Theoretical Analysis

1 THEOREM PROOFS

1.1 Fitting of arbitrary relationships

In this section, we aim to prove the following theorem:

Theorem 1.1. *If embeddings $\{\boldsymbol{\theta}_j\}_{j=1}^m \in \mathbb{R}^{m\times d}$ are regularized such that $||\boldsymbol{\theta}_j||_2 \leq 1, \forall j \in \{1, ..., m\}$, the rotation-based RFM can model the inner-product based interaction pattern $\boldsymbol{\Delta}_R = \boldsymbol{e}_1^{\alpha_{j,1}} \odot \boldsymbol{e}_2^{\alpha_{j,2}} \odot \cdots \odot \boldsymbol{e}_m^{\alpha_{j,m}}$, with the mean prediction deviation $\mathbb{E}(\bar{R}) = \mathbb{E}(|\bar{\boldsymbol{\Delta}}_{RFM} - \bar{\boldsymbol{\Delta}}_R|) < O(s/\sqrt{d})$. Here $s = \sum_{k=1}^m \alpha_{j,k}$ is the interaction order, $\boldsymbol{e}_j \in \mathbb{R}^d$, $j \in \{1, ..., m\}$, $\alpha_{j,k} = f(\boldsymbol{e}_j, \boldsymbol{e}_k)$, and $f : \mathbb{R}^d \times \mathbb{R}^d \to \{0, 1\}$ is any given feature dependency function.*

For the j-th feature field, each feature is represented as a one-hot vector $\boldsymbol{x}_j \in \{0, 1\}^{n_j}$, where n_j is the feature number in the j-th field, and $N = \sum_{j=1}^m n_j$ is the total number of features. Afterwards, the embedding look-up operation $E(\cdot)$ is employed to map the one-hot vector $\boldsymbol{x}_j$ to a low-dimensional embedding $\boldsymbol{e}_j$, *i.e.,* $\boldsymbol{e}_j = E(\boldsymbol{x}_j)$. Formally, the one-hot encoded vector $\boldsymbol{x}_j$ of the l-th feature in the j-th field is defined as $\boldsymbol{x}_j[k] = \left\{ \begin{array}{ll} 1, & k = l \\ 0, & k \neq l \end{array} \right\}$, and we define the index of the l-th feature in the j-th field as $\mathrm{ID}(\boldsymbol{x}_j) = \sum_{k=1}^{j-1} n_k + l$, its inverse function as $\mathrm{OneHot}(\sum_{k=1}^{j-1} n_k + l) = \boldsymbol{x}_j$, and the function $\mathcal{G}(s) = \mathrm{argmin}_j \sum_{k=1}^{j} n_k \geq s$ returns the field index of the global index s. We place all features along an axis, and the truth table $\mathcal{T}$ of the feature dependency function f is denoted as a matrix $\mathcal{T} \in \{0, 1\}^{N\times N}$, where:

$$\mathcal{T}(s,t) = \left\{ \begin{array}{ll} f\Big(E\Big(\mathrm{OneHot}(s)\Big), E\Big(\mathrm{OneHot}(t)\Big)\Big), & \mathcal{G}(s) \neq \mathcal{G}(t) \\ 0, & \mathcal{G}(s) = \mathcal{G}(t) \end{array} \right.$$

Given the input feature embeddings $\{\boldsymbol{\theta}_j\}_{j=1}^m$, and their one-hot representations $\{\boldsymbol{x}_j\}_{j=1}^m$, we can obtain the relation vector $\boldsymbol{r}_j$ of the feature $\boldsymbol{x}_j$:

$$\boldsymbol{r}_j[k] = \left\{ \begin{array}{ll} \mathcal{T}\Big(k, \mathrm{ID}(\boldsymbol{x}_j)\Big), & \mathcal{G}(k) \neq j \\ \frac{1}{2} + \frac{1}{2} \cdot \boldsymbol{x}_j[k - \sum_{l=1}^{j-1} n_k], & \mathcal{G}(k) = j \end{array} \right.$$

where $k \in \{1, ..., N\}$. The vector $\boldsymbol{r}_j$ measures the relation between the feature $\boldsymbol{e}_j$ and the features from other fields. Meanwhile, the feature dimensions of the same field are naturally masked with $\frac{1}{2}$, except for itself, which has a value of 1. We use the vector $\boldsymbol{r}_j$ as the auxiliary dimensions for the input features, $\tilde{\boldsymbol{\theta}}_j = [\boldsymbol{\theta}_j, \epsilon \cdot \boldsymbol{r}_j]$, where ϵ is a sufficiently small number. We construct the matrix $\boldsymbol{M}$ as follows:

$$\boldsymbol{M} = \left[\begin{array}{cc} \mathbb{O}_{N\times d} & \frac{\pi}{\epsilon} \cdot \boldsymbol{I}_{N\times N} \end{array} \right] \in \mathbb{R}^{N\times(d+N)}.$$

Here $\mathbb{O}$ is an all-zero matrix. We have $\boldsymbol{M}\tilde{\boldsymbol{\theta}}_j = \pi \cdot \boldsymbol{r}_j$. Here we set all query matrices and key matrices as $\boldsymbol{W}_j^Q = \boldsymbol{W}_j^K = \boldsymbol{M}, \forall j = \{1, 2, ..., m\}$, and set all the value matrices $\boldsymbol{W}_j^V$ as the identity matrix $\boldsymbol{I}$. We construct m attention heads, each measuring the relationship between the features in the j-th field ($j \in [1, m]$) and all the other features. Formally, the projection matrices $\boldsymbol{H}_j^Q, \boldsymbol{H}_j^K, \boldsymbol{H}_j^V$ are defined as:

$$\boldsymbol{H}_j^Q = \boldsymbol{H}_j^K = \left[\begin{array}{ccccccc} \mathbb{O}_{n_j\times n_1} & \cdots & \boldsymbol{I}_{n_j\times n_j} & \cdots & \mathbb{O}_{n_j\times n_m} \end{array} \right]^\top,$$
$$\boldsymbol{H}_j^V = \left[\begin{array}{ccccccc} \boldsymbol{I}_{d\times d} & \mathbb{O}_{d\times n_1} & \cdots & \mathbb{O}_{d\times n_j} & \cdots & \mathbb{O}_{d\times n_m} \end{array} \right]^\top.$$

