

1. DartQuant剩余问题分析：

上一个PDF我们讨论了DartQuant的一个基本假设：激活值分布都是laplacian distributed的。然后我们引入了SWD loss来自动的衡量任意均值为0的分布和均匀分布之间的距离。

但是还有一些核心的问题没有解决：

1. 为什么要target均匀分布？为什么不能target高斯分布？
2. 为什么 R_3, R_4 要使用随机的Hadamard矩阵？

1.1 均匀假设：

上个PDF里面我回答（巧妙地避开了）问题1，我当时的回答是：“因为我们想要通过让激活值均匀分布，从而最大化信息熵来最小化量化误差”。这个理论上是没什么问题的，但是我们可以看到在LLM里面，有很多dead connections(激活值为0)。也就是说现实和理论是相违背的，尽管理论上均匀分布是最好的，但是现有的LLM激活值分布不能满足理想情况。**那为什么我们不能像QLoRA那样假设激活值分布就是高斯分布的呢？理论上把Laplace分布的激活值转换成高斯分布，比转换到均匀分布容易多了。**

1.2 Random Hadamard矩阵：

对于问题2，作者的回答是： R_3 是作用于RoPE之后，而RoPE是旋转位置编码，他是位置敏感的，且不是普通的线性变化。我们无法把 R_3 穿过RoPE融合在前面的线性层权重里面。 R_4 是作用与SiLU/SwiGLU激活函数之后。非线性激活函数无法让 R_4 像 $R_{1,2}$ 那样作用在前面的权重。所以 $R_{3,4}$ 必须在推理阶段在线的计算矩阵乘法。

Hadamard矩阵具有理论最小的互不相干性，

$$\mu(H) = \max_{i,j} |H_{ij}| = 1/\sqrt{d}$$

以上是从矩阵元素视角考虑，当然更直观的理解是从向量内积（基变换）视角考虑：

$$\mu(H) = \max_{i,j} |\langle h_i, e_j \rangle| = \frac{1}{\sqrt{d}}$$

这意味着Hadamard矩阵能够把原来在基态方向的数值均匀分散到所有维度。但实际上，只有当数据的协方差矩阵接近于 $\sigma^2 I$ （各向同性）时，互不相干性才能最优发挥作用。

但是RoPE在 R_3 处，输入数据经过了RoPE之后，被引入了块对角相关性。

Definition 1(RoPE):

$$RoPE(\mathbf{x}, m) = \begin{pmatrix} \cos(m\theta) & -\sin(m\theta) \\ \sin(m\theta) & \cos(m\theta) \end{pmatrix}$$

随机的Hadamard本质上是对维度的随机加减组合，如果块对角之间存在这种强相关性的时候，随机的加减法可能导致相关维度上的Outliers叠加，反而增大了Outliers。

Question 2: 那为什么要用随机的Hadamard矩阵？没有别的支持在线推理办法了么？

因为你需要在推理阶段在线的实时的计算矩阵乘法，所以你的 $R_{3,4}$ 时间复杂度肯定要越快越好，要不然就违背量化的本质和初衷了。

2. 改进方案：

这里介绍可学习的旋转矩阵和针对于高斯分布的SWD Loss

2.1: Learnable Structured Rotation

为了解决 R_3, R_4 无法融合导致的在线计算效率问题，同时并且随机矩阵的次优越性，我们提出使用可学习的旋转矩阵进行优化。

符号约定 (Symbol Convention): 设隐藏层维度为 d 。蝶形旋转 $R_3 \in \mathbb{R}^{d \times d}$ 直接作用于完整的 d 维激活向量。定义 $K = \log_2 d$ 层蝶形操作，每层 $d/2$ 个 Givens 配对，总共 $\frac{d}{2} \log_2 d$ 个可学习角度参数。

Definition 3(Butterfly Operation): Butterfly Operation指的是将输入序列拆分后，通过特定的加减法和旋转因子乘法组合在一起的过程。

Example 4: 这种算法常用于快速傅里叶变换(FFT)中。在FFT的矩阵表示中，一个大型的离散傅里叶变换矩阵可以被分解为多个稀疏矩阵的乘积。通过这种分解，把原来需要 $O(N^2)$ 的乘法运算离散傅里叶变化简化为了 $O(N \log N)$ 次。

