An equivalent formulation of matrix kernels (III)

Miaoyan Wang, Aug 7, 2020

1 Equivalence

- **Concatenated mapping**.

$$\begin{aligned}\Phi_{\text{con}} : \mathbb{R}^{d_1\times d_2} &\to \mathcal{H}_r^{d_1} \times \mathcal{H}_c^{d_2}\\ \boldsymbol{X} &\mapsto (\Phi_r(\boldsymbol{X}), \Phi_c(\boldsymbol{X})) \stackrel{\text{def}}{=} (\underbrace{\phi_r(\boldsymbol{X}_{1:}), \ldots, \phi_r(\boldsymbol{X}_{d_1:})}_{\text{row vectors, denoted } \boldsymbol{R}}, \underbrace{\phi_c(\boldsymbol{X}_{:1}), \ldots,\ \phi_c(\boldsymbol{X}_{:d_2})}_{\text{col vectors, denoted } \boldsymbol{L}})\end{aligned}$$

- **Bilinear mapping**.

$$\begin{aligned}\Phi_{\text{bi}} : \mathbb{R}^{d_1\times d_2} &\to (\mathcal{H}_r \times \mathcal{H}_c)^{d_1\times d_2}\\ \boldsymbol{X} &\mapsto [\Phi_{\text{bi}}(\boldsymbol{X})_{ij}], \quad \text{where } \Phi_{\text{bi}}(\boldsymbol{X})_{ij} \stackrel{\text{def}}{=} (\phi_c(\boldsymbol{X}_{i:}),\ \phi_r(\boldsymbol{X}_{:j})) = (\boldsymbol{R}_i, \boldsymbol{L}_j),\end{aligned}$$

  where $\boldsymbol{R}_i \in \mathcal{H}_r$ (respectively, $\boldsymbol{L}_j \in \mathcal{H}_c$) denotes the i-th (respectively, j-th) element in $\boldsymbol{R} \in \mathcal{H}_r^{d_1}$ (respectively, $\boldsymbol{L} \in \mathcal{H}_c^{d_2}$).

Using symbolic computation, it is easy to verify the following two properties.

1. There exists an one-to-one correspondence between the two mapped features.
2. There exists an one-to-one correspondence between the two low-rank coefficients. Specifically,

$$\begin{aligned}\text{concatenated function induced by } (\boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{P}_1, \boldsymbol{P}_2) &\cong \text{bilinear function induced by } (\boldsymbol{C}, \boldsymbol{P}_1, \boldsymbol{P}_2)\\ &\cong \text{bilinear function induced by } (\tilde{\boldsymbol{C}}, \boldsymbol{P}_1, \boldsymbol{P}_2),\end{aligned} \tag{1}$$

   where $\boldsymbol{C}, \tilde{\boldsymbol{C}}, \boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{P}_1, \boldsymbol{P}_2$ are parameters related to the low-rank coefficients (to be specified below).

Proof of Property 2. Note: I typically use subscripts "1" and "2" to distinguish quantities relevant to rows and columns. When there are clumped notations in sub/super-scripts, I omit the subscripts and instead use superscripts "row" and "col".

"$\Rightarrow$" The decision function under the concatenated mapping is

$$f_{\text{con}}(\boldsymbol{X}) = \langle \underbrace{(\boldsymbol{B}_1, \boldsymbol{B}_2)}_{\text{coefficients of interest}}, \underbrace{(\boldsymbol{R}, \boldsymbol{L})}_{\text{mapped feature } \Phi_1(\boldsymbol{X})} \rangle = \langle \boldsymbol{B}_1, \boldsymbol{R}\rangle + \langle \boldsymbol{B}_2, \boldsymbol{L}\rangle.$$