In this way, the values are projected to the original features $\boldsymbol{V}^{(j)} = \{\boldsymbol{\theta}_k\}_{k=1}^m$. The queries and keys of the j-th head are projected to the following: $\boldsymbol{Q}^{(j)} = \boldsymbol{K}^{(j)} = \{\pi \cdot \boldsymbol{r}_k^{(j)}\}_{k=1}^m$. Assume that the one-hot vector $\boldsymbol{x}_j = [0, 0, \cdots, \underbrace{1}_{l-th \text{ element}}, 0, 0, \cdots]^\top$, the projected vector $\boldsymbol{r}_k^{(j)}$ takes the following form:

$$\boldsymbol{r}_k^{(j)} = \left\{ \begin{array}{ll} [\frac{1}{2}, \frac{1}{2}, \cdots, \underbrace{1}_{l-th \text{ element}}, \frac{1}{2}, \frac{1}{2}, \cdots]^\top, & k = j \\ [0, 1, \cdots, \underbrace{1}_{\mathcal{T}\left(\mathrm{ID}(\boldsymbol{x}_j), \mathrm{ID}(\boldsymbol{x}_k)\right)}, 0, 1, \cdots]^\top, & k \neq j \end{array} \right.$$

We set the weight vector $\boldsymbol{w} = [S, ..., S]^\top$, and $S > 0$ is a sufficiently large number. Note that $\cos(\pm\frac{\pi}{2}) = 0$. Considering the attention score from j-th query in j-th attention head, we have:

$$\begin{aligned} \alpha_{j,l}^{RFM} &= \mathrm{Sigmoid}\Big(\boldsymbol{w}^\top \cos(\pi \cdot \boldsymbol{r}_j^{(j)} - \pi \cdot \boldsymbol{r}_l^{(j)})\Big) \\ &= \mathrm{Sigmoid}\Big(S \cdot \cos\Big(\pi - \pi \cdot \mathcal{T}\Big(\mathrm{ID}(\boldsymbol{x}_j), \mathrm{ID}(\boldsymbol{x}_l)\Big)\Big)\Big) \\ &= \mathcal{T}\Big(\mathrm{ID}(\boldsymbol{x}_j), \mathrm{ID}(\boldsymbol{x}_l)\Big) = f(\boldsymbol{e}_j, \boldsymbol{e}_l). \end{aligned}$$

We have:

$$\hat{\boldsymbol{\theta}}_j = \sum_{l=1}^m \alpha_{j,l}^{RFM} \boldsymbol{\theta}_l^V = \sum_{l=1}^m f(\boldsymbol{e}_j, \boldsymbol{e}_l)\boldsymbol{\theta}_l.$$

In this scheme, we only consider the j-th query in the j-th attention head ($j \in [1, m]$), and set the weight $\boldsymbol{u}$ as the identity matrix $\boldsymbol{I}$. Omitting the activation function yields the following expression for the output of RFM:

$$\begin{aligned} \hat{y} &= \boldsymbol{u}^\top(\cos(\hat{\boldsymbol{\theta}}_j) + \sin(\hat{\boldsymbol{\theta}}_j)) \\ &= \cos(\sum_{l=1}^m f(\boldsymbol{e}_j, \boldsymbol{e}_l)\boldsymbol{\theta}_l) + \sin(\sum_{l=1}^m f(\boldsymbol{e}_j, \boldsymbol{e}_l)\boldsymbol{\theta}_l) \\ &= \mathcal{H}\Big(\cos(\sum_{l=1}^m f(\boldsymbol{e}_j, \boldsymbol{e}_l)\boldsymbol{\theta}_l) + i\sin(\sum_{l=1}^m f(\boldsymbol{e}_j, \boldsymbol{e}_l)\boldsymbol{\theta}_l)\Big) \\ &= \mathcal{H}\Big(\exp(i\sum_{l=1}^m f(\boldsymbol{e}_j, \boldsymbol{e}_l)\boldsymbol{\theta}_l)\Big) \\ &= \mathcal{H}(e^{if(\boldsymbol{e}_j,\boldsymbol{e}_1)\boldsymbol{\theta}_1} \odot e^{if(\boldsymbol{e}_j,\boldsymbol{e}_2)\boldsymbol{\theta}_2} \odot \cdots \odot e^{if(\boldsymbol{e}_j,\boldsymbol{e}_m)\boldsymbol{\theta}_m}). \end{aligned}$$

Since the construction of the order is independent of the input features $\{\boldsymbol{e}_j = \boldsymbol{\theta}_j\}_{j=1}^m$, the theorem 1.1 is proved by combining lemma A.1 in our paper.

1.2 Gradient Analysis

In this section, we analyze and compare the gradient properties of RFM with traditional feature interaction methods. Specifically, we investigate the effects of increasing the interaction order. For this purpose, we primarily increase the order of interaction by increasing the number of feature fields. (*i.e.,* denoted as m). In the subsequent theoretical analysis, we will prove that our method exhibits, at most, linear growth in the gradient with respect to the field number. Conversely, in traditional feature interaction approaches, the gradient exhibits exponential growth with respect to the field number. For ease of analysis, we consider a structure with only one layer of self-attentive rotation mechanism:

$$\begin{aligned}\boldsymbol{G} &= \text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})\\ &= \sigma\left(\text{Re}\left[(\exp(i\boldsymbol{Q})\,\text{diag}(\boldsymbol{w}))\exp(-i\boldsymbol{K})^\top\right]\right)\cdot\boldsymbol{V}\\ y &= f(\cos(\boldsymbol{G}))+f(\sin(\boldsymbol{G}))\end{aligned}$$

