|_{\mathrm{F}}\|\boldsymbol{\Lambda}^{(3)}-\widetilde{\boldsymbol{\Lambda}}^{(3)}\|_{\mathrm{F}} + 2\|\boldsymbol{\Lambda}^{(3)}\|_{\mathrm{F}}\|\mathbf{E}-\widetilde{\mathbf{E}}\|_{\mathrm{F}} + 2\|\mathbf{E}\|_{\mathrm{F}}\|\widetilde{\boldsymbol{\Lambda}}^{(3)}-\boldsymbol{\Lambda}^{(3)}\|_{\mathrm{F}}
\end{aligned}$$

~其中，在最后一步变换中，我们使用了 $\|\mathbf{E}-\mathbf{E}\|_{\mathrm{F}} \le \|\mathbf{E}\|_{\mathrm{F}}$，这依据的是 (53)。我们再次利用 (53) 延续这一不等式链：

$$\mathrm{Trace}\left(\mathbf{E}\mathbf{Q}^{**}\mathbf{N}(\mathbf{Q}^{**})^\top\right) \le \sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\boldsymbol{\Lambda}^{(3)}_{l,l} + \frac{2}{12}\delta + \frac{2}{12}\delta + \frac{2}{12}\delta = \sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\boldsymbol{\Lambda}^{(3)}_{l,l} + \frac{\delta}{2} < \sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\boldsymbol{\Lambda}^{(3)}_{l,l} + \delta.$$

考虑到$\mathbf{Q}^{**}$的定义 (54)，这与 (52) 矛盾。因此，(50) 得证。 □

令$\mathbf{Q}^*$ 按照引理 A.4 的陈述定义。此外，我们记 $\boldsymbol{\pi}(\mathbf{A}) = \boldsymbol{\pi}, \boldsymbol{\Pi}(\mathbf{A}) = \boldsymbol{\Pi}$，其中 $\boldsymbol{\pi}, \boldsymbol{\Pi}$ 按照引理 A.4 的陈述定义。也就是说， $\boldsymbol{\pi}(\mathbf{A})$ 表示某个置换，它将$\mathbf{E}$ 的对角元素按非升序排列。由于$\mathbf{E}$ 是 (41) 中定义的$\mathbf{A}$ 的函数，因此它是$\mathbf{A}$ 的函数。事实上，根据 (41)，我们可以看出， $\boldsymbol{\pi}(\mathbf{A})$ 表示某个置换，它将$\mathbf{A}$ 的对角元素按非降序排列。 $\boldsymbol{\Pi}(\mathbf{A})$ 表示与 $\boldsymbol{\pi}(\mathbf{A})$ 对应的置换矩阵。也就是说， $\boldsymbol{\Pi}(\mathbf{A})\mathbf{A}\boldsymbol{\Pi}(\mathbf{A})^\top$ 的对角元素按非降序排列。

令 $\mathcal{G}(\mathbf{A})$ 表示 (43) 的右侧，其中将$\mathbf{Q} = \mathbf{Q}^*$代入。也就是说， $\mathcal{G}(\mathbf{A})$ 是在$\mathbf{A}$ 固定时 $\mathcal{L}(\boldsymbol{\theta}_{\mathrm{SDE}}; \mathcal{X}, \mathcal{Y}, \mathcal{T}_{\mathrm{SDE}})$ 的最优值：

