# Factorization Machine

Steffen Rendle

# Content

1. Instruction
   - Purpose(Advantage)
   - Backgrounds

2. FM
   - Core
   - Explanation

# Enormous sparse data

Feature vector **x**

| | User | | | | Movie | | | | | Other Movies rated | | | | | Time | Last Movie rated | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | B | C | ... | TI | NH | SW | ST | ... | TI | NH | SW | ST | ... | Time | TI | NH | SW | ST | ... |
| $\mathbf{x}^{(1)}$ | 1 | 0 | 0 | ... | 1 | 0 | 0 | 0 | ... | 0.3 | 0.3 | 0.3 | 0 | ... | 13 | 0 | 0 | 0 | 0 | ... |
| $\mathbf{x}^{(2)}$ | 1 | 0 | 0 | ... | 0 | 1 | 0 | 0 | ... | 0.3 | 0.3 | 0.3 | 0 | ... | 14 | 1 | 0 | 0 | 0 | ... |
| $\mathbf{x}^{(3)}$ | 1 | 0 | 0 | ... | 0 | 0 | 1 | 0 | ... | 0.3 | 0.3 | 0.3 | 0 | ... | 16 | 0 | 1 | 0 | 0 | ... |
| $\mathbf{x}^{(4)}$ | 0 | 1 | 0 | ... | 0 | 0 | 1 | 0 | ... | 0 | 0 | 0.5 | 0.5 | ... | 5 | 0 | 0 | 0 | 0 | ... |
| $\mathbf{x}^{(5)}$ | 0 | 1 | 0 | ... | 0 | 0 | 0 | 1 | ... | 0 | 0 | 0.5 | 0.5 | ... | 8 | 0 | 0 | 1 | 0 | ... |
| $\mathbf{x}^{(6)}$ | 0 | 0 | 1 | ... | 1 | 0 | 0 | 0 | ... | 0.5 | 0 | 0.5 | 0 | ... | 9 | 0 | 0 | 0 | 0 | ... |
| $\mathbf{x}^{(7)}$ | 0 | 0 | 1 | ... | 0 | 0 | 1 | 0 | ... | 0.5 | 0 | 0.5 | 0 | ... | 12 | 1 | 0 | 0 | 0 | ... |

# Purpose

**SVM and collaborative filter models:**

- Work not well in sparse data X
- Work only on very restricted data X
- Non-linear complexity X

**Factorization Machine:**

- Allow parameters estimation under very sparse data √
- Linear complexity √
- General predictor √

# Backgrounds

- Prediction function:

(W needs to be learned)