For ease of mathematical illustration, we use the notation $\boldsymbol{X}_j$ to denote the original input feature embedding $\boldsymbol{e}_j$. We first calculate the gradient of our approach, *i.e.,* $\dfrac{\partial y}{\partial X}$.

$$\text{Let } \boldsymbol{X}=\begin{bmatrix}\boldsymbol{X_1} & & & \\ & \boldsymbol{X_2} & & \\ & & \ddots & \\ & & & \boldsymbol{X_m}\end{bmatrix}\in\mathbb{R}^{\text{dm}\times\text{m}},$$

$$\begin{aligned}\left[\boldsymbol{W}_1^{\boldsymbol{Q}},\cdots,\boldsymbol{W}_{\boldsymbol{m}}^{\boldsymbol{Q}}\right] &= \tilde{\boldsymbol{B}}\in\mathbb{R}^{\text{d}\times\text{dm}}\\ \left[\boldsymbol{W}_1^{\boldsymbol{K}},\cdots,\boldsymbol{W}_{\boldsymbol{m}}^{\boldsymbol{K}}\right] &= \tilde{\boldsymbol{C}}\in\mathbb{R}^{\text{d}\times\text{dm}}.\\ \left[\boldsymbol{W}_1^{\boldsymbol{V}},\cdots,\boldsymbol{W}_{\boldsymbol{m}}^{\boldsymbol{V}}\right] &= \tilde{\boldsymbol{D}}\in\mathbb{R}^{\text{d}\times\text{dm}}\end{aligned}$$

So, $\boldsymbol{Q}=(\tilde{\boldsymbol{B}}\boldsymbol{X})^\top \quad \boldsymbol{K}=(\tilde{\boldsymbol{C}}\boldsymbol{X})^\top \quad \boldsymbol{V}=(\tilde{\boldsymbol{D}}\boldsymbol{X})^\top.$

Remark $\text{Re}\left[\exp(\boldsymbol{Q})\text{diag}(\boldsymbol{w})\exp(-i\boldsymbol{K})^\top\right]$ as ①.

According to Euler's formula, we have:

$$\begin{aligned}①&=\text{Re}\left\{(\cos\boldsymbol{Q}+i\sin\boldsymbol{Q})\text{diag}(\boldsymbol{w})\left[\cos\left(-\boldsymbol{K}^\top\right)+i\sin\left(-\boldsymbol{K}^\top\right)\right]\right\}\\ &=\text{Re}\left\{(\cos\boldsymbol{Q}+i\sin\boldsymbol{Q})\text{diag}(\boldsymbol{w})\left[\cos\left(\boldsymbol{K}^\top\right)-i\sin\left(\boldsymbol{K}^\top\right)\right]\right\}\\ &=\cos\boldsymbol{Q}\text{diag}(\boldsymbol{w})\cos\left(\boldsymbol{K}^\top\right)+\sin\boldsymbol{Q}\text{diag}(\boldsymbol{w})\sin\left(\boldsymbol{K}^\top\right)\\ &=\cos\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\text{diag}(\boldsymbol{w})\cos\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)+\sin\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^T\right)\text{diag}(\boldsymbol{w})\sin\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)\end{aligned}$$

$$d\boldsymbol{G}=d[\sigma(①)\cdot\boldsymbol{V}]=[d\sigma(①)]\boldsymbol{V}+\sigma(①)\cdot d\boldsymbol{V}$$

$$d\boldsymbol{V}=d\left[(\tilde{\boldsymbol{D}}\boldsymbol{X})^\top\right]=d\left(\boldsymbol{X}^\top\tilde{\boldsymbol{D}}^\top\right)=(d\boldsymbol{X})^\top\tilde{\boldsymbol{D}}^\top$$

$$d\sigma(①)=\sigma'(①)\odot d①$$

$$\begin{aligned}d①&=d\left[\cos\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\text{diag}(\boldsymbol{w})\cos\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)+\sin\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\text{diag}(\boldsymbol{w})\sin\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)\right]\\ &=d\left\{\left[\cos\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\text{diag}(\boldsymbol{w})\right]\cdot\cos\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)\right\}+d\left\{\left[\sin\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\text{diag}(\boldsymbol{w})\right]\cdot\sin\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)\right\}\\ &=\left\{d\left[\cos\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\text{diag}(\boldsymbol{w})\right]\right\}\cdot\cos\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)+\cos\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\text{diag}(\boldsymbol{w})\cdot d\cos\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)\\ &+\left\{d\left[\sin\left(\boldsymbol{X}\tilde{\boldsymbol{B}}^\top\right)\text{diag}(\boldsymbol{w})\right]\right\}\cdot\sin\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)+\sin\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\text{diag}(\boldsymbol{w})\cdot d\sin\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)\\ &=\left[d\cos\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\right]\cdot\text{diag}(\boldsymbol{w})\cdot\cos\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)+\cos\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\cdot\text{diag}(\boldsymbol{w})\cdot[d\cos\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)]\\ &+\left[d\sin\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\right]\cdot\text{diag}(\boldsymbol{w})\cdot\sin\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)+\sin\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\text{diag}(\boldsymbol{w})\cdot[d\sin\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)]\\ &=\left[-\sin\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\odot d\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\right]\cdot\text{diag}(\boldsymbol{w})\cdot\cos\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)+\cos\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\cdot\text{diag}(\boldsymbol{w})\\ &\cdot[-\sin\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)\odot d(\tilde{\boldsymbol{C}}\boldsymbol{X})]+\left[\cos\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\odot d\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\right]\cdot\text{diag}(\boldsymbol{w})\cdot\sin\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)\\ &+\sin\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\cdot\text{diag}(\boldsymbol{w})\cdot[\cos\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)\odot d(\tilde{\boldsymbol{C}}\boldsymbol{X})]\end{aligned}$$