Definition 5 (Givens Rotation):

$$G(i, j, \theta) = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix}$$

以下是完整的从RoPE定义，到解决方案的分析：

2.1.1 RoPE分析

设 $A \in \mathbb{R}^{L \times d}$ 为输入激活矩阵，其中 L 为序列长度， d 为隐藏层维度。

我们将 A 视为 L 个行向量的集合： $A = [\mathbf{a}_1, \dots, \mathbf{a}_L]^T$ 。

对于第 m 个位置的 token \mathbf{a}_m ，根据Definition 1, RoPE 定义为：

$$\text{RoPE}(\mathbf{a}_m) = \mathcal{R}_m \mathbf{a}_m$$

其中 $\mathcal{R}_m \in \mathbb{R}^{d \times d}$ 是一个块对角矩阵 (Block Diagonal Matrix)：

$$\mathcal{R}_m = \text{diag} \left(R_m^{(0)}, R_m^{(1)}, \dots, R_m^{(d/2-1)} \right)$$

每一个 2×2 的子块 $R_m^{(k)}$ 对应第 k 个频率带：

$$R_m^{(k)} = \begin{bmatrix} \cos(m\theta_k) & -\sin(m\theta_k) \\ \sin(m\theta_k) & \cos(m\theta_k) \end{bmatrix}, \quad \theta_k = 10000^{-2k/d}$$

Remark 5: $\text{RoPE}(A)$ 的第 m 行是 $\mathbf{a}_m^T \mathcal{R}_m^T$ 。它保持了每个 2×2 子空间的模长不变 (正交变换)，但改变了其方向。

2.1.2 RoPE带来的问题

因为我们关心的是量化难度，量化难度取决于在给定数据集上特征通道的整体分布范围。因此，我们要看 $\text{RoPE}(A)$ 在整个序列 L 上的 **经验协方差矩阵 (Empirical Covariance Matrix)**。

设 $\tilde{\mathbf{a}}$ 是 RoPE 后的随机变量。其协方差矩阵其协方差矩阵 $\Sigma_{\text{rope}} \in \mathbb{R}^{d \times d}$ 为：

$$\Sigma_{\text{rope}} = \frac{1}{L} \sum_{m=1}^L (\mathcal{R}_m \mathbf{a}_m)(\mathcal{R}_m \mathbf{a}_m)^T = \frac{1}{L} \sum_{m=1}^L \mathcal{R}_m (\mathbf{a}_m \mathbf{a}_m^T) \mathcal{R}_m^T$$

固定一个频率索引 k (其余频率带类似)。假设输入在不同位置上的局部协方差近似不变 (平稳性假设，stationarity assumption)，即 $\mathbb{E}[\mathbf{a}_m^{(k)} \mathbf{a}_m^{(k)T}] \approx \Sigma_{\text{in}}^{(k)}$ 对所有 m 成立，其中 $\Sigma_{\text{in}}^{(k)} = \begin{bmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{bmatrix}$ 。

Proposition 6(RoPE第k个子块协方差矩阵的对角元素): 第 k 个子块协方差 $\Sigma_{\text{rope}}^{(k)}$ 的对角元素 (即通道方差) 为：

$$\text{Var}(\tilde{a}_{2k}) \approx \underbrace{\frac{\sigma_1^2 + \sigma_2^2}{2}}_{\text{均值项}} + \underbrace{\frac{\sigma_1^2 - \sigma_2^2}{2} \cdot \frac{1}{L} \sum_{m=1}^L \cos(2m\theta_k)}_{\text{震荡项}} \quad (*)$$

Proof: 为了书写简洁，下文令 $c = \cos(m\theta)$, $s = \sin(m\theta)$

1. 计算 $\mathcal{R}_m \Sigma_{\text{in}}$:

$$\begin{bmatrix} c & -s \\ s & c \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} c\sigma_1^2 - s\rho & c\rho - s\sigma_2^2 \\ s\sigma_1^2 + c\rho & s\rho + c\sigma_2^2 \end{bmatrix}$$

2. 乘以转置矩阵 \mathcal{R}_m^T :

展开矩阵中的每一项:

- 元素 (1,1):

$$\begin{aligned} (c\sigma_1^2 - s\rho)c + (c\rho - s\sigma_2^2)(-s) &= c^2\sigma_1^2 - sc\rho - sc\rho + s^2\sigma_2^2 \\ &= \sigma_1^2c^2 + \sigma_2^2s^2 - 2\rho sc \end{aligned}$$

