Near-Orthogonality Regularization in Kernel Methods – Supplementary Material

1 ADMM-Based Algorithms

1.1 Detailed Derivation of Solving A

Given $\mathbf{H} \in \mathbb{R}^{K\times K}$ where $H_{ij} = \langle f_i, \hat{f}_j \rangle$, the sub-problem defined over $\mathbf{A}$ is

$$\min_{\mathbf{A}} \quad \lambda D_\phi(\mathbf{A}, \mathbf{I}) - \langle \mathbf{P}, \mathbf{A}\rangle - \langle \mathbf{Q}^\top, \mathbf{A}\rangle + \tfrac{\rho}{2}\|\mathbf{H} - \mathbf{A}\|_F^2 + \tfrac{\rho}{2}\|\mathbf{H}^\top - \mathbf{A}\|_F^2. \tag{1}$$

$D_\phi(\mathbf{A}, \mathbf{I})$ has three cases, which we discuss separately.
When $D_\phi(\mathbf{A}, \mathbf{I})$ is the squared Frobenius norm, the problem becomes

$$\min_{\mathbf{A}} \quad \lambda\|\mathbf{A} - \mathbf{I}\|_F^2 - \langle \mathbf{P}, \mathbf{A}\rangle - \langle \mathbf{Q}^\top, \mathbf{A}\rangle + \tfrac{\rho}{2}\|\mathbf{H} - \mathbf{A}\|_F^2 + \tfrac{\rho}{2}\|\mathbf{H}^\top - \mathbf{A}\|_F^2. \tag{2}$$

Taking the derivative and setting it to zero, we get the optimal solution for $\mathbf{A}$:

$$\mathbf{A} = (2\lambda\mathbf{I} + \mathbf{P} + \mathbf{Q}^\top + \rho(\mathbf{H} + \mathbf{H}^\top))/(2\lambda + 2\rho). \tag{3}$$

When $D_\phi(\mathbf{A}, \mathbf{I})$ is the log-determinant divergence, the problem is specialized to

$$\begin{array}{ll} \min_{\mathbf{A}} & \lambda(\mathrm{tr}(\mathbf{A}) - \log\det(\mathbf{A})) - \langle \mathbf{P}, \mathbf{A}\rangle - \langle \mathbf{Q}^\top, \mathbf{A}\rangle + \frac{\rho}{2}\|\mathbf{H} - \mathbf{A}\|_F^2 + \frac{\rho}{2}\|\mathbf{H}^\top - \mathbf{A}\|_F^2 \\ s.t. & \mathbf{A} \succ 0 \end{array} \tag{4}$$

Taking the derivative of the objective function w.r.t $\mathbf{A}$ and setting it to zero, we get

$$\mathbf{A}^2 + \tfrac{1}{\rho}(\lambda\mathbf{I} - \mathbf{P} - \mathbf{Q}^\top - \rho(\mathbf{H} + \mathbf{H}^\top))\mathbf{A} - \tfrac{\lambda}{\rho}\mathbf{I} = 0. \tag{5}$$

Let $\mathbf{B} = \frac{1}{\rho}(\lambda\mathbf{I} - \mathbf{P} - \mathbf{Q}^\top - \rho(\mathbf{H} + \mathbf{H}^\top))$, $\mathbf{C} = -\frac{\lambda}{\rho}\mathbf{I}$, Eq.(5) can be written as

$$\mathbf{A}^2 + \mathbf{B}\mathbf{A} + \mathbf{C} = 0. \tag{6}$$

According to (Higham and Kim, 2001), since $\mathbf{B}$ and $\mathbf{C}$ commute, the solution of this equation is

$$\mathbf{A} = -\tfrac{1}{2}\mathbf{B} + \tfrac{1}{2}\sqrt{\mathbf{B}^2 - 4\mathbf{C}}. \tag{7}$$

Taking an eigendecomposition of $\mathbf{B} = \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^{-1}$, we can compute $\mathbf{A}$ as $\mathbf{A} = \mathbf{\Phi}\widehat{\mathbf{\Sigma}}\mathbf{\Phi}^{-1}$, where $\widehat{\mathbf{\Sigma}}$ is a diagonal matrix with

$$\widehat{\Sigma}_{kk} = -\frac{1}{2}\Sigma_{kk} + \frac{1}{2}\sqrt{\Sigma_{kk}^2 + \frac{4\lambda}{\rho}}.$$