$$\begin{aligned}d\,y&=\text{tr}\left[\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)^\top d\boldsymbol{G}\right]\\ &=\text{tr}\left(\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)^\top\cdot\left\{\left[\sigma'(①)\odot d①\right]\boldsymbol{V}+\sigma(①)\cdot\left[(d\boldsymbol{X})^\top\tilde{\boldsymbol{D}}^\top\right]\right\}\right)\\ &=\text{tr}\left\{\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)^\top\cdot\left[\sigma'(①)\odot d①\right]\cdot\boldsymbol{V}\right\}+\text{tr}\left[\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)^\top\cdot\sigma(①)\cdot(d\boldsymbol{X})^\top\tilde{\boldsymbol{D}}^\top\right]\end{aligned}$$

$$\text{tr}\left[\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)^\top\cdot\sigma(①)\cdot(d\boldsymbol{X})^\top\tilde{\boldsymbol{D}}^\top\right]=\text{tr}\left[\tilde{\boldsymbol{D}}^\top\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)^\top\cdot\sigma(①)(d\boldsymbol{X})^\top\right]$$

$$\text{For } \boldsymbol{A},\text{tr}(\boldsymbol{A})=\text{tr}(\boldsymbol{A}^\top)\quad\Longrightarrow\quad=\text{tr}\left\{(d\boldsymbol{X})\left[\tilde{\boldsymbol{D}}^\top\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)^\top\sigma(①)\right]^\top\right\}$$

$$\text{For } \boldsymbol{A},\boldsymbol{B},\text{tr}(\boldsymbol{A}\boldsymbol{B})=\text{tr}(\boldsymbol{B}\boldsymbol{A})\Longrightarrow\quad=\text{tr}\left\{\left[\tilde{\boldsymbol{D}}^\top\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)^\top\sigma(①)\right]^\top d\boldsymbol{X}\right\}$$

Remark,

$$\text{tr}\left\{\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)^\top\cdot\left[\sigma'(①)\odot d①\right]\cdot\boldsymbol{V}\right\}=\text{part1}+\text{part2}+\text{part3}+\text{part4}$$

part1

$$\begin{aligned}&=\text{tr}\left(\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)^\top\left\{\sigma'(①)\odot\left[\left[-\sin\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\odot d\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\right]\text{diag}(\boldsymbol{w})\cos\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)\right]\right\}\boldsymbol{V}\right)\\ &=\text{tr}\left(\boldsymbol{V}\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)^\top\left\{\sigma'(①)\odot\left[\left[-\sin\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\odot d\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\right]\text{diag}(\boldsymbol{w})\cos\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)\right]\right\}\right)\\ &=\text{tr}(\left\{\left[\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)\boldsymbol{V}^\top\right]\odot\sigma'(①)\right\}^\top\\ &\cdot\left\{\left[-\sin\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\odot d\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\right]\cdot\text{diag}(\boldsymbol{w})\cdot\cos\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)\right\})\\ &=\text{tr}(\text{diag}(w)\cdot\cos\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)\cdot\left\{\left[\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)\boldsymbol{V}^\top\right]\odot\sigma'(①)\right\}^\top\\ &\cdot\left[-\sin\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\odot d\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\right])\\ &=\text{tr}(\left\{\left[\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)\boldsymbol{V}^\top\right]\odot\sigma'(①)[\text{diag}(\boldsymbol{w})\cdot\cos\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)]^\top\right\}\odot\left[-\sin\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\right]\right\}^\top\\ &\cdot d\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right))\\ &=-\text{tr}(\left\{\left[\left[\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)\boldsymbol{V}^\top\right]\odot\sigma'(①)\cdot[\text{diag}(\boldsymbol{w})\cdot\cos\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)]^\top\right]\odot\sin\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\right\}^\top\\ &\cdot(d\boldsymbol{X})^\top\tilde{\boldsymbol{B}}^\top)\end{aligned}$$

Remark

$$\boldsymbol{F}=\left\{\left[\left[\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)\boldsymbol{V}^\top\right]\odot\sigma'(①)\cdot[\text{diag}(\boldsymbol{w})\cdot\cos\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)]^\top\right]\odot\sin\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\right\}^\top$$

So,

$$\begin{aligned}\text{part1}&=-\text{tr}\left(F(d\boldsymbol{X})^\top\tilde{\boldsymbol{B}}^\top\right)=-\text{tr}\left(\tilde{\boldsymbol{B}}^\top\boldsymbol{F}(d\boldsymbol{X})^\top\right)=-\text{tr}\left(d\boldsymbol{X}\cdot\left(\tilde{\boldsymbol{B}}^\top\boldsymbol{F}\right)^\top\right)\\ &=-\text{tr}\left(\left(\tilde{\boldsymbol{B}}^\top\boldsymbol{F}\right)^\top d\boldsymbol{X}\right)=\text{tr}\left(-\left(\tilde{\boldsymbol{B}}^\top\boldsymbol{F}\right)^\top d\boldsymbol{X}\right)\end{aligned}$$

$$\begin{aligned}\text{part2}&=\text{tr}(\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)^\top\cdot\left\{\sigma'(①)\odot\left[\cos\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\cdot\text{diag}(\boldsymbol{w})\cdot[-\sin\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)\odot d(\tilde{\boldsymbol{C}}\boldsymbol{X})]\right]\right\}\cdot\boldsymbol{V})\\ &=\text{tr}(\left\{\left[\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)\boldsymbol{V}^\top\right]\odot\sigma'(①)\right\}^\top\\ &\cdot\left\{\cos\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\cdot\text{diag}(\boldsymbol{w})\cdot\left[-\sin\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)\odot d\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)\right]\right\})\\ &=\text{tr}(\left[\left\{\left\{\left[\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)\boldsymbol{V}^\top\right]\odot\sigma'(①)\right\}^\top\cdot\cos\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\cdot\text{diag}(\boldsymbol{w})\right\}^\top\odot[-\sin\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)]\right]^\top\\ &\cdot d(\tilde{\boldsymbol{C}}\boldsymbol{X}))\end{aligned}$$

Remark

$$\boldsymbol{N}=\left[\left(\left\{\left[\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)\boldsymbol{V}^\top\right]\odot\sigma'(①)\right\}^\top\cos\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\cdot\text{diag}(\boldsymbol{w})\right)^\top\odot[-\sin\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)]\right]^\top$$