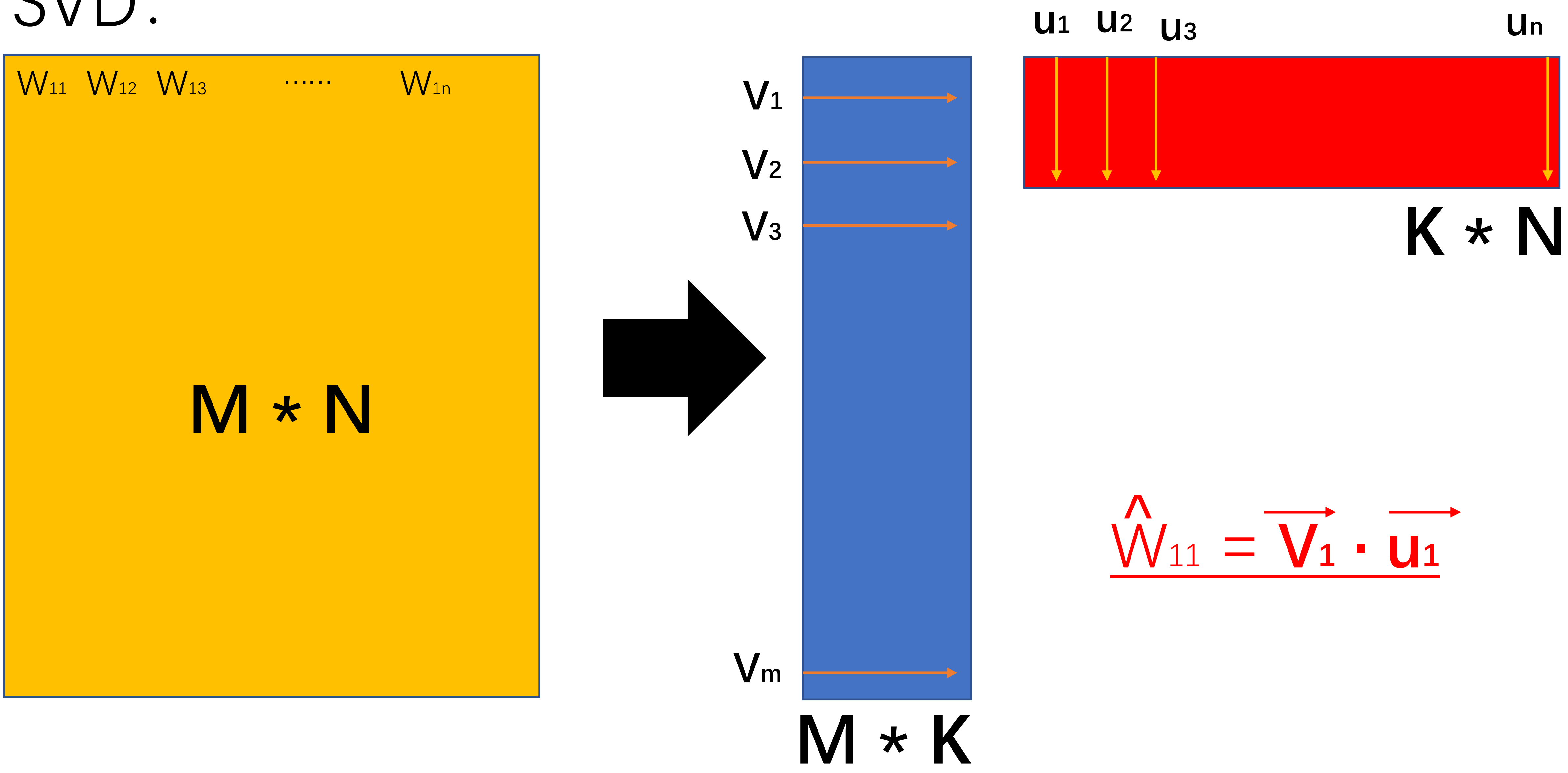
polynomial p=1: $$\hat{y}(\mathbf{x}) = w_0 + \sum_{i=1}^{n} w_i \, x_i, \quad w_0 \in \mathbb{R}, \quad \mathbf{w} \in \mathbb{R}^n$$

polynomial p=2: $$\hat{y(x)} = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_{i,j} x_i x_j$$

$W_{ij}$ is difficult to calculate if the matrix is sparse

# Backgrounds

SVD：




$$\hat{W}_{11} = \vec{V_1} \cdot \vec{u_1}$$

# FM

## Core:

$$\hat{y}(\mathbf{x}) := w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j$$

**X**: a single input, including n variable, **X** = $(x_1,x_2,x_3\cdots\cdots x_n)$

$w_0$ Global bias

$w_i$ Weight of single variable

$\hat{w}_{i,j} = \langle \mathbf{v}_i, \mathbf{v}_j \rangle$ Predicted weight of two variable

# Core

Matrix of $W_{ij}$(s)