Since $\sqrt{\Sigma_{kk}^2 + \frac{4\lambda}{\rho}} > \Sigma_{kk}$, we know $\widehat{\Sigma}_{kk} > 0$. Hence $\mathbf{A}$ is positive definite.
When $D_\phi(\mathbf{A}, \mathbf{I})$ is the von Neumann divergence, the problem becomes

$$\begin{array}{ll} \min_{\mathbf{A}} & \lambda\mathrm{tr}(\mathbf{A}\log\mathbf{A} - \mathbf{A}) - \langle \mathbf{P}, \mathbf{A}\rangle - \langle \mathbf{Q}^\top, \mathbf{A}\rangle + \frac{\rho}{2}\|\mathbf{H} - \mathbf{A}\|_F^2 + \frac{\rho}{2}\|\mathbf{H}^\top - \mathbf{A}\|_F^2 \\ s.t. & \mathbf{A} \succ 0 \end{array} \tag{8}$$

Setting the gradient of the objective function w.r.t $\mathbf{A}$ to zero, we get

$$\lambda \log \mathbf{A} + 2\rho\mathbf{A} = \mathbf{P} + \mathbf{Q}^\top + \rho(\mathbf{H} + \mathbf{H}^\top). \tag{9}$$

Let $\mathbf{D} = \mathbf{P} + \mathbf{Q}^\top + \rho(\mathbf{H} + \mathbf{H}^\top)$. We perform an eigendecomposition of $\mathbf{D} = \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^{-1}$ and parameterize $\mathbf{A}$ as $\mathbf{A} = \mathbf{\Phi}\widehat{\mathbf{\Sigma}}\mathbf{\Phi}^{-1}$, then we obtain

$$\lambda \log \mathbf{A} + 2\rho\mathbf{A} = \mathbf{\Phi}(\lambda \log \widehat{\mathbf{\Sigma}} + 2\rho\widehat{\mathbf{\Sigma}})\mathbf{\Phi}^{-1}. \tag{10}$$

Plugging this equation into Eq.(9), we get the following equation:

$$\lambda \log \widehat{\mathbf{\Sigma}} + 2\rho\widehat{\mathbf{\Sigma}} = \mathbf{\Sigma} \tag{11}$$

which amounts to solving K independent one-variable equations taking the form

$$\lambda \log \widehat{\Sigma}_{ii} + 2\rho\widehat{\Sigma}_{ii} = \Sigma_{ii} \tag{12}$$

where $i = 1, \cdots, K$. This equation has a closed-form solution:

$$\widehat{\Sigma}_{ii} = \frac{\lambda\omega(\frac{\Sigma_{ii}}{\lambda} - \log(\frac{\lambda}{2\rho}))}{2\rho} \tag{13}$$

where $\omega(\cdot)$ is the Wright omega function Gorenflo et al. (2007). Due to the presence of $\log$, $\widehat{\Sigma}_{ii}$ is required to be positive and the solution always exists since the range of $\lambda \log \widehat{\Sigma}_{ii} + 2\rho\widehat{\Sigma}_{ii}$ and Σ_{ii} are both $(-\infty, \infty)$. Hence $\mathbf{A}$ is guaranteed to be positive definite.

1.2 ADMM-Based Algorithm for BMD-KSC

BMD-KSC has two set of parameters: sparse codes $\{\mathbf{a}_n\}_{n=1}^N$ and a dictionary of RKHS functions $\{f_i\}_{i=1}^K$. We use a coordinate descent algorithm to learn these two parameter sets, which iteratively performs the following two steps: (1) fixing $\{f_i\}_{i=1}^K$, solving $\{\mathbf{a}_n\}_{n=1}^N$; (2) fixing $\{\mathbf{a}_n\}_{n=1}^N$, solving $\{f_i\}_{i=1}^K$, until convergence. We first discuss step (1). The sub-problem defined over $\mathbf{a}_n$ is

$$\min_{\mathbf{a}_n} \quad \tfrac{1}{2}\|k(\mathbf{x}_n, \cdot) - \textstyle\sum_{i=1}^K a_{ni} f_i\|_{\mathcal{H}}^2 + \lambda_1 \|\mathbf{a}_n\|_1. \tag{14}$$