Suppose we impose low-rank structure $\boldsymbol{B}_k = \boldsymbol{C}_k \boldsymbol{P}_k^T$, where $\boldsymbol{C}_k \in \mathcal{H}^{r_k}$, and $\boldsymbol{P}_k \in \mathbb{R}^{d_k \times r}$ are matrices for $k = 1, 2$. In particular, $\boldsymbol{P}_k$ has full column rank but is not necessarily column-orthonormal. Denote $(\boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{P}_1, \boldsymbol{P}_2)$ the parameters for the decision function under the concatenated mapping. Then, we have

$$\begin{aligned} f_{\text{con}}(\boldsymbol{X}) &= \langle \boldsymbol{C}_1, \boldsymbol{R}\boldsymbol{P}_1 \rangle + \langle \boldsymbol{C}_2, \boldsymbol{L}\boldsymbol{P}_2 \rangle \\ &= \sum_{(i,s)\in[r]\times[d_1]} \boldsymbol{P}_{si}^{\text{row}} \underbrace{\langle \boldsymbol{c}_i^{\text{row}}, \boldsymbol{R}_s \rangle}_{\text{in mapped row space}} + \sum_{(i,s)\in[r]\times[d_2]} \boldsymbol{P}_{si}^{\text{col}} \underbrace{\langle \boldsymbol{c}_i^{\text{col}}, \boldsymbol{L}_s \rangle}_{\text{in mapped col space}}, \end{aligned} \tag{2}$$

where the subscripts, i, s, is, denote the i-th, s-th, and (i, s)-th element in the corresponding vector/matrices.

Now, consider the decision function under the bilinear mapping. We prove the equivalence (1) by construction. Define a triplet $(\boldsymbol{C}, \boldsymbol{P}_1, \boldsymbol{P}_2)$ based on $(\boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{P}_1, \boldsymbol{P}_2)$,

$$\boldsymbol{C} \leftarrow [\![(\gamma_1 \boldsymbol{c}_i^{\text{row}},\ \gamma_2 \boldsymbol{c}_j^{\text{col}})]\!] \in (\mathcal{H}_1 \times \mathcal{H}_2)^{r\times r}, \quad \boldsymbol{P}_k \leftarrow \boldsymbol{P}_k, \quad k = 1, 2, \tag{3}$$

where $\gamma_1 = \frac{1}{\sum_{i,s} \boldsymbol{P}_{si}^{\text{col}}}, \gamma_2 = \frac{1}{\sum_{i,s} \boldsymbol{P}_{si}^{\text{row}}}$ are two normalizing constants (assuming non-zero denominators), and $\{\boldsymbol{c}_i^{\text{row}}\}$, $\{\boldsymbol{c}_j^{\text{col}}\}$ are elements of $\boldsymbol{C}_1$, $\boldsymbol{C}_2$, respectively. Define a low-rank coefficient "matrix" $\boldsymbol{B} = \boldsymbol{P}_1 \boldsymbol{C} \boldsymbol{P}_2^T \in (\mathcal{H}_1 \times \mathcal{H}_2)^{d_1 \times d_2}$.

With this choice, the decision function under the bilinear mapping is

$$\begin{aligned} f_{\text{bi}}(\boldsymbol{X}) &= \langle \boldsymbol{B}, \Phi_{\text{bi}}(\boldsymbol{X}) \rangle = \langle \boldsymbol{P}_1 \boldsymbol{C} \boldsymbol{P}_2^T,\ \Phi_{\text{bi}}(\boldsymbol{X}) \rangle \\ &= \sum_{s,s',i,j} \boldsymbol{P}_{si}^{\text{row}} \boldsymbol{P}_{s'j}^{\text{col}} \langle (\boldsymbol{c}_i^{\text{row}}, \boldsymbol{c}_j^{\text{col}}),\ (\boldsymbol{R}_s, \boldsymbol{L}_{s'}) \rangle \\ &= \sum_{i,s} \boldsymbol{P}_{si}^{\text{row}} \langle \boldsymbol{c}_i^{\text{row}}, \boldsymbol{R}_s \rangle + \sum_{s,i} \boldsymbol{P}_{si}^{\text{col}} \langle \boldsymbol{c}_i^{\text{col}}, \boldsymbol{R}_s \rangle, \end{aligned} \tag{4}$$

where the last line follows from the definition of γ_1 and γ_2. Comparing (4) and (2), we have shown the correspondence from concatenated function to bilinear function.