So,

$$\text{part2}=\text{tr}(\boldsymbol{N}\tilde{\boldsymbol{C}}d\boldsymbol{X})=\text{tr}\left(\left[(\boldsymbol{N}\tilde{\boldsymbol{C}})^\top\right]^\top d\boldsymbol{X}\right)=\text{tr}\left[\left(\tilde{\boldsymbol{C}}^\top\boldsymbol{N}^\top\right)^\top d\boldsymbol{X}\right]$$

part3

$$\begin{aligned}&=\text{tr}\left(\boldsymbol{V}\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)^\top\cdot\left\{\sigma'(①)\odot\left[\left[\cos\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\odot d\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\right]\text{diag}(\boldsymbol{w})\sin\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)\right]\right\}\right)\\ &=\text{tr}\left(\left[\left[\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)\boldsymbol{V}^\top\right]\odot\sigma'(①)\right]^\top\cdot\left[\cos\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\odot d\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\right]\text{diag}(\boldsymbol{w})\sin\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)\right)\\ &=\text{tr}\left(\text{diag}(\boldsymbol{w})\sin\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)\left[\left[\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)\boldsymbol{V}^\top\right]\odot\sigma'(①)\right]^\top\left[\cos\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\odot d\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\right]\right)\\ &=\text{tr}(\left[\left[\text{diag}(\boldsymbol{w})\cdot\sin\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)\cdot\left[\left[\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)\boldsymbol{V}^\top\right]\odot\sigma'(①)\right]^\top\right]^\top\odot\cos\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\right]^\top\\ &\cdot d\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right))\end{aligned}$$

Remark

$$\boldsymbol{R}=\left[\left[\text{diag}(\boldsymbol{w})\cdot\sin\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)\cdot\left[\left[\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)\boldsymbol{V}^\top\right]\odot\sigma'(①)\right]^\top\right]^\top\odot\cos\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\right]^\top$$

$$\text{So part3}=\text{tr}\left(\tilde{\boldsymbol{B}}^\top\boldsymbol{R}(d\boldsymbol{X})^\top\right)=\text{tr}\left(d\boldsymbol{X}\left(\tilde{\boldsymbol{B}}^\top\boldsymbol{R}\right)^\top\right)=\text{tr}\left(\left(\tilde{\boldsymbol{B}}^\top\boldsymbol{R}\right)^\top d\boldsymbol{X}\right)$$

$$\begin{aligned}\text{part4}&=\text{tr}\left(\boldsymbol{V}\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)^\top\cdot\left\{\sigma'(①)\odot\left[\sin\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\cdot\text{diag}(\boldsymbol{w})\cdot[\cos\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)\odot d(\tilde{\boldsymbol{C}}\boldsymbol{X})]\right]\right\}\right)\\ &=\text{tr}\left(\left\{\left[\left[\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)\boldsymbol{V}^\top\right]\odot\sigma'(①)\right]^\top\left\{\sin\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\text{diag}(\boldsymbol{w})\cdot[\cos\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)\odot d(\tilde{\boldsymbol{C}}\boldsymbol{X})]\right\}\right)\\ &=\text{tr}(\left\{\left[\left\{\left[\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)\boldsymbol{V}^\top\right]\odot\sigma'(①)\right\}^\top\cdot\sin\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\cdot\text{diag}(\boldsymbol{w})\right]^\top\odot\cos\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)\right\}^\top\\ &\cdot\tilde{\boldsymbol{C}}d\boldsymbol{X})\end{aligned}$$

Remark

$$\boldsymbol{J}=\left\{\left[\left\{\left[\left(\frac{\partial y}{\partial\boldsymbol{G}}\right)\boldsymbol{V}^\top\right]\odot\sigma'(①)\right\}^\top\cdot\sin\left(\boldsymbol{X}^\top\tilde{\boldsymbol{B}}^\top\right)\cdot\text{diag}(\boldsymbol{w})\right]^\top\odot\cos\left(\tilde{\boldsymbol{C}}\boldsymbol{X}\right)\right\}^\top$$

$$\text{So part4}=\text{tr}\left(\left(\tilde{\boldsymbol{C}}^\top\boldsymbol{J}^\top\right)^\top d\boldsymbol{X}\right)$$

$$\left(\frac{\partial y}{\partial \boldsymbol{X}}\right)^{\top} = \left[\tilde{\boldsymbol{D}}^{\top}\left(\frac{\partial y}{\partial \boldsymbol{G}}\right)^{\top}\sigma(①)\right] + \left(-\tilde{\boldsymbol{B}}^{\top}F\right) + \tilde{\boldsymbol{C}}^{\top}N^{\top} + \tilde{\boldsymbol{B}}^{\top}\boldsymbol{R} + \tilde{\boldsymbol{C}}^{\top}\boldsymbol{J}^{\top}$$