$$\begin{aligned}
\mathcal{G}(\mathbf{A}) &= \log\det(\mathbf{I}_d - 4\mathbf{A}) - \frac{1}{2}\log\det(\mathbf{I}_d - 8\mathbf{A}) - L^{-1}\sum_{i=1}^{L}\|\mathbf{x}^{(i)}\|^2 - L^{-1}\sum_{j=1}^{L}\|\mathbf{y}^{(j)}\|^2 - 2\sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}(\mathbf{A})_l,\boldsymbol{\pi}(\mathbf{A})_l}\boldsymbol{\Lambda}^{(3)}_{l,l}\\
&= \log\det(\mathbf{I}_d - 4\mathbf{A}) - \frac{1}{2}\log\det(\mathbf{I}_d - 8\mathbf{A}) - L^{-1}\sum_{i=1}^{L}\|\mathbf{x}^{(i)}\|^2 - L^{-1}\sum_{j=1}^{L}\|\mathbf{y}^{(j)}\|^2\\
&+ \sum_{l=1}^{d}\left(1 + (1 - 8\mathbf{A}_{\boldsymbol{\pi}(\mathbf{A})_l,\boldsymbol{\pi}(\mathbf{A})_l})^{-1}\right)\boldsymbol{\Lambda}^{(3)}_{l,l}
\end{aligned} \tag{57}$$

其中我们使用了$\mathbf{E}$的定义（41）。令 $\boldsymbol{\pi}^{-1}(\mathbf{A}) \in \mathbb{N}^d$ 表示 $\boldsymbol{\pi}(\mathbf{A})$的逆置换。通过重新排列求和中的各项，我们得到：

$$\sum_{l=1}^{d}\left(1 + (1 - 8\mathbf{A}_{\boldsymbol{\pi}(\mathbf{A})_l,\boldsymbol{\pi}(\mathbf{A})_l})^{-1}\right)\boldsymbol{\Lambda}^{(3)}_{l,l} = \sum_{l=1}^{d}\left(1 + (1 - 8\mathbf{A}_{l,l})^{-1}\right)\boldsymbol{\Lambda}^{(3)}_{\boldsymbol{\pi}^{-1}(\mathbf{A})_l,\boldsymbol{\pi}^{-1}(\mathbf{A})_l}.$$

因此，我们得到：

$$\mathcal{G}(\mathbf{A}) = \log\det(\mathbf{I}_d - 4\mathbf{A}) - \frac{1}{2}\log\det(\mathbf{I}_d - 8\mathbf{A}) - L^{-1}\sum_{i=1}^{L}\|\mathbf{x}^{(i)}\|^2 - L^{-1}\sum_{j=1}^{L}\|\mathbf{y}^{(j)}\|^2 + \sum_{l=1}^{d}\left(1 + (1-8\mathbf{A}_{l,l})^{-1}\right)\mathbf{\Lambda}^{(3)}_{\boldsymbol{\pi}^{-1}(\mathbf{A})_l,\boldsymbol{\pi}^{-1}(\mathbf{A})_l}. \tag{58}$$

定义一个新函数 $\mathcal{G}(\mathbf{A},\mathbf{A})$，其中 $\mathbf{A} \in \mathbb{D}_d$ 满足 $8\mathbf{A} \prec \mathbf{I}_d$，如下所示：

$$\mathcal{G}(\widehat{\mathbf{A}}, \mathbf{A}) = \log\det(\mathbf{I}_d - 4\widehat{\mathbf{A}}) - \frac{1}{2}\log\det(\mathbf{I}_d - 8\widehat{\mathbf{A}}) + \sum_{l=1}^{d}(1-8\widehat{\mathbf{A}}_{l,l})^{-1}\mathbf{\Lambda}^{(3)}_{\boldsymbol{\pi}^{-1}(\mathbf{A})_l,\boldsymbol{\pi}^{-1}(\mathbf{A})_l}.$$

根据 $\mathcal{G}(\mathbf{A},\mathbf{A})$ 的定义，我们有：

$$\mathcal{G}(\mathbf{A}) = \mathcal{G}(\mathbf{A},\mathbf{A}) - L^{-1}\sum_{i=1}^{L}\|\mathbf{x}^{(i)}\|^2 - L^{-1}\sum_{j=1}^{L}\|\mathbf{y}^{(j)}\|^2 + \sum_{l=1}^{d}\mathbf{\Lambda}^{(3)}_{\boldsymbol{\pi}^{-1}(\mathbf{A})_l,\boldsymbol{\pi}^{-1}(\mathbf{A})_l}.$$

