一个最优化问题的简化

Uwni 縈鴉 uwni@example.com

我的研究过程中，遇到了如下的优化问题．现在考虑下面这种情形，我有任意的 n 阶幺正矩阵 $\boldsymbol{U} \in U(n)$ 对应了可实现的物理构造，然后有 n 阶输入方阵为 $\boldsymbol{X}$, $n = 2^m$, $\boldsymbol{X}$ 的前 m 行是 m bit 的 BPSK 信号的所有组合，数学上就是 Rademacher 空间，后 $n - m$ 行为每行都相同的 $n - m$ 个自由未定参数．也就是说 $\boldsymbol{X}$ 的每一列代表了一种 BPSK 编码的输入状态，即 Rademacher 空间 $\{\pm 1\}^m$ 中的一个元素，如若 $m = 3$ 的时候的话，$\boldsymbol{X}$ 可写成

$$\begin{pmatrix} 1 & 1 & \cdots & -1 & -1 \\ \vdots & \vdots & & \vdots & \vdots \\ 1 & -1 & \ldots & 1 & -1 \\ a_1 & a_1 & \cdots & a_1 & a_1 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{n-m} & a_{n-m} & \cdots & a_{n-m} & a_{n-m} \end{pmatrix}$$

每一行代表了一个输入信号通道．

写成分块矩阵为

$$\boldsymbol{X}(\boldsymbol{a}) = \begin{pmatrix} \boldsymbol{B} \\ \boldsymbol{A} \end{pmatrix}$$

其中 $\boldsymbol{B}$ 为 m 行的 BPSK 信号矩阵，其列枚举了所有的 m bit BPSK 信号组合，$\boldsymbol{A}$ 为 $n - m$ 行的自由参数矩阵．其各列皆为 $n - m$ 元向量 $\boldsymbol{a}$．即 $\boldsymbol{A} = \boldsymbol{a}\boldsymbol{1}_n^\mathsf{T} \in \mathbf{C}^{(n-m)\times n}$．考虑输入信号的每一种状态，即 $\boldsymbol{X}$ 的每一列 $\boldsymbol{x}_j = \left(\boldsymbol{b}_j; \boldsymbol{a}\right)^\mathsf{T}$，其中 $\boldsymbol{b}_j$ 为 $\boldsymbol{B}$ 的第 j 列．现在考虑功率矩阵 $\boldsymbol{P}$ 为输出 $\boldsymbol{U}\boldsymbol{X}$ 各元素的功率，即有

$$p_{ij} = \left|(\boldsymbol{U}\boldsymbol{X})_{ij}\right|^2 = \left|\langle \boldsymbol{u}_i, \boldsymbol{x}_j \rangle\right|^2, \quad \boldsymbol{u}_i \in \mathbf{C}^{1\times n}, \boldsymbol{x}_j \in \mathbf{C}^{n\times 1}$$

并对 $\boldsymbol{P}$ 的每一列向量归一化后成为 $\boldsymbol{P}_{\text{norm}}$． 我现在要这个物理器件在对以 $\boldsymbol{a}$ 为额外输入参数时，成为一个译码器（decoder）的功能．数学上，即对幺正矩阵 $\boldsymbol{U}$ 和向量 $\boldsymbol{a}$ 对 $\|\boldsymbol{P}_{\text{norm}} - I\|_F^2$ 求最小值优化，其中 $\boldsymbol{I}$ 为同阶的单位阵．也就是说要输出尽可能为 one-hot 的分布．通过数值计算我发现所有的局部最优值都是全局最优 (即有多个局部最优但结果都一样)，现在要通过理论分析来证明这一点．

0.1 整理目標函數

现在有几个重要的事实 显然有一点对于 $\boldsymbol{X}$ 的每一列其功率都是一样的，即对于每一列 $\boldsymbol{x}_j = \left(b_j; \boldsymbol{a}\right)^\mathsf{T}$ 有

$$L \coloneqq \left\|\boldsymbol{x}_j\right\|^2 = \left\|\boldsymbol{b}_j\right\|^2 + \|\boldsymbol{a}\|^2 = m + \|\boldsymbol{a}\|^2$$

即每一列的功率都同为 L． 那么优化问题自然便成了

$$
\begin{aligned}
\min_{\substack{\boldsymbol{U}\in U(n)\\ \boldsymbol{a}\in\mathbf{C}^{n-m}}} \|\boldsymbol{P}-L\boldsymbol{I}\|_F^2 &= \sum_{i=1}^{n}\sum_{j=1}^{n}\left(p_{ij}-L\delta_{ij}\right)^2 \\
&= \sum_{i,j}\left|p_{ij}\right|^2 - 2L\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij}\delta_{ij} + nL^2 \\
&= \sum_{i,j}\left|p_{ij}\right|^2 - 2L\sum_{i=1}^{n} p_{ii} + nL^2
\end{aligned}
$$

把 $\boldsymbol{U}$ 分块，第 i 行 $\boldsymbol{u}_i \in \mathbf{C}^{1\times n}$ 拆成

$$
\boldsymbol{u}_i = \begin{pmatrix}\boldsymbol{v}_i & \boldsymbol{w}_i\end{pmatrix}
$$

其中 $\boldsymbol{v}_i \in \mathbf{C}^{1\times m}$, $\boldsymbol{w}_i \in \mathbf{C}^{1\times(n-m)}$. 因为 $\boldsymbol{U}$ 幺正，$\|\boldsymbol{v}_i\|^2 + \|\boldsymbol{w}_i\|^2 = 1$ 而 i 行, j 列，所對應的輸出元素為

$$
y_{ij} = \langle \boldsymbol{u}_i, \boldsymbol{x}_j\rangle = \langle \boldsymbol{v}_i, \boldsymbol{b}_j\rangle + \langle \boldsymbol{w}_i, \boldsymbol{a}\rangle
$$

觀察到 $\sum_j \boldsymbol{b}_j = 0$, $\sum_j \boldsymbol{b}_j\boldsymbol{b}_j^H = n\boldsymbol{I}$， 定義 $r_i := \|\boldsymbol{v}_i\|^2$，$\boldsymbol{r} := (r_i)$, 注意到 $0 \le r_i \le 1$

$$
\sum_{i=1}^{n} r_i = \sum_{i=1}^{n} \|\boldsymbol{v}_i\|^2 = m
$$

因爲 $\boldsymbol{W}$ 为 $\boldsymbol{U}$ 的右 $(n-m)$ 列块，幺正矩阵的列正交归一化意味着 $\boldsymbol{W}^{\mathrm{H}}\boldsymbol{W} = \boldsymbol{I}_{n-m}$，所以定义

$$
\sum_{i=1}^{n} \left|\langle \boldsymbol{w}_i, \boldsymbol{a}\rangle\right|^2 = \|\boldsymbol{W}\boldsymbol{a}\|^2 = \boldsymbol{a}^{\mathrm{H}}\boldsymbol{W}^{\mathrm{H}}\boldsymbol{W}\boldsymbol{a} = \|\boldsymbol{a}\|^2
$$

對於固定的输出端口 i 在各个状态 j 下的平均功率有

$$
\frac{1}{n}\sum_{j=1}^{n} p_{ij} = \frac{1}{n}\sum_{j=1}^{n}\left|y_{ij}\right|^2 = \frac{1}{n}\Big(\sum \left|\langle \boldsymbol{v}_i, \boldsymbol{b}_j\rangle\right|^2 + |\langle \boldsymbol{w}_i, \boldsymbol{a}\rangle|^2\Big) =
$$

0.2 可行域的結構