"$\Leftarrow$" Suppose that we have a triplet of parameters, $(\boldsymbol{C}, \boldsymbol{P}_1, \boldsymbol{P}_2)$, for the decision function under the bilinear mapping. Let $(\boldsymbol{c}_{ij}^{\text{row}}, \boldsymbol{c}_{ij}^{\text{col}})$ denote the (i, j)-th entry of $\boldsymbol{C}$. Define a new matrix $\tilde{\boldsymbol{C}}$ whose (i, j)-th entry is $(\tilde{\boldsymbol{c}}_i^{\text{row}}, \tilde{\boldsymbol{c}}_j^{\text{col}})$,

$$\tilde{\boldsymbol{c}}_i^{\text{row}} = \frac{1}{r} \sum_j \boldsymbol{c}_{ij}^{\text{row}}, \quad \tilde{\boldsymbol{c}}_j^{\text{col}} = \frac{1}{r} \sum_i \boldsymbol{c}_{ij}^{\text{row}}, \quad \text{for all } (i, j) \in [r] \times [r].$$

The following computation shows that $(\tilde{\boldsymbol{C}}, \boldsymbol{P}_1, \boldsymbol{P}_2)$ induces the same function as $(\boldsymbol{C}, \boldsymbol{P}_1, \boldsymbol{P}_2)$.

$$f_{\text{bi}}(\boldsymbol{X}) = \langle \boldsymbol{C},\ \boldsymbol{Y} \rangle = \sum_{ij} \langle \boldsymbol{c}_{ij}^{\text{row}}, y_i^{\text{row}} \rangle + \sum_{ij} \langle \boldsymbol{c}_{ij}^{\text{col}}, y_j^{\text{col}} \rangle$$

$$
\begin{aligned}
&= r \sum_i \langle \tilde{\boldsymbol{c}}_i^{\text{row}}, y_i^{\text{row}} \rangle + r \sum_j \langle \tilde{\boldsymbol{c}}_j^{\text{col}}, y_j^{\text{col}} \rangle \\
&= \langle \tilde{\boldsymbol{C}}, \boldsymbol{Y} \rangle
\end{aligned}
$$

where, for notational convenience, we have denoted the matrix $\boldsymbol{Y} \stackrel{\text{def}}{=} \boldsymbol{P}_1^T \Phi_{\text{bi}}(\boldsymbol{X}) \boldsymbol{P}_2 = [\![(y_i^{\text{row}}, y_j^{\text{col}})]\!]$. Hence, the second correspondence in (1) is proved. The first correspondence in (1) is shown by a similar argument as in "$\Rightarrow$" in combination with the relationship (3). $\square$

2 Algorithm under bilinear mapping

Consider the bilinear mapping,

$$
\begin{aligned}
\Phi \colon \mathbb{R}^{d_1 \times d_2} &\to (\mathcal{H}_r \times \mathcal{H}_c)^{d_1 \times d_2} \\
\boldsymbol{X} &\mapsto [\Phi(\boldsymbol{X})_{ij}], \quad \text{where } \Phi(\boldsymbol{X})_{ij} \stackrel{\text{def}}{=} (\phi_c(\boldsymbol{X}_{i:}),\ \phi_r(\boldsymbol{X}_{:j})).
\end{aligned}
$$

We solve the optimization problem

$$
\begin{aligned}
&\min_{\boldsymbol{B}} \frac{1}{2} \|\boldsymbol{C}\|_F^2 + c \sum_{i=1}^n \xi_i, \\
&\text{subject to } y_i \langle \boldsymbol{P}_r \boldsymbol{C} \boldsymbol{P}_c^T, \Phi(\boldsymbol{X}_i) \rangle \leq 1 - \xi_i \text{ and } \xi_i \geq 0,\ i = 1, \ldots, n
\end{aligned}
\tag{5}
$$

where $\boldsymbol{P}_r \in \mathbb{R}^{d_1 \times r}$ and $\boldsymbol{P}_c \in \mathbb{R}^{d_2 \times r}$ are column-orthonormal matrices, and $\boldsymbol{C} = [\![(\boldsymbol{c}_i^{\text{row}},\ \boldsymbol{c}_j^{\text{col}})]\!] \in (\mathcal{H}_r \times \mathcal{H}_c)^{r \times r}$ are linear coefficients. Note that $\boldsymbol{C} \cong \mathcal{H}_r \times \mathcal{H}_c$ and $\|\boldsymbol{C}\|_F = \sum_{i=1}^r (\boldsymbol{c}_i^{\text{row}})^2 + \sum_{i=1}^r (\boldsymbol{c}_j^{\text{col}})^2$.