因此，有：

$$\mathcal{G}(\mathbf{A}) \geq \inf_{\widehat{\mathbf{A}}\in\mathbb{D}_d,\, 8\widehat{\mathbf{A}}\prec\mathbf{I}_d} \mathcal{G}(\widehat{\mathbf{A}},\mathbf{A}) - L^{-1}\sum_{i=1}^{L}\|\mathbf{x}^{(i)}\|^2 - L^{-1}\sum_{j=1}^{L}\|\mathbf{y}^{(j)}\|^2 + \sum_{l=1}^{d}\mathbf{\Lambda}^{(3)}_{\boldsymbol{\pi}^{-1}(\mathbf{A})_l,\boldsymbol{\pi}^{-1}(\mathbf{A})_l}. \tag{59}$$

接下来，我们证明 $\inf_{\widehat{\mathbf{A}}\in\mathbb{D}_d, 8\widehat{\mathbf{A}}\prec\mathbf{I}_d} \mathcal{G}(\mathbf{A},\mathbf{A})$ 的解存在闭式表达式。由于 $\mathbf{A} \in \mathbb{D}_d$，我们有：$\log\det(\mathbf{I}_d - 4\widehat{\mathbf{A}}) = \sum_{l=1}^{d}\log(1-4\widehat{\mathbf{A}}_{l,l})$，$\log\det(\mathbf{I}_d - 8\widehat{\mathbf{A}}) = \sum_{l=1}^{d}\log(1-8\widehat{\mathbf{A}}_{l,l})$。此外，我们还有：

$$\mathcal{G}(\widehat{\mathbf{A}},\mathbf{A}) = \sum_{l=1}^{d}\left(\log(1-4\widehat{\mathbf{A}}_{l,l}) - \frac{1}{2}\log(1-8\widehat{\mathbf{A}}_{l,l}) + (1-8\widehat{\mathbf{A}}_{l,l})^{-1}\mathbf{\Lambda}^{(3)}_{\boldsymbol{\pi}^{-1}(\mathbf{A})_l,\boldsymbol{\pi}^{-1}(\mathbf{A})_l}\right). \tag{60}$$

由 (60) 可知，最小化 $\inf_{\widehat{\mathbf{A}}\in\mathbb{D}_d, 8\widehat{\mathbf{A}}\prec\mathbf{I}_d} \mathcal{G}(\mathbf{A},\mathbf{A})$ 可化简为关于 $\widehat{\mathbf{A}}_{l,l}$ 的 d 个相互独立的最小化问题，其中 $8\widehat{\mathbf{A}}_{l,l} < 1$。第 l 个问题 $1 \leq l \leq d$ 使用引理 A.1 求解，其中令 $\phi = \mathbf{\Lambda}^{(3)}_{\boldsymbol{\pi}^{-1}(\mathbf{A})_l,\boldsymbol{\pi}^{-1}(\mathbf{A})_l}$。令 $\mathbf{A}^{**} \in \mathbb{D}_d$ 表示相应的解。那么，对所有 $1 \leq l \leq d$，我们有：

$$\mathbf{A}^{**}_{l,l} = \frac{1}{16}\left(1 - 2\mathbf{\Lambda}^{(3)}_{\boldsymbol{\pi}^{-1}(\mathbf{A})_l,\boldsymbol{\pi}^{-1}(\mathbf{A})_l} - \sqrt{\left(2\mathbf{\Lambda}^{(3)}_{\boldsymbol{\pi}^{-1}(\mathbf{A})_l,\boldsymbol{\pi}^{-1}(\mathbf{A})_l} + 1\right)^2 + 8\mathbf{\Lambda}^{(3)}_{\boldsymbol{\pi}^{-1}(\mathbf{A})_l,\boldsymbol{\pi}^{-1}(\mathbf{A})_l}}\right). \tag{61}$$