- 元素 (1,2):

$$\begin{aligned} (c\sigma_1^2 - s\rho)s + (c\rho - s\sigma_2^2)c &= sc\sigma_1^2 - s^2\rho + c^2\rho - sc\sigma_2^2 \\ &= (\sigma_1^2 - \sigma_2^2)sc + \rho(c^2 - s^2) \end{aligned}$$

- 元素 (2,1): (由于是对称矩阵, 结果同 (1,2))

$$= (\sigma_1^2 - \sigma_2^2)sc + \rho(c^2 - s^2)$$

- 元素 (2,2):

$$\begin{aligned} (s\sigma_1^2 + c\rho)s + (s\rho + c\sigma_2^2)c &= s^2\sigma_1^2 + sc\rho + sc\rho + c^2\sigma_2^2 \\ &= \sigma_1^2s^2 + \sigma_2^2c^2 + 2\rho sc \end{aligned}$$

3. 第 m 个位置的协方差贡献 Σ_m (整体经验协方差为 $\frac{1}{L} \sum_m \Sigma_m$)

$$\Sigma_m = \begin{bmatrix} \sigma_1^2 \cos^2(m\theta) + \sigma_2^2 \sin^2(m\theta) - \rho \sin(2m\theta) & \frac{1}{2}(\sigma_1^2 - \sigma_2^2) \sin(2m\theta) + \rho \cos(2m\theta) \\ \frac{1}{2}(\sigma_1^2 - \sigma_2^2) \sin(2m\theta) + \rho \cos(2m\theta) & \sigma_1^2 \sin^2(m\theta) + \sigma_2^2 \cos^2(m\theta) + \rho \sin(2m\theta) \end{bmatrix}$$

4. Variance:

第一个variance和第二个variance是对称的, 我们选择一个分析即可

$$\text{Var}_1 = \sigma_1^2 \cos^2(m\theta) + \sigma_2^2 \sin^2(m\theta) - \rho \sin(2m\theta)$$

5. 用trig formula:

- $\cos^2 \alpha = \frac{1+\cos 2\alpha}{2}$
- $\sin^2 \alpha = \frac{1-\cos 2\alpha}{2}$

$$\text{Var}(\tilde{a}_{2k}) = \sigma_1^2 \left(\frac{1 + \cos 2m\theta}{2} \right) + \sigma_2^2 \left(\frac{1 - \cos 2m\theta}{2} \right) - \rho \sin 2m\theta$$

将含 σ 的项按常数项和 \cos 项合并：

$$\text{Var}(\tilde{a}_{2k}) = \underbrace{\frac{\sigma_1^2 + \sigma_2^2}{2}}_{\text{均值项}} + \underbrace{\frac{\sigma_1^2 - \sigma_2^2}{2} \cos(2m\theta)}_{\text{震荡项}} - \underbrace{\rho \sin(2m\theta)}_{\text{相关性项}}$$

Remark 7: 因为我们关注的是整体的方差，而不是仅仅只关注这一个块的方差。所以我们需要取均值。

在整个序列 L 上取均值时：

- 均值项：由于是常数，平均值保持不变。
- 震荡项： $\cos(2m\theta) \rightarrow \frac{1}{L} \sum^L \cos(2m\theta)$
- 相关性项：与震荡项的 $\cos(2m\theta_k)$ 同理， $\sin(2m\theta_k)$ 在序列上同样振荡；当 L 足够大时， $\frac{1}{L} \sum_{m=1}^L \sin(2m\theta_k) \approx 0$ ，故该项在取均值后自然消失，无需额外假设 $\rho = 0$ 。

最后得到

$$\text{Var}(\tilde{a}_{2k}) \approx \underbrace{\frac{\sigma_1^2 + \sigma_2^2}{2}}_{\text{均值项}} + \underbrace{\frac{\sigma_1^2 - \sigma_2^2}{2} \cdot \frac{1}{L} \sum_{m=1}^L \cos(2m\theta_k)}_{\text{震荡项}}$$