$\|k(\mathbf{x}_n, \cdot) - \sum_{i=1}^K a_{ni} f_i\|_{\mathcal{H}}^2 = k(\mathbf{x}_n, \mathbf{x}_n) - 2\mathbf{a}_n^\top \mathbf{h} + \mathbf{a}_n^\top \mathbf{G}\mathbf{a}_n$ where $\mathbf{h} \in \mathbb{R}^K$, $h_i = \langle f_i, k(\mathbf{x}_n, \cdot)\rangle$, $\mathbf{G} \in \mathbb{R}^{K\times K}$ and $G_{ij} = \langle f_i, f_j\rangle$. This is a standard Lasso (Tibshirani, 1996) problem and can be solved using many algorithms.
Next we discuss step (2), which learns $\{f_i\}_{i=1}^K$ using the ADMM-based algorithm outlined in the main paper. The updates of all variables are the same as those in BMD-KDML, except f_i. Let $b_{ni} = k(\mathbf{x}_n, \cdot) - \sum_{j\neq i}^K a_{nj} f_j$, the sub-problem defined over f_i is:

$$\begin{aligned}\min_{f_i} \quad & \tfrac{1}{2}\textstyle\sum_{n=1}^N \|b_{ni} - a_{ni} f_i\|_{\mathcal{H}}^2 + \frac{\lambda_2}{2}\|f_i\|_{\mathcal{H}}^2 + \langle g_i, f_i\rangle + \sum_{j=1}^K P_{ij}\langle f_i, \hat{f}_j\rangle + \sum_{j=1}^K Q_{ij}\langle f_i, \hat{f}_j\rangle \\ & + \tfrac{\rho}{2}\textstyle\sum_{j=1}^K (\langle f_i, \hat{f}_j\rangle - A_{ij})^2 + \frac{\rho}{2}\sum_{j=1}^K (\langle f_i, \hat{f}_j\rangle - A_{ji})^2\end{aligned} \tag{15}$$

This problem can be solved with functional gradient descent. The functional gradient of the objective function w.r.t f_i is

$$\sum_{n=1}^N a_{ni}(a_{ni} f_i - b_{ni}) + \lambda_2 f_i + g_i + \sum_{j=1}^K (P_{ij} + Q_{ij} + 2\rho\langle f_i, \hat{f}_j\rangle - \rho(A_{ij} + A_{ji}))\hat{f}_j \tag{16}$$

2 Proof of Lemma 1

For the ease of notation, we use $\langle\cdot,\cdot\rangle$ to denote the inner product in the RKHS and $\|\cdot\|$ to denote the Hilbert norm. To prove Lemma 1, we need the following Lemma, which is an extension of Lemma 4 in (Xie et al., 2015a) from vectors to RKHS functions.

Lemma 2. *Let $\mathcal{F} = \{f_i\}_{i=1}^K$ and $\mathbf{G}$ be the Gram matrix defined on $\mathcal{F}$. Let g_i' be the functional gradient of* $\det(\mathbf{G})$ *w.r.t f_i, then $\langle g_i', f_j\rangle = 0$ for all $j \neq i$, and $\langle g_i', f_i\rangle > 0$.*

Proof. We decompose f_i into $f_i = f_i^{\parallel} + f_i^{\perp}$. $f_i^{\parallel}$ is in the span of $\mathcal{F}/\{f_i\}$: $f_i^{\parallel} = \sum_{j\neq i}^{K} a_j f_j$, where $\{a_j\}_{j\neq i}^{K}$ are the linear coefficients. $f_i^{\perp}$ is orthogonal to $\mathcal{F}/\{f_i\}$: $\langle f_i^{\perp}, f_j\rangle = 0$ for all $j \neq i$.
Let $\mathbf{c}_j$ denote the j-th column of $\mathbf{G}$. Subtracting $\sum_{j\neq i}^{K} a_j \mathbf{c}_j$ from the i-th column, we get

$$\det(\mathbf{G}) = \begin{vmatrix} \langle f_1, f_1\rangle & \cdots & 0 & \cdots & \langle f_1, f_K\rangle \\ \langle f_2, f_1\rangle & \cdots & 0 & \cdots & \langle f_2, f_K\rangle \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \langle f_i, f_1\rangle & \cdots & \langle f_i^{\perp}, f_i\rangle & \cdots & \langle f_i, f_K\rangle \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \langle f_K, f_1\rangle & \cdots & 0 & \cdots & \langle f_K, f_K\rangle \end{vmatrix} \tag{17}$$