由 (59) 可得

$$\begin{aligned}\mathcal{G}(\mathbf{A}) &\geq \mathcal{G}(\mathbf{A}^{**},\mathbf{A}) - L^{-1}\sum_{i=1}^{L}\|\mathbf{x}^{(i)}\|^2 - L^{-1}\sum_{j=1}^{L}\|\mathbf{y}^{(j)}\|^2 + \sum_{l=1}^{d}\mathbf{\Lambda}^{(3)}_{\boldsymbol{\pi}^{-1}(\mathbf{A})_l,\boldsymbol{\pi}^{-1}(\mathbf{A})_l} \\ &= \log\det(\mathbf{I}_d - 4\mathbf{A}^{**}) - \frac{1}{2}\log\det(\mathbf{I}_d - 8\mathbf{A}^{**}) - L^{-1}\sum_{i=1}^{L}\|\mathbf{x}^{(i)}\|^2 - L^{-1}\sum_{j=1}^{L}\|\mathbf{y}^{(j)}\|^2 \\ &\quad + \sum_{l=1}^{d}\left(1 + (1-8\mathbf{A}^{**}_{l,l})^{-1}\right)\mathbf{\Lambda}^{(3)}_{\boldsymbol{\pi}^{-1}(\mathbf{A})_l,\boldsymbol{\pi}^{-1}(\mathbf{A})_l}.\end{aligned} \tag{62}$$

记 $\mathbf{E}^{**} = -\frac{1}{2}\mathbf{I}_d + \frac{1}{2}(\mathbf{I}_d - 8\mathbf{A}^{**})^{-1}$。则有：

$$\sum_{l=1}^{d}\left(1 + (1-8\mathbf{A}^{**}_{l,l})^{-1}\right)\mathbf{\Lambda}^{(3)}_{\boldsymbol{\pi}^{-1}(\mathbf{A})_l,\boldsymbol{\pi}^{-1}(\mathbf{A})_l} = -2\mathrm{Trace}\left(\mathbf{E}^{**}\mathbf{\Pi}(\mathbf{A})^{-1}\mathbf{\Lambda}^{(3)}\left(\mathbf{\Pi}(\mathbf{A})^{-1}\right)^{\top}\right)$$

$$\leq -2\sum_{l=1}^{d}\mathbf{E}^{**}_{\boldsymbol{\pi}(\mathbf{A}^{**})_l,\boldsymbol{\pi}(\mathbf{A}^{**})_l}\boldsymbol{\Lambda}^{(3)}_{l,l} \tag{63}$$

其中，第二个推导步骤由引理 A.4 以及以下事实得出： $\boldsymbol{\pi}(\mathbf{A}^{**})$ 按非递减顺序对 $\mathbf{A}^{**}$ 的对角元素排序，因此它也按非递减顺序对$\mathbf{E}^{**}$ 的对角元素排序（回顾 $\boldsymbol{\pi}(\mathbf{A}^{**})$ 和 $\mathbf{E}^{**}$ 的定义）。记 $\mathbf{E}^{*}=\boldsymbol{\Pi}(\mathbf{A}^{**})\mathbf{E}^{**}\boldsymbol{\Pi}(\mathbf{A}^{**})^{\top}$。则$\mathbf{E}^{**}_{\boldsymbol{\pi}(\mathbf{A}^{**})_l,\ \boldsymbol{\pi}(\mathbf{A}^{**})_l}=\mathbf{E}^{*}_{l,\ l}$ 对所有 $1\leq l\leq d$ 和

$$\sum_{l=1}^{d}\mathbf{E}^{**}_{\boldsymbol{\pi}(\mathbf{A}^{**})_l,\boldsymbol{\pi}(\mathbf{A}^{**})_l}\boldsymbol{\Lambda}^{(3)}_{l,l}=\sum_{l=1}^{d}\mathbf{E}^{*}_{l,l}\boldsymbol{\Lambda}^{(3)}_{l,l}. \tag{64}$$