□

回到我们的分析，我们现在对公式*进行极端值分析：

1. Small $k, \theta_k \approx 1$:

$\cos(2m\theta_k)$ 在 $[-1, 1]$ 间快速震荡。当 L 足够大时， $\frac{1}{L} \sum \cos \rightarrow 0$ 。

$$\text{Var}_{\text{high}} \approx \frac{\sigma_1^2 + \sigma_2^2}{2}$$

2. $k \rightarrow d/2, \theta_k \rightarrow 0 \implies \cos(2m\theta_k) \approx 1$

$$\text{Var}_{\text{low}} \approx \sigma_1^2$$

也就是说 Σ_{rope} 的对角线元素（Variance），随着索引 k 增加（频率降低），方差从“均匀均值”逐渐变成“原始极端值”。

2.1.3 New Objective - 新的优化目标

因为“只有当数据的协方差矩阵接近于 $\sigma^2 I$ （各向同性）时，互不相干性才能最优发挥作用”，所以我们要找一个正交矩阵 R ，使得：

$$\text{diag}(R\Sigma_{\text{rope}}R^T) \approx \lambda \mathbf{1}$$

2.1.4 蝶形拓扑的构造

Remark (从子空间分析到维度操作的衔接): Section 2.1.2 的分析在 $d/2$ 个 RoPE 子空间的粒度上揭示了方差的非均匀性。由于每个 RoPE 子空间对应两个相邻维度 $(2k, 2k + 1)$, 子空间 k 的方差近似等于这两个维度的平均方差。因此, 子空间级别的方差不均匀性直接映射为维度级别的方差不均匀性——具体地, 维度 $2k$ 和 $2k + 1$ 共享近似相同的方差 $\text{Var}^{(k)}$, 且该方差随 k 单调变化。标准的 d 维蝶形 Givens 拓扑在前几层 (小步长) 会自然地配对这些“同方差”的相邻维度, 而在高层 (大步长) 实现跨频率带的方差迁移, 这正是我们需要的。

Lemma 8: 对于任意两个具有不同方差 λ_i, λ_j 的独立通道, 总是存在一个 2×2 的 Givens 旋转 $G(\theta)$, 使得旋转后的两个通道方差相等。

Proof: 旋转后的方差 $\lambda'_i = \lambda_i \cos^2 \theta + \lambda_j \sin^2 \theta$ 。令其等于 $(\lambda_i + \lambda_j)/2$, 解出 θ 即可 (通常 $\theta = \pi/4$)。□

推论: 只要我们能让“高能量通道”和“低能量通道”相遇 (Pairing), 我们就能中和它们。

Lemma 8 告诉我们可以用 Givens 旋转均衡任意一对通道。但我们有 d 个激活维度, 方差沿维度索引单调变化。我们需要一种**系统性的配对策略**, 使得所有通道在有限步内均衡。

Definition 9 (Butterfly Givens Layer): 设通道总数为 $N = d$ (假设 $d = 2^K$)。定义第 ℓ 层蝶形操作 B_ℓ ($\ell = 0, 1, \dots, K - 1$) 为: 将所有 d 个维度按步长 $s = 2^\ell$ 进行配对, 每对施加一个独立的 Givens 旋转。具体地, 第 ℓ 层配对索引为:

$$(i, i + 2^\ell), \quad i = 0, 1, \dots, N - 1, \quad i \bmod 2^{\ell+1} < 2^\ell$$

每一对 (i, j) 上作用一个可学习的 Givens 旋转 $G(\theta_{ij}^{(\ell)})$ 。

Lemma 10 (蝶形拓扑的完全混合性): 经过 $K = \log_2 N$ 层蝶形 Givens 旋转后, 若每层所有角度均取 $\theta = \pi/4$, 则所有通道方差收敛到全局均值:

$$\lambda_{\text{eq}} = \frac{1}{N} \sum_{k=0}^{N-1} \lambda_k$$

Proof: 归纳法。记第 ℓ 层操作后, 通道方差的集合为 $\{\lambda_i^{(\ell)}\}$ 。

Base case ($\ell = 0$): 配对 $(2k, 2k + 1)$, 由 Lemma 8, 旋转后

$$\lambda_{2k}^{(1)} = \lambda_{2k+1}^{(1)} = \frac{\lambda_{2k}^{(0)} + \lambda_{2k+1}^{(0)}}{2}$$

此时方差值最多有 $N/2$ 个不同的值 (每对内部已相等)。

Inductive step: 假设经过 ℓ 层后，每 2^ℓ 个连续通道共享相同方差（即有 $N/2^\ell$ 个不同的值）。第 $\ell + 1$ 层以步长 2^ℓ 配对，将这些不同的值两两取均值。操作后，每 $2^{\ell+1}$ 个连续通道共享相同方差，不同值的数量减半为 $N/2^{\ell+1}$ 。