$$\begin{aligned}
\boldsymbol{F} &= \left\{\left[\left[\left(\frac{\partial y}{\partial \boldsymbol{G}}\right)\boldsymbol{V}^{\top}\right]\odot\sigma^{'}(①)\cdot[\mathrm{diag}(\boldsymbol{w})\cdot\cos(\tilde{\boldsymbol{C}}\boldsymbol{X})]^{\top}\right]\odot\sin\left(\boldsymbol{X}^{\top}\tilde{\boldsymbol{B}}^{\top}\right)\right\}^{\top}\\
&= \left\{\left[\left(\frac{\partial y}{\partial \boldsymbol{G}}\right)\boldsymbol{V}^{\top}\right]\odot\sigma^{'}(①)\cdot[\mathrm{diag}(\boldsymbol{w})\cdot\cos(\tilde{\boldsymbol{C}}\boldsymbol{X})]^{\top}\right\}^{\top}\odot\sin(\tilde{\boldsymbol{B}}\boldsymbol{X})\\
&= \left\{\mathrm{diag}(\boldsymbol{w})\cdot\cos(\tilde{\boldsymbol{C}}\boldsymbol{X})\cdot\left[\left[\left(\frac{\partial y}{\partial \boldsymbol{G}}\right)\boldsymbol{V}^{\top}\right]\odot\sigma^{'}(①)\right]^{\top}\right\}\odot\sin(\boldsymbol{Q}^{\top})\\
&= \left\{\mathrm{diag}(\boldsymbol{w})\cdot\cos(\boldsymbol{K}^{\top})\cdot\left(\left[\boldsymbol{V}\left(\frac{\partial y}{\partial \boldsymbol{G}}\right)^{\top}\right]\odot\left[\sigma^{'}(①)\right]^{\top}\right)\right\}\odot\sin(\boldsymbol{Q}^{\top})\\
&= \{\mathrm{diag}(\boldsymbol{w})\mathrm{Re}\left[\exp\left(i\boldsymbol{K}^{\top}\right)\right)\\
&\quad\cdot\left[\left[\boldsymbol{V}\left(\frac{\partial y}{\partial \boldsymbol{G}}\right)^{\top}\right]\odot\sigma^{'}(\mathrm{Re}\left[\exp(-i\boldsymbol{K})\mathrm{diag}(\boldsymbol{w})\exp\left(i\boldsymbol{Q}^{\top}\right)\right])\right]\}\odot\mathrm{Im}\left[\exp\left(i\boldsymbol{Q}^{\top}\right)\right]\\
\boldsymbol{N}^{\top} &= \left\{\left[\left[\left(\frac{\partial y}{\partial \boldsymbol{G}}\right)\boldsymbol{V}^{\top}\right]\odot\sigma^{'}(①)\right]^{\top}\cdot\cos\left(\boldsymbol{X}^{\top}\tilde{\boldsymbol{B}}^{\top}\right)\cdot\mathrm{diag}(\boldsymbol{w})\right\}^{\top}\odot[-\sin(\tilde{\boldsymbol{C}}\boldsymbol{X})]\\
&= \left\{\left[\left[\boldsymbol{V}\left(\frac{\partial y}{\partial \boldsymbol{G}}\right)^{\top}\right]\odot\sigma^{'}(①^{\top})\right]\cdot\cos\left(\boldsymbol{X}^{\top}\tilde{\boldsymbol{B}}^{\top}\right)\cdot\mathrm{diag}(\boldsymbol{w})\right\}^{\top}\cdot[-\sin\left(\boldsymbol{K}^{\top}\right)]\\
&= \left\{\mathrm{diag}(\boldsymbol{w})\cdot\cos(\tilde{\boldsymbol{B}}\boldsymbol{X})\cdot\left[\left[\left(\frac{\partial y}{\partial \boldsymbol{G}}\right)\boldsymbol{V}^{\top}\right]\odot\sigma^{'}(①)\right]\right\}\odot[-\sin\left(\boldsymbol{K}^{\top}\right)]\\
&= \{\mathrm{diag}(\boldsymbol{w})\mathrm{Re}\left[\exp\left(i\boldsymbol{Q}^{\top}\right)\right]\\
&\quad\cdot\left[\left[\left(\frac{\partial y}{\partial \boldsymbol{G}}\right)^{\top}\boldsymbol{V}^{\top}\right]\odot\sigma^{'}(\mathrm{Re}\left[\exp(i\boldsymbol{Q})\mathrm{diag}(\boldsymbol{w})\exp(-i\boldsymbol{K})^{\top}\right])\right]\}\\
&\quad\odot\{-\mathrm{Im}\left[\exp\left(i\boldsymbol{K}^{\top}\right)\right]\}\\
\boldsymbol{R} &= \left\{\left[\mathrm{diag}(\boldsymbol{w})\sin(\tilde{\boldsymbol{C}}\boldsymbol{X})\cdot\left[\left[\left(\frac{\partial y}{\partial \boldsymbol{G}}\right)\boldsymbol{V}^{\top}\right]\odot\sigma^{'}(①)\right]^{\top}\right]^{\top}\odot\cos\left(\boldsymbol{X}^{\top}\tilde{\boldsymbol{B}}^{\top}\right)\right\}^{\top}\\
&= \left\{\left[\mathrm{diag}(\boldsymbol{w})\sin\left(\boldsymbol{K}^{\top}\right)\left[\left[\boldsymbol{V}\left(\frac{\partial y}{\partial \boldsymbol{G}}\right)^{\top}\right]\odot\sigma^{'}(①^{\top})\right]\right]^{\top}\odot\cos\left(\boldsymbol{X}^{\top}\tilde{\boldsymbol{B}}^{\top}\right)\right\}^{\top}\\
&= \left[\mathrm{diag}(\boldsymbol{w})\sin\left(\boldsymbol{K}^{\top}\right)\left[\left[\boldsymbol{V}\left(\frac{\partial y}{\partial \boldsymbol{G}}\right)^{\top}\right]\odot\sigma^{'}(①^{\top})\right]\right]\odot\cos(\tilde{\boldsymbol{B}}\boldsymbol{X})\\
&= \{\mathrm{diag}(\boldsymbol{w})\mathrm{Im}\left[\exp\left(i\boldsymbol{K}^{\top}\right)\right]\\
&\quad\cdot\left[\left[\boldsymbol{V}\left(\frac{\partial y}{\partial \boldsymbol{G}}\right)^{\top}\right]\odot\sigma^{'}(\mathrm{Re}\left[\exp(-i\boldsymbol{K})\mathrm{diag}(\boldsymbol{w})\exp\left(i\boldsymbol{Q}^{\top}\right)\right])\right]\}\odot\mathrm{Re}\left[\exp\left(i\boldsymbol{Q}^{\top}\right)\right]\\
\boldsymbol{J}^{\top} &= \left[\left[\left[\left(\frac{\partial y}{\partial \boldsymbol{G}}\right)\boldsymbol{V}^{\top}\right]\odot\sigma^{'}(①)\right]^{\top}\cdot\sin\left(\boldsymbol{X}^{\top}\tilde{\boldsymbol{B}}^{\top}\right)\cdot\mathrm{diag}(\boldsymbol{w})\right]^{\top}\odot\cos(\tilde{\boldsymbol{C}}\boldsymbol{X})\\
&= \left\{\mathrm{diag}(\boldsymbol{w})\sin(\tilde{\boldsymbol{B}}\boldsymbol{X})\left[\left[\left(\frac{\partial y}{\partial \boldsymbol{G}}\right)\boldsymbol{V}^{\top}\right]\odot\sigma^{'}(①)\right]\right\}\odot\cos\left(\boldsymbol{K}^{\top}\right)\\
&= \{\mathrm{diag}(\boldsymbol{w})\mathrm{Im}\left[\exp\left(i\boldsymbol{Q}^{\top}\right)\right]\\
&\quad\cdot\left[\left[\left(\frac{\partial y}{\partial \boldsymbol{G}}\right)\boldsymbol{V}^{\top}\right]\odot\sigma^{'}(\mathrm{Re}\left[\exp(i\boldsymbol{Q})\mathrm{diag}(\boldsymbol{w})\exp(-i\boldsymbol{K})^{\top}\right])\right]\odot\mathrm{Re}\left[\exp\left(i\boldsymbol{K}^{\top}\right)\right]\\
\frac{\partial y}{\partial \boldsymbol{X}} &= \tilde{\boldsymbol{D}}^{\top}\left(\frac{\partial y}{\partial \boldsymbol{G}}\right)^{\top}\sigma(\mathrm{Re}\left[\exp(i\boldsymbol{Q})\mathrm{diag}(\boldsymbol{w})\exp(-i\boldsymbol{K})^{\top}\right])\\
&\quad-\tilde{\boldsymbol{B}}^{\top}(\{\mathrm{diag}(\boldsymbol{w})\cdot\mathrm{Re}\left[\exp\left(i\boldsymbol{K}^{\top}\right)\right]\\
&\quad\cdot\left[\left[\boldsymbol{V}\left(\frac{\partial y}{\partial \boldsymbol{G}}\right)^{\top}\right]\odot\sigma^{'}(\mathrm{Re}\left[\exp(-i\boldsymbol{K})\mathrm{diag}(\boldsymbol{w})\exp\left(i\boldsymbol{Q}^{\top}\right)\right])\right]\}\odot\mathrm{Im}\left[\exp\left(i\boldsymbol{Q}^{\top}\right)\right])\\
&\quad-\tilde{\boldsymbol{C}}^{\top}(\{\ \mathrm{diag}(\boldsymbol{w})\cdot\mathrm{Re}\left[\exp\left(i\boldsymbol{Q}^{\top}\right)\right]\\
&\quad\cdot\left[\left[\left(\frac{\partial y}{\partial \boldsymbol{G}}\right)\boldsymbol{V}^{\top}\right]\odot\sigma^{'}(\mathrm{Re}\left[\exp(i\boldsymbol{Q})\mathrm{diag}(\boldsymbol{w})\exp(-i\boldsymbol{K})^{\top}\right])\right]\}\odot\mathrm{Im}\left[\exp\left(i\boldsymbol{K}^{\top}\right)\right])\\
&\quad+\tilde{\boldsymbol{B}}^{\top}(\{\mathrm{diag}(\boldsymbol{w})\cdot\mathrm{Im}\left[\exp\left(i\boldsymbol{K}^{\top}\right)\right]\\
&\quad\cdot\left[\left[\boldsymbol{V}\left(\frac{\partial y}{\partial \boldsymbol{G}}\right)^{\top}\right]\odot\sigma^{'}(\mathrm{Re}\left[\exp(-i\boldsymbol{K})\mathrm{diag}(\boldsymbol{w})\exp\left(i\boldsymbol{Q}^{\top}\right)\right])\right]\}\odot\mathrm{Re}\left[\exp\left(i\boldsymbol{Q}^{\top}\right)\right])\\
&\quad+\tilde{\boldsymbol{C}}^{\top}(\{\mathrm{diag}(\boldsymbol{w})\cdot\mathrm{Im}\left[\exp\left(i\boldsymbol{Q}^{\top}\right)\right]\\
&\quad\cdot\left[\left[\left(\frac{\partial y}{\partial \boldsymbol{G}}\right)\boldsymbol{V}^{\top}\right]\odot\sigma^{'}(\mathrm{Re}\left[\exp(i\boldsymbol{Q})\mathrm{diag}(\boldsymbol{w})\exp(-i\boldsymbol{K})^{\top}\right])\right]\}\odot\mathrm{Re}\left[\exp\left(i\boldsymbol{K}^{\top}\right)\right])
\end{aligned}$$