此外，我们有：

$$\begin{aligned}\mathbf{E}^{*}&=\boldsymbol{\Pi}(\mathbf{A}^{**})\left(-\frac{1}{2}\mathbf{I}_d-\frac{1}{2}(\mathbf{I}_d-8\mathbf{A}^{**})^{-1}\right)\boldsymbol{\Pi}(\mathbf{A}^{**})^{\top}=-\frac{1}{2}\mathbf{I}_d-\frac{1}{2}(\mathbf{I}_d-8\boldsymbol{\Pi}(\mathbf{A}^{**})\mathbf{A}^{**}\boldsymbol{\Pi}(\mathbf{A}^{**})^{\top})^{-1}\\&=-\frac{1}{2}\mathbf{I}_d-\frac{1}{2}(\mathbf{I}_d-8\mathbf{A}^{*})^{-1}\end{aligned}$$

其中，我们记 $\mathbf{A}^{*}=\boldsymbol{\Pi}(\mathbf{A}^{**})\mathbf{A}^{**}\boldsymbol{\Pi}(\mathbf{A}^{**})^{\top}$，即 $\mathbf{A}^{*}_{l,l}=\mathbf{A}^{**}_{\boldsymbol{\pi}(\mathbf{A}^{**})_l,\boldsymbol{\pi}(\mathbf{A}^{**})_l}$ 对所有 $1\leq l\leq d$ 都成立。根据 $\mathbf{A}^{**}$ (61) 的定义，对所有 $1\leq l\leq d$，我们有：

$$\mathbf{A}^{*}_{l,l}=\frac{1}{16}\left(1-2\boldsymbol{\Lambda}^{(3)}_{l,l}-\sqrt{\left(2\boldsymbol{\Lambda}^{(3)}_{l,l}+1\right)^2+8\boldsymbol{\Lambda}^{(3)}_{l,l}}\right). \tag{65}$$

也就是说，A^{*} 与 A 相互独立。根据 (65)，我们可以看到， $\boldsymbol{\Lambda}^{(3)}_{l\ l}$ 的值越小，$\mathrm{A}^{*}_{l,l}$的值就越大。由于$\boldsymbol{\Lambda}^{(3)}_{1\ 1},\ldots,\boldsymbol{\Lambda}^{(3)}_{d\ d}$ 按非升序排列，我们推知 $\mathrm{A}^{*}_{1,1},\ldots,\mathrm{A}^{*}_{d,d}$ 按非降序排列。根据 $\boldsymbol{\pi}(\mathrm{A}^{*})$ 的定义，对所有 $1\leq l\leq d$，我们有 $\mathrm{A}^{*}_{l,\ l}=\mathrm{A}^{*}_{\boldsymbol{\pi}(\mathrm{A}^{*})_l,\boldsymbol{\pi}(\mathrm{A}^{*})_l}$。因此，

$$\sum_{l=1}^{d}\mathbf{E}^{*}_{l,l}\boldsymbol{\Lambda}^{(3)}_{l,l}=-\frac{1}{2}\sum_{l=1}^{d}(1+(1-8\mathbf{A}^{*}_{l,l})^{-1})\boldsymbol{\Lambda}^{(3)}_{l,l}=-\frac{1}{2}\sum_{l=1}^{d}(1+(1-8\mathbf{A}^{*}_{\boldsymbol{\pi}(\mathbf{A}^{*})_l,\boldsymbol{\pi}(\mathbf{A}^{*})_l})^{-1})\boldsymbol{\Lambda}^{(3)}_{l,l}.$$