1. First, we update $\boldsymbol{C}\boldsymbol{P}_c^T$ holding $\boldsymbol{P}_r$ fixed. Under the orthonormal condition, the dual problem of (5) is

$$
\begin{aligned}
&\min_{\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)} - \sum_{i=1}^n \beta_i + \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{P}_r^T \Phi(\boldsymbol{X}_i),\ \boldsymbol{P}_r^T \Phi(\boldsymbol{X}_j) \rangle \\
&\text{subject to} \sum_i y_i \alpha_i = 0, \text{ and } 0 \leq \beta_i \leq c,\ i = 1, \ldots, n.
\end{aligned}
\tag{6}
$$

   Using kernel tricks, we solve (6) without the explicit feature mapping. The updating scheme of $\boldsymbol{C}\boldsymbol{P}_c^T$ is

$$
\boldsymbol{C}\boldsymbol{P}_c^T = \sum_i \alpha_i y_i \boldsymbol{P}_r^T \Phi(\boldsymbol{X}_i) \in (\mathcal{H}_r \times \mathcal{H}_c)^{r \times d_2}. \tag{7}
$$

2. Second, we update $\boldsymbol{P}_r$ holding $\boldsymbol{C}\boldsymbol{P}_c^T$ fixed. The dual problem of (5) is

$$\min_{\boldsymbol{\beta}=(\beta_1,\ldots,\beta_n)} -\sum_{i=1}^{n}\beta_i+\frac{1}{2}\sum_{i,j}\beta_i\beta_j y_i y_j\langle\Phi(\boldsymbol{X}_i)\boldsymbol{P}_c\boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{C}^T)^{-1/2},\ \Phi(\boldsymbol{X}_j)\boldsymbol{P}_c\boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{C}^T)^{-1/2}\rangle \tag{8}$$
$$\text{subject to} \sum_i y_i\beta_i=0,\ \text{and } 0\le\beta_i\le c,\ i=1,\ldots,n.$$

Using kernel tricks, we can solve (8) without explicit feature mapping. To show this, notice that by plugging (7) to (8), we have

$$\boldsymbol{C}\boldsymbol{C}^T=\boldsymbol{C}\boldsymbol{P}_c^T\boldsymbol{P}_c\boldsymbol{C}^T=\sum_{i,j}\alpha_i\alpha_j y_i y_j\boldsymbol{P}_r^T\Phi(\boldsymbol{X}_i)\Phi^T(\boldsymbol{X}_j)\boldsymbol{P}_r\in\mathbb{R}^{r\times r},$$
$$\Phi(\boldsymbol{X}_i)\boldsymbol{P}_c\boldsymbol{C}^T=\sum_j\Phi(\boldsymbol{X}_i)\alpha_j y_j\Phi^T(\boldsymbol{X}_j)\boldsymbol{P}_r=\sum_j\alpha_j y_j\Phi(\boldsymbol{X}_i)\Phi^T(\boldsymbol{X}_j)\boldsymbol{P}_r\in\mathbb{R}^{d_1\times r}.$$

Hence, the kernel in (8) can be expressed without explicit feature mapping.

We update $\boldsymbol{P}_r$ by

$$\boldsymbol{P}_r=\sum_i\beta_i y_i\underbrace{\Phi(\boldsymbol{X}_i)\boldsymbol{P}_c\boldsymbol{C}^T}_{\in\mathbb{R}^{d_1\times r}}\underbrace{(\boldsymbol{C}\boldsymbol{C}^T)^{-1}}_{\in\mathbb{R}^{r\times r}}\in\mathbb{R}^{d_1\times r}$$

The output $\boldsymbol{P}_r$ may not have orthonormal columns. We postprocess $\boldsymbol{P}_r$ by updating $\boldsymbol{P}_r\leftarrow$ Left singular space of $\boldsymbol{P}_r$, if $\boldsymbol{P}_r$ is not orthonormal.