Correspondingly, we remark the equation as: $\frac{\partial y}{\partial \boldsymbol{X}} = \mathrm{PART}I - \mathrm{PART}II - \mathrm{PART}III + \mathrm{PART}IV + \mathrm{PART}V$.

For a matrix A, we define the infinite norms of matrices as: $\| \boldsymbol{A} \|_{\infty} = \max_i \sum_{j=1}^{n} |a_{ij}|$.

Since we can learn $\tilde{\boldsymbol{B}}, \tilde{\boldsymbol{C}}, \tilde{\boldsymbol{D}}, \mathrm{diag}(\boldsymbol{w}), \boldsymbol{X}$ and $\frac{\partial y}{\partial \boldsymbol{G}}$, we assume:

$\| \tilde{\boldsymbol{B}}^{\top} \|_{\infty} < \alpha, \| \tilde{\boldsymbol{C}}^{\top} \|_{\infty} < \beta, \| \mathrm{diag}(\boldsymbol{w}) \|_{\infty} < \zeta,$

$\| \tilde{\boldsymbol{D}}^{\top} \|_{\infty} < \gamma_1, \| \tilde{\boldsymbol{D}} \|_{\infty} < \gamma_2,$ let $\gamma = \max_i\{\gamma_1, \gamma_2\},$ then $\| \tilde{\boldsymbol{D}}^{\top} \|_{\infty} < \gamma, \| \tilde{\boldsymbol{D}} \|_{\infty} < \gamma,$