将此结果与 (64)、(63) 结合，我们可以继续推导不等式链 (62)：

$$\begin{aligned}\mathcal{G}(\mathbf{A})&\geq\log\det(\mathbf{I}_d-4\mathbf{A}^{**})-\frac{1}{2}\log\det(\mathbf{I}_d-8\mathbf{A}^{**})-L^{-1}\sum_{i=1}^{L}\|\mathbf{x}^{(i)}\|^2-L^{-1}\sum_{j=1}^{L}\|\mathbf{y}^{(j)}\|^2\\&\quad+\sum_{l=1}^{d}\left(1+(1-8\mathbf{A}^{*}_{\boldsymbol{\pi}(\mathbf{A}^{*})_l,\boldsymbol{\pi}(\mathbf{A}^{*})_l})^{-1}\right)\boldsymbol{\Lambda}^{(3)}_{l,l}\\&=\log\det(\mathbf{I}_d-4\mathbf{A}^{*})-\frac{1}{2}\log\det(\mathbf{I}_d-8\mathbf{A}*)-L^{-1}\sum_{i=1}^{L}\|\mathbf{x}^{(i)}\|^2-L^{-1}\sum_{j=1}^{L}\|\mathbf{y}^{(j)}\|^2\\&\quad+\sum_{l=1}^{d}\left(1+(1-8\mathbf{A}^{*}_{\boldsymbol{\pi}(\mathbf{A}^{*})_l,\boldsymbol{\pi}(\mathbf{A}^{*})_l})^{-1}\right)\boldsymbol{\Lambda}^{(3)}_{l,l}=\mathcal{G}(\mathbf{A}^{*})\end{aligned} \tag{66}$$

其中，在第二步变换中，我们利用了以下事实：

$$\det(\mathbf{I}_d-4\mathbf{A}^{*})=\det\left(\boldsymbol{\Pi}(\mathbf{A}^{**})(\mathbf{I}_d-4\mathbf{A}^{**})\boldsymbol{\Pi}(\mathbf{A}^{**})^{\top}\right)=\det\left(\mathbf{I}_d-4\mathbf{A}^{**}\right)$$

并且，类似地，$\det(\mathrm{I}_d-8\mathrm{A}^{*})=\det(\mathrm{I}_d-8\mathrm{A}^{**})$。在第三步变换中，我们使用了 $\mathcal{G}(\cdot)$ (57) 的定义。注意，(66) 对所有满足 $8\mathrm{A}\prec\mathrm{I}_d$ 且 $8\mathrm{A}^{*}\prec\mathrm{I}_d$ 的 $\mathrm{A}\in\mathbb{D}_d$ 都成立，因为 $8\mathrm{A}^{**}\prec\mathrm{I}_d$。我们得出结论，当针对给定的 A 最优地选择 B、C D 时， $\mathcal{L}(\boldsymbol{\theta}_{\mathrm{SDE}};\mathcal{X},\mathcal{Y},\mathcal{T}_{\mathrm{SDE}})$ 的最小值在 $\mathrm{A}=\mathrm{A}^{*}$ 时取得。正如我们已经推导出的，$\mathrm{A}=\mathrm{A}^{*}$ 的对角元素按非降序排列。因此，使用引理 A.4 的记号，E 的对角元素已经按非升序排列，并且 $\boldsymbol{\pi}=(1,\ldots,d)$, $\boldsymbol{\Pi}=\mathrm{I}_d$ 满足引理的要求。因此，当 $\mathrm{A}=\mathrm{A}^{*}$ 时，最优的 B 具有 $(\mathrm{I}_d-4\mathrm{A})^{1/2}\mathrm{Q}^{*}$ 的形式，其中 $\mathrm{Q}^{*}=\mathrm{I}_d(\mathrm{Q}^{(3)})^{\top}=(\mathrm{Q}^{(3)})^{\top}$。最优的 C 和 D 由 (39) 进一步确定。(17) 由 (66) 以及如下事实得出：如上所述，我们可以在 (66) 中用 l 替换 $\boldsymbol{\pi}(\mathrm{A}^{*})_l$， $1\leq l\leq d$。