经过 $K = \log_2 N$ 层后，不同值数量为 $N/2^K = 1$ ，即所有通道方差相等。由于每次操作保持方差总和不变（Givens 旋转是正交变换），最终值必为 $\frac{1}{N} \sum_k \lambda_k$ 。□

Example 11: 以 $N = 8$ （即 $d = 8, K = 3$ ）为例，三层蝶形的配对模式为：

- B_0 (**stride 1**): $(\sigma_0^2, \sigma_1^2), (\sigma_2^2, \sigma_3^2), (\sigma_4^2, \sigma_5^2), (\sigma_6^2, \sigma_7^2)$ — 相邻频率带内部均衡。

- 此时，原来的 8 个独立方差变成了 4 组“局部均值”。

- 通道 0, 1 的新方差: $V_{01}^{(1)} = \frac{\sigma_0^2 + \sigma_1^2}{2}$
 - 通道 2, 3 的新方差: $V_{23}^{(1)} = \frac{\sigma_2^2 + \sigma_3^2}{2}$
 - 通道 4, 5 的新方差: $V_{45}^{(1)} = \frac{\sigma_4^2 + \sigma_5^2}{2}$
 - 通道 6, 7 的新方差: $V_{67}^{(1)} = \frac{\sigma_6^2 + \sigma_7^2}{2}$

- B_1 (**stride 2**): $(0, 2), (1, 3), (4, 6), (5, 7)$ — 跨相邻频率组均衡

- 配对逻辑：跳过 1 个位置配对。注意，现在的输入已经是上一层的 $V^{(1)}$ 了。

- 通道 0 和 2 配对 → 混合 $V_{01}^{(1)}$ 和 $V_{23}^{(1)}$
 - 通道 1 和 3 配对 → 混合 $V_{01}^{(1)}$ 和 $V_{23}^{(1)}$ (结果同上)
 - 通道 4 和 6 配对 → 混合 $V_{45}^{(1)}$ 和 $V_{67}^{(1)}$
 - 通道 5 和 7 配对 → 混合 $V_{45}^{(1)}$ 和 $V_{67}^{(1)}$ (结果同上)。

- 通道 0, 1, 2, 3 的新方差：

$$V_{0-3}^{(2)} = \frac{V_{01}^{(1)} + V_{23}^{(1)}}{2} = \frac{\frac{\sigma_0^2 + \sigma_1^2}{2} + \frac{\sigma_2^2 + \sigma_3^2}{2}}{2} = \frac{\sigma_0^2 + \sigma_1^2 + \sigma_2^2 + \sigma_3^2}{4}$$

- 通道 4, 5, 6, 7 的新方差：

$$V_{4-7}^{(2)} = \frac{V_{45}^{(1)} + V_{67}^{(1)}}{2} = \frac{\frac{\sigma_4^2 + \sigma_5^2}{2} + \frac{\sigma_6^2 + \sigma_7^2}{2}}{2} = \frac{\sigma_4^2 + \sigma_5^2 + \sigma_6^2 + \sigma_7^2}{4}$$

- B_2 (**stride 4**): $(0, 4), (1, 5), (2, 6), (3, 7)$ — 最高频与最低频直接配对

所有通道的方差此刻全部统一：

$$V_{\text{final}}^{(3)} = \frac{V_{0-3}^{(2)} + V_{4-7}^{(2)}}{2} = \frac{\frac{\sigma_0^2 + \sigma_1^2 + \sigma_2^2 + \sigma_3^2}{4} + \frac{\sigma_4^2 + \sigma_5^2 + \sigma_6^2 + \sigma_7^2}{4}}{2} = \frac{\sum_{k=0}^7 \sigma_k^2}{8}$$

Remark 12: Lemma 11 中 $\theta = \pi/4$ 是均衡的**充分条件**，但不是最优条件。实际中 RoPE 后的方差分布并非理想的单调序列，且通道间可能存在残余相关性。因此我们不固定 $\theta = \pi/4$ ，而是让所有角度 $\{\theta_{ij}^{(\ell)}\}$ 作为可学习参数，通过数据驱动的方式优化。