How to read off $\boldsymbol{P}_r$ and $\boldsymbol{P}_c$ from the algorithm outputs?

(All the quantities below are outputs from the step 1.)

The row projection matrix $\boldsymbol{P}_r$ is readily available from the second step of the algorithm. To obtain the column projection matrix $\boldsymbol{P}_c$, we notice that the matrix $\boldsymbol{P}_c\boldsymbol{P}_c^T$ can be expressed without explicit feature mapping (see calculation below). Hence, $\boldsymbol{P}_c\leftarrow$ Singular space of $(\boldsymbol{P}_c\boldsymbol{P}_c^T)$.

Calculation of $\boldsymbol{P}_c\boldsymbol{P}_c^T$ without feature mapping:

$$\begin{aligned}\boldsymbol{P}_c\boldsymbol{P}_c^T&=\boldsymbol{P}_c\boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{P}_c^T\boldsymbol{P}_c\boldsymbol{C}^T)^{-1}\boldsymbol{C}\boldsymbol{P}_c^T\\&\overset{\text{c.f. (7)}}{=}\left(\sum_i\alpha_i y_i\Phi^T(\boldsymbol{X}_i)\right)\underbrace{\boldsymbol{P}_r\left(\sum_{i,j}\alpha_i\alpha_j y_i y_j\boldsymbol{P}_r^T\Phi(\boldsymbol{X}_i)\Phi^T(\boldsymbol{X}_j)\boldsymbol{P}_r\right)^{-1}\boldsymbol{P}_r^T}_{\text{computable without explicit feature mapping; denoted }\boldsymbol{W}}\left(\sum_i\alpha_i y_i\Phi(\boldsymbol{X}_i)\right)\\&=\sum_{i,j}\alpha_i\alpha_j y_i y_j\underbrace{\left[\Phi^T(\boldsymbol{X}_i)\boldsymbol{W}\Phi(\boldsymbol{X}_j)\right]}_{\text{a }d_2\text{-by-}d_2\text{ matrix over }\mathbb{R}}\end{aligned}$$

The matrix $\Phi^T(\boldsymbol{X})\boldsymbol{W}\Phi(\boldsymbol{Y})\in\mathbb{R}^{d_2\times d_2}$ can be expressed without explicit feature mapping. Specifically, the (i,j)-entry of $\Phi^T(\boldsymbol{X})\boldsymbol{W}\Phi(\boldsymbol{Y})$ is

$$[\Phi^T(\boldsymbol{X})\boldsymbol{W}\Phi(\boldsymbol{Y})]_{i,j} = \sum_{s,s'} w_{ss'}\langle \Phi(\boldsymbol{X})_{s,i},\ \Phi(\boldsymbol{Y})_{s',j}\rangle = \sum_{i,j} w_{ss'}K_r(s,s') + K_c(i,j)(\sum_{s,s'} w_{ss'}),$$

for all $(i,j) \in [d_2] \times [d_2]$.

How to read off the decision function from the algorithm outputs?

$$\begin{aligned}
f(\boldsymbol{X}_{\text{new}}) &= \text{trace}\left(\Phi^T(\boldsymbol{X}_{\text{new}})\boldsymbol{P}_r\boldsymbol{C}\boldsymbol{P}_c^T\right) \\
&\overset{\text{c.f. (7)}}{=} \text{trace}\left(\boldsymbol{P}_r\boldsymbol{P}_r^T\sum_i \alpha_i y_i \underbrace{\Phi(\boldsymbol{X}_i)\Phi^T(\boldsymbol{X}_{\text{new}})}_{\in\mathbb{R}^{d_1\times d_2}}\right) \\
&=: \sum_i \alpha_i y_i \left[K_r(i,\text{new}) + \tilde{K}_c(i,\text{new})\right]
\end{aligned}$$