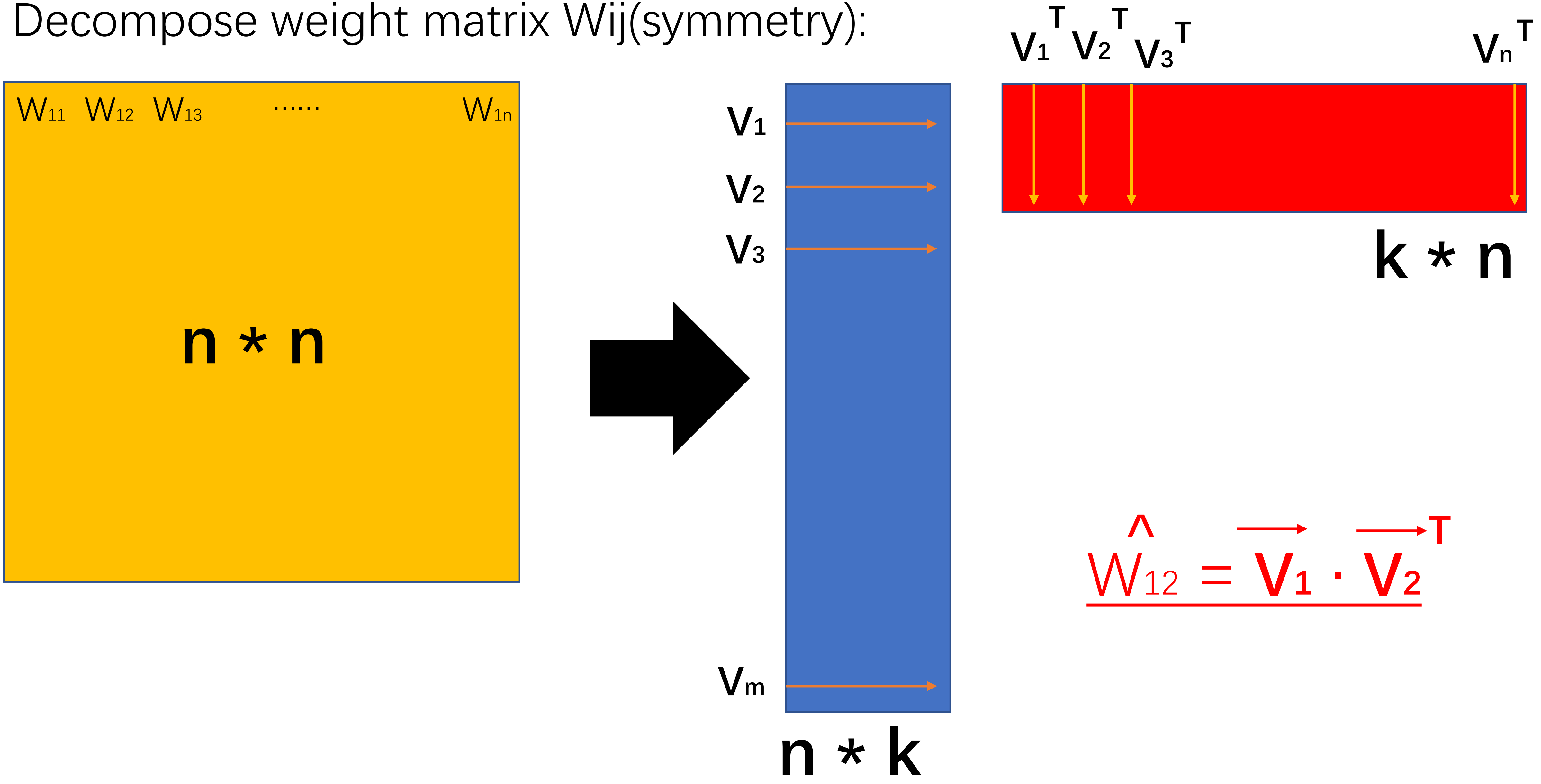
| | $X_1$ | $X_2$ | $X_3$ | ...... |
|---|---|---|---|---|
| $X_1$ | $W_{11}$ | $W_{12}$ | $W_{13}$ | ...... |
| $X_2$ | $W_{21}$ | $W_{22}$ | $W_{23}$ | ...... |
| $X_3$ | $W_{31}$ | $W_{32}$ | $W_{33}$ | ...... |
| ...... | ...... | ...... | ...... | ...... |

(Totally n records)

**n*n**

# Core

Decompose weight matrix Wij(symmetry):



# Core


n * k

$$\sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle \, x_i \, x_j$$

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle := \sum_{f=1}^{k} v_{i,f} \cdot v_{j,f}$$ (K: a chosen number)

| | | | | | |
|---|---|---|---|---|---|
| $V_1$ | 3 | 1 | 2 | 2 | 3 |
| $V_4$ | 5 | 2 | 4 | 2 | 1 |

$$<\mathbf{v}_1,\mathbf{v}_4> = 3*5 + 1*2 + 2*4 + 2*2 + 3*1$$
$$= 25$$

# Explanation

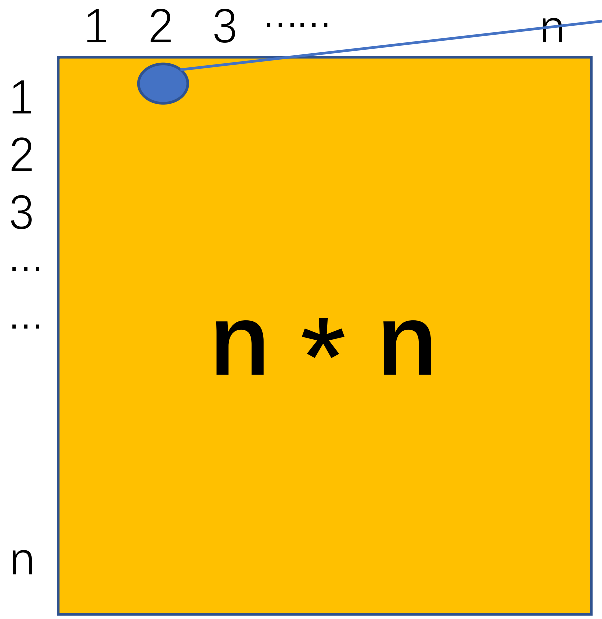
- Linear Complexity :

$$\hat{y}(\mathbf{x}) := w_0 + \sum_{i=1}^{n} w_i \, x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i , \mathbf{v}_j \rangle \, x_i \, x_j$$

linear

linear

?