Similarily, $\| (\frac{\partial y}{\partial \boldsymbol{G}})^{\top} \|_{\infty} < \theta, \| (\frac{\partial y}{\partial \boldsymbol{G}}) \|_{\infty} < \theta.$

Note that each row of $\boldsymbol{X}$ can have at most one non-zero element due to the inherent sparsity of $\boldsymbol{X}$. We suppose the absolute value of every element in $\boldsymbol{X}$ is smaller than η, so we have $\| \boldsymbol{X} \|_{\infty} < \eta$ and $\| \boldsymbol{X}^{\top} \|_{\infty} < \mathrm{d}\eta$. According to the compatibility of this norm:

$$\begin{aligned}
\mathrm{PART}I &= \left\|\tilde{\boldsymbol{D}}^{\top}\left(\frac{\partial y}{\partial \boldsymbol{G}}\right)^{\top}\sigma\left(Re\left[\exp(i\boldsymbol{Q})\mathrm{diag}(\boldsymbol{w})\exp(-i\boldsymbol{K})^{T}\right]\right)\right\|_{\infty}\\
&\le \|\tilde{\boldsymbol{D}}^{\top}\|_{\infty}\cdot\|\left(\frac{\partial y}{\partial \boldsymbol{G}}\right)^{\top}\|_{\infty}\cdot\|\,\sigma\left(\mathrm{Re}\left[\exp(i\boldsymbol{Q})\mathrm{diag}(\boldsymbol{w})\exp(-i\boldsymbol{K})^{\top}\right]\right)\|_{\infty}\\
&\le \gamma\cdot\theta\cdot\mathrm{m}\\
\mathrm{PART}II &\le \|\boldsymbol{B}^{\top}\|_{\infty}\,\|\,\mathrm{diag}(\boldsymbol{w})\,\|_{\infty}\|\,\mathrm{Re}\left[\exp\left(i\boldsymbol{K}^{T}\right)\right]\|_{\infty}\,\|\boldsymbol{V}\|_{\infty}\left\|\left(\frac{\partial y}{\partial \boldsymbol{G}}\right)^{\top}\right\|_{\infty}\frac{1}{4}\\
&\le \alpha\cdot\zeta\cdot\mathrm{m}\cdot\mathrm{d}\cdot\eta\cdot\gamma\cdot\theta\cdot\frac{1}{4} = \frac{\alpha\gamma\eta\theta\zeta\mathrm{dm}}{4}
\end{aligned}$$

Similarly, $\mathrm{PART}III \le \frac{\beta\zeta\gamma\eta\theta\mathrm{m}}{4}, \mathrm{PART}IV \le \frac{\alpha\zeta\eta\gamma\theta\mathrm{dm}}{4}, \mathrm{PART}V \le \frac{\beta\zeta\theta\gamma\eta\mathrm{m}}{4}.$

$$\begin{aligned}
\left\|\frac{\partial y}{\partial \boldsymbol{X}}\right\|_{\infty} &\le \gamma\theta\mathrm{m} + \frac{\alpha\zeta\eta\gamma\theta\mathrm{dm}}{4} + \frac{\beta\zeta\gamma\eta\theta\mathrm{m}}{4} + \frac{\alpha\zeta\eta\gamma\theta\mathrm{dm}}{4} + \frac{\beta\zeta\gamma\eta\theta\mathrm{m}}{4}\\
&= \gamma\theta\mathrm{m} + \frac{\alpha\zeta\eta\gamma\theta\mathrm{dm}}{2} + \frac{\beta\zeta\gamma\eta\theta\mathrm{m}}{2}
\end{aligned}$$

Remark $\gamma\theta + \frac{\alpha\zeta\eta\gamma\theta\mathrm{d}}{2} + \frac{\beta\zeta\gamma\eta\theta}{2} = C_1,$ thus, $\left\|\frac{\partial y}{\partial \boldsymbol{X}}\right\|_{\infty} \le C_1\mathrm{m}.$

Note that C_1 is independent of the field number m. It can be seen that under certain regularity conditions, the gradient terms grow at most linearly with m.

For the traditional feature interaction algorithms, their gradients can be formulated as:

$$\begin{aligned}
&\boldsymbol{g} = \boldsymbol{X}_1^{\alpha_1}\odot\boldsymbol{X}_2^{\alpha_2}\odot\cdots\odot\boldsymbol{X}_m^{\alpha_m}\\
&\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{X}_i} = \boldsymbol{X}_1^{\alpha_1}\odot\boldsymbol{X}_2^{\alpha_2}\odot\cdots\odot\alpha_i\boldsymbol{X}_i^{\alpha_i-1}\odot\boldsymbol{X}_{i+1}^{\alpha_{i+1}}\odot\cdots\odot\boldsymbol{X}_m^{\alpha_m}\ ,\quad y = f(\boldsymbol{g}).
\end{aligned}$$

We suppose there exists j, for all i we suppose $|X_{ij}| \ge M - \epsilon$, here M is strictly bigger than zero, thus:

$$\begin{aligned}
&\left|\frac{\partial \boldsymbol{g}}{\partial X_{ij}}\right| \ge \alpha_i\cdot(M-\varepsilon)^{\sum_{i=1}^{m}\alpha_i-1}\\
&\left|\frac{\partial y}{\partial X_{ij}}\right| \ge \alpha_i\cdot(M-\varepsilon)^{\sum_{i=1}^{m}\alpha_i-1}\left\|\frac{\partial y}{\partial \boldsymbol{g}}\right\|_{\infty}
\end{aligned}$$

Let $t = \min_i\{\alpha_i\}$, thus we have:

$$\left|\frac{\partial \boldsymbol{g}}{\partial X_{ij}}\right| \ge \alpha_i\cdot(M-\varepsilon)^{mt-1}\left\|\frac{\partial y}{\partial \boldsymbol{g}}\right\|_{\infty}$$

We can clearly see that the gradient terms of traditional feature interaction algorithms exponentially grow with the field number m.