□

B. 补充实验细节

我们使用 NumPy（Harris et 等，2020年）在 Google Colaboratory 中进行的方差比较和核分类实验 t. 对于 Transformer 设置 s，我们使用 TPU 集群和 JAX（Bradbury et al., 2018）库。

B.1. 方差比较

我们基本沿用 (Likhosherstov等人，2022) 的设置：抽取 5 对集合 $\{\mathbf{x}^{(i)}\}_{1\leq i\leq L}, \{\mathbf{y}^{(j)}\}_{1\leq j\leq L}, L=1024$。对于每一对集合，我们计算所有点对以及所有指定随机特征方法的相对方差。此外，针对 GERF、ADERF 和 SDERF，分别在每对集合上优化平移对数方差。

我们取 $M=1$，因为在本实验中 M 的值并不重要：更大的 M 只会使下方曲线发生平移。报告的曲线是对所有点对以及全部 5 对集合取均值得到的。

B.2. 核分类

与 (Likhosherstov 等人，2022) 相同，我们通过打乱原始数据集，分别取 90%、5%、5% 的对象作为训练集、验证集和测试集划分。所有随机特征方法都使用相同的数据划分。我们在 $[10^{-2}, 10^{2}]$ 上通过包含 10 个值的对数网格调整 σ。对于每个 σ 和每种随机特征类型，我们在验证和测试期间尝试 50 个随机种子来生成随机特征。测试仅针对最佳 σ 进行。图 3 报告了对 50 个随机种子取平均后的结果。我们按照 (Likhosherstov 等人，2022) 中的描述，对所有类型的随机特征都使用正交 ω，因为正交随机特征在实践中效果更好 (Choromanski等，2021年；Likhosherstov 等人，2022)。

B.3. 长序列 Transformer 的 DERF

B.3.1. 语音建模

我们的 Conformer-Transducer 变体具有以下特征：**20** 个 Conformer层，模型维度为 $=\mathbf{512}$，相对位置嵌入维度 rped 为 $=\mathbf{512}$，以及 $h=\mathbf{8}$ 个注意力头。我们使用批次大小 bs $=\mathbf{2048}$，并在 TPU 上使用 Adam优化器进行训练。对于常规的 Conformer-Transducer 训练，我们针对不同数量的随机特征进行消融研究：$m=\mathbf{8}$，**32**、**128**。在 NST 设置下，我们使用 $m=\mathbf{8}$进行实验。我们报告了常用指标：归一化词错误率（NWER）。

B.3.2. 自然语言处理

我们在两个公开可用的数据集上对 BERT 模型进行了预训练（见：表3）。遵循原始 BERT 训练方式，我们将这两个数据集中 15% 的词元进行掩码，并训练模型预测这些掩码。BERT 预训练经过 1M 次迭代后，所有方法均从完全相同的预训练检查点开始进行热启动。我们为所有基线方法（FAVOR++（Likhosherstov 等人，2022）、FAVOR+ （Choromanski等，2021年）、ELU（Katharopoulos 等人，2020）、ReLU（Choromanski等，2021年））和 FAVOR++ 使用完全相同的超参数设置。预训练的超参数见表2。我们在 GLUE 任务上进行微调，并使用预训练模型的权重进行热启动。该设置与原始 BERT 论文中的设置类似。

表2. 所有方法预训练所用基础模型的超参数

参数	值
注意力头数量	12
隐藏层数量	12
隐藏层大小	768
词元数	512
批次大小	256
M	256
预训练步数	$1M$
Loss	MLM
激活层	gelu
丢弃概率	0.1
注意力丢弃概率	0.1
优化器	Adam
学习率	10^{-4}
计算资源	8×8 TPUv3

表3. 预训练所用的数据集。

数据集	# 词元	平均文档长度
书籍（Zhu等人，2015年）	1.0B	37K
维基百科	3.1B	592