<v1,v2>*x1*x2

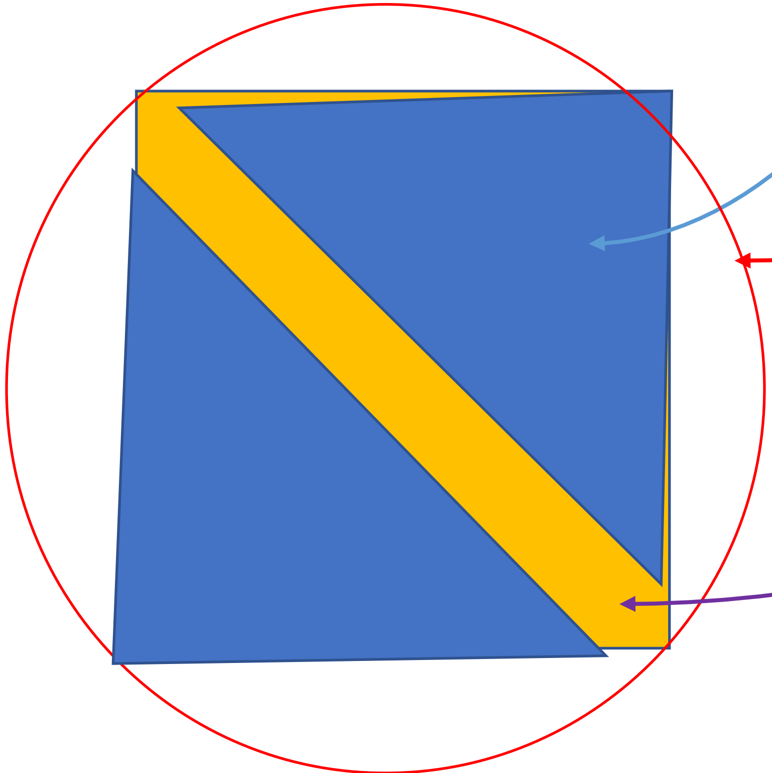
$$\sum_{i=1}^{n}\sum_{j=i+1}^{n}\langle \mathbf{v}_i,\mathbf{v}_j\rangle\, x_i\, x_j$$

$$=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\langle \mathbf{v}_i,\mathbf{v}_j\rangle\, x_i\, x_j-\frac{1}{2}\sum_{i=1}^{n}\langle \mathbf{v}_i,\mathbf{v}_i\rangle\, x_i\, x_i$$

$$=\frac{1}{2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{f=1}^{k} v_{i,f}\, v_{j,f}\, x_i\, x_j-\sum_{i=1}^{n}\sum_{f=1}^{k} v_{i,f}\, v_{i,f}\, x_i\, x_i\right)$$

$$=\frac{1}{2}\sum_{f=1}^{k}\left(\left(\sum_{i=1}^{n} v_{i,f}\, x_i\right)\left(\sum_{j=1}^{n} v_{j,f}\, x_j\right)-\sum_{i=1}^{n} v_{i,f}^2\, x_i^2\right)$$

$$=\frac{1}{2}\sum_{f=1}^{k}\left(\left(\sum_{i=1}^{n} v_{i,f}\, x_i\right)^2-\sum_{i=1}^{n} v_{i,f}^2\, x_i^2\right)$$



Linear !

# Explanation

## Parameters: gradient descent

$$\hat{y}(\mathbf{x}) := w_0 + \sum_{i=1}^{n} w_i \, x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle \, x_i \, x_j$$

$$\frac{\partial}{\partial \theta} \hat{y}(\mathbf{x}) = \begin{cases} 1, & \text{if } \theta \text{ is } w_0 \\ x_i, & \text{if } \theta \text{ is } w_i \\ x_i \sum_{j=1}^{n} v_{j,f} \, x_j - v_{i,f} \, x_i^2, & \text{if } \theta \text{ is } v_{i,f} \end{cases}$$