2.1.6 完整旋转矩阵的构造与复杂度

Definition 13 (Learnable Butterfly Rotation): 定义 R_3 (或 R_4) 为 K 层蝶形 Givens 旋转的乘积：

$$R = B_{K-1} \cdot B_{K-2} \cdots B_1 \cdot B_0$$

其中每层 B_ℓ 由 $d/2$ 个并行的 2×2 Givens 旋转组成，总共有

$$\frac{d}{2} \times K = \frac{d}{2} \log_2 d$$

个可学习的角度参数。

Proposition 14 (复杂度分析):

	随机 Hadamard (DartQuant)	Butterfly Givens (Ours)
时间复杂度	$O(d \log d)$	$O(d \log d)$
参数量	0 (固定随机)	$O(d \log d)$, 具体为 $\frac{d}{2} \log_2 d$
正交性	✓ (by construction)	✓ (Givens 旋转之积)
可学习	✗	✓
硬件友好性	需要随机种子 + 在线生成	稀疏结构, 可并行

Proof of orthogonality: 每个 $G(\theta)$ 满足 $G^T G = I$ 。由于 B_ℓ 中的 Givens 旋转作用于不相交的维度对, B_ℓ 本身是正交矩阵。正交矩阵之积仍为正交矩阵, 故 R 正交。□

Remark 15: 时间复杂度与 Hadamard 相同, 但蝶形结构的关键优势在于:

1. **可学习性:** 角度参数通过梯度下降优化, 能适应具体模型和数据的方差分布;
2. **结构保证:** 无论参数如何更新, R 始终保持正交性, 无需额外投影或正则化;
3. **针对性:** 蝶形拓扑天然适配 RoPE 的频率带结构——低层处理子空间内部, 高层处理跨频率带的能量迁移。

2.1.7 完整算法流程:

第一阶段：定义结构 (The Architecture)

首先, 我们要搭起那个“方差混合管道”的骨架。

1. 确定维度:

设隐藏层维度为 d (例如 LLaMA-7B 中 $d = 4096$)。计算层数 $K = \log_2 d$ (例如 $\log_2 4096 = 12$ 层)。

2. 定义可学习参数集合 Θ : 在这个管道的每一个“交叉点”上，都装一个可调节的Givens 旋转角度。

- 总共有 K 层。
- 每一层有 $d/2$ 个配对。
- 所以，我们需要初始化一个参数张量 Θ ，大小为 $[K, d/2]$ 。
- 初始化策略: ButterflyQuant论文推荐使用 Identity Initialization (即所有角度 $\theta \approx 0$)，这比随机初始化或 Hadamard 初始化效果更好，因为它允许网络从“不旋转”开始慢慢学习如何旋转。

第二阶段：构建矩阵 R_3 (The Construction Algorithm)

这是核心算法逻辑。给定当前的参数 Θ ，我们如何算出对应的旋转矩阵 R_3 ? 它是 K 个稀疏矩阵的连乘:

$$R_3 = B_{K-1} \cdot B_{K-2} \cdots B_1 \cdot B_0$$

具体的构建循环如下 (伪代码逻辑):

输入: 参数集合 Θ

输出: 旋转矩阵 R_3

1. **初始化累乘矩阵:** $R_{\text{total}} = I$ (单位矩阵)。

2. **循环每一层 ℓ 从 0 到 $K - 1$:**

- 获取步长 (Stride): $stride = 2^\ell$ (即 1, 2, 4, 8...)。

- 构建当前层的稀疏矩阵 B_ℓ :

- B_ℓ 初始化为零矩阵。

- 遍历配对: 对于每一个配对 (i, j) , 其中 $j = i + s$ (根据蝶形拓扑规则生成):

- 从参数集 Θ 中取出对应的角度 $\theta = \Theta[\ell, \text{pair_index}]$ 。

- 计算 $c = \cos \theta, s = \sin \theta$ 。

- 在 B_ℓ 的 (i, i) 和 (j, j) 位置填入 c 。

- 在 B_ℓ 的 (i, j) 位置填入 s 。

- 在 B_ℓ 的 (j, i) 位置填入 $-s$ 。

- 累乘: $R_{\text{total}} = B_\ell \cdot R_{\text{total}}$ 。

3. 返回: $R_3 = R_{\text{total}}$ 。

(注: 在实际代码部署推理时, 我们不会真的把 R_3 乘出来变成一个巨大的 $d \times d$ 稠密矩阵, 而是利用刚才的循环直接用 CUDA Kernel 对输入向量 X 进行操作, 这样复杂度才是 $O(d \log d)$ 而不是 $O(d^2)$)。

第三阶段：训练优化

现在骨架有了, 参数 θ 也有了初值, 我们怎么确定这些 θ 到底应该是多少度才能把方差混得最好? 我们需要用数据来训练它。

输入:

- 一批校准数据 (Calibration Data)，例如 128 个样本的 WikiText-2。
- 经过 RoPE 后的激活值输入 X_{in} 。

算法流程：

1. 前向传播 (Forward):

- 利用当前的 Θ 构建 (或隐式应用) 旋转 R_3 。
-计算旋转后的激活值: $X_{\text{out}} = R_3 \cdot X_{\text{in}}$ 。 - 对 X_{out} 进行模拟量化 (Quantize) 和反量化 (Dequantize)，得到 \hat{X}_{out} 。

2. 计算损失 (Loss Calculation): ButterflyQuant论文使用了两个 Loss 来指导优化：

- 重构损失 ($\mathcal{L}_{\text{recon}}$): 量化前后的误差要小。

$$\|X_{\text{in}} - R_3^T \cdot \hat{X}_{\text{out}}\|^2$$

- 均匀性正则化 ($\mathcal{L}_{\text{uniform}}$): 强迫 X_{out} 的分布接近均匀分布 (或高斯分布，取决于具体设置)，这直接对应了你之前的“方差均衡”目标。

$$D_{KL}(P(X_{\text{out}}) || \text{Uniform})$$

3. 反向传播 (Backward):

- 计算 Loss 对每个角度 θ 的梯度 $\nabla_{\theta} \mathcal{L}$ 。
- 由于 Givens 旋转是连续可导的 (\sin/\cos)，梯度可以顺畅地传导回 θ 。

4. 更新参数:

$$\theta \leftarrow \theta - \eta \cdot \nabla_{\theta} \mathcal{L} \text{ (SGD更新)}.$$

2.2 针对高斯分布的 SWD Loss

回到 1.1 节的问题：为什么要 target 均匀分布？我们论证了对于 LLM 中存在大量 dead connections 的情况，将 Laplace 分布的激活值转换为高斯分布比转换为均匀分布更自然。

Proposition 16 (高斯量化的理论动机): 设量化器为均匀量化器 (uniform quantizer)，输入分布为 $p(x)$ 。对于 b -bit 量化，均方量化误差 (MSQE) 的高分辨率近似为：

$$D \approx \frac{\Delta^2}{12}, \quad \Delta = \frac{x_{\max} - x_{\min}}{2^b}$$

对于均匀分布，这已经是最优的。但对于存在 dead connections 的 LLM 激活值，有效支撑集远小于 $[x_{\min}, x_{\max}]$ ，导致大量量化 bin 被浪费。高斯分布的钟形集中性意味着我们可以用更紧凑的 clipping range，从而减小 Δ 。

Definition 17 (Gaussian SWD Loss): 给定一批激活值 $\{x_i\}_{i=1}^n$ (零均值)，将其排序得到 $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ 。目标高斯分位数为：

$$q_i = \Phi^{-1} \left(\frac{i - 0.5}{n} \right) \cdot \hat{\sigma}, \quad \hat{\sigma} = \sqrt{\frac{1}{n} \sum_i x_i^2}$$

其中 Φ^{-1} 是标准正态分布的逆 CDF。Gaussian SWD Loss 定义为：

$$\mathcal{L}_{\text{G-SWD}} = \frac{1}{n} \sum_{i=1}^n (x_{(i)} - q_i)^2$$

Remark 18: 与 DartQuant 的 uniform SWD 相比：

- Uniform SWD 的目标分位数是等间距的： $q_i^{\text{unif}} = x_{\min} + (x_{\max} - x_{\min}) \cdot \frac{i-0.5}{n}$
- Gaussian SWD 的目标分位数在中心密集、两端稀疏，自然适应了 LLM 激活值的尖峰分布
- $\hat{\sigma}$ 的使用确保了 loss 是尺度不变的：我们只要求分布的**形状**接近高斯，而非匹配某个特定的均值和方差

Remark 19 (与 QLoRA 的联系): QLoRA 的 NF4 量化方案正是基于高斯假设设计的非均匀量化 bin。我们的 Gaussian SWD Loss 可以看作是在**旋转阶段**就主动将分布塑造为高斯形状，为后续的高斯感知量化（如 NF4）提供更好的输入条件。这两者是互补的：一个优化量化器设计，一个优化输入分布。