Expanding the determinant according to the i-th column, we get

$$\det(\mathbf{G}) = \det(\mathbf{G}_{-i})\langle f_i^{\perp}, f_i\rangle \tag{18}$$

where $\mathbf{G}_{-i}$ is the Gram matrix defined on $\mathcal{F}/\{f_i\}$. Then the functional gradient g_i' of $\det(\mathbf{G})$ w.r.t f_i is $\det(\mathbf{G}_{-i})f_i^{\perp}$, which is orthogonal to f_j for all $j \neq i$. Since $\mathbf{G}$ is full rank, we know $\det(\mathbf{G}) > 0$. From Eq.(18) we know $\langle f_i^{\perp}, f_i\rangle$ is non-negative. Hence $\langle g_i', f_i\rangle = \det(\mathbf{G}_{-i})\langle f_i^{\perp}, f_i\rangle > 0$. □

Now we are ready to prove Lemma 1. Some of the proof techniques draw inspiration from (Xie et al., 2015a). We first compute $\hat{s}_{ij}$:

$$\hat{s}_{ij} = \frac{|\langle \widehat{f}_i, \widehat{f}_j\rangle|}{\|\widehat{f}_i\|\|\widehat{f}_j\|} = \frac{|\langle f_i+\eta g_i, f_j+\eta g_j\rangle|}{\sqrt{\|f_i+\eta g_i\|^2}\sqrt{\|f_j+\eta g_j\|^2}} \tag{19}$$

The functional gradient of $\log\det(\mathbf{G})$ w.r.t f_i is computed as $g_i'' = \frac{1}{\det(\mathbf{G})} g_i'$. According to Lemma 1 and the fact that $\det(\mathbf{G}) > 0$, we know $\langle g_i'', f_j\rangle = 0$ for all $j \neq i$, and $\langle g_i'', f_i\rangle > 0$. Then we have

$$\begin{aligned} &|\langle f_i + \eta g_i, f_j + \eta g_j\rangle| \\ &= |\langle f_i + \eta(2f_i - 2g_i''), f_j + \eta(2f_j - 2g_j'')\rangle| \\ &= |\langle (1+2\eta)f_i - 2\eta g_i'', (1+2\eta)f_j - 2\eta g_j''\rangle| \\ &= |(1+2\eta)^2\langle f_i, f_j\rangle + 4\eta^2\langle g_i'', g_j''\rangle| \\ &= |\langle f_i, f_j\rangle||(1+2\eta)^2 + \tfrac{4\eta^2\langle g_i'', g_j''\rangle}{\langle f_i, f_j\rangle}| \end{aligned} \tag{20}$$

and

$$\begin{aligned} &\frac{1}{\sqrt{\|f_i+\eta g_i\|^2}} \\ &= \frac{1}{\sqrt{\|f_i\|^2 + 2\eta\langle f_i, g_i\rangle + \eta^2\|g_i\|^2}} \\ &= \frac{1}{\sqrt{\|f_i\|^2(1 + \frac{2\eta\langle f_i, g_i\rangle}{\|f_i\|^2} + \frac{\eta^2\|g_i\|^2}{\|f_i\|^2})}} \\ &= \frac{1}{\|f_i\|\sqrt{1 + \frac{2\eta\langle f_i, g_i\rangle}{\|f_i\|^2} + \frac{\eta^2\|g_i\|^2}{\|f_i\|^2}}} \end{aligned} \tag{21}$$

According to the Taylor expansion, we have

$$\frac{1}{\sqrt{1+x}} = 1 - \frac{1}{2}x + o(x) \tag{22}$$

Then

$$\frac{1}{\sqrt{1 + \frac{2\eta\langle f_i, g_i\rangle}{\|f_i\|^2} + \frac{\eta^2\|g_i\|^2}{\|f_i\|^2}}} = 1 - \frac{1}{2}\left(\frac{2\eta\langle f_i, g_i\rangle}{\|f_i\|^2} + \frac{\eta^2\|g_i\|^2}{\|f_i\|^2}\right) + o\left(\frac{2\eta\langle f_i, g_i\rangle}{\|f_i\|^2} + \frac{\eta^2\|g_i\|^2}{\|f_i\|^2}\right) \tag{23}$$

where

$$\frac{2\eta\langle f_i, g_i\rangle}{\|f_i\|^2} = \frac{2\eta\langle f_i, 2f_i - 2g_i''\rangle}{\|f_i\|^2} = 4\eta - 4\eta\frac{\langle f_i, g_i''\rangle}{\|f_i\|^2} \tag{24}$$

Hence

$$\frac{1}{\sqrt{1 + \frac{2\eta\langle f_i, g_i\rangle}{\|f_i\|^2} + \frac{\eta^2\|g_i\|^2}{\|f_i\|^2}}} = 1 - 2\eta + 2\eta\frac{\langle f_i, g_i''\rangle}{\|f_i\|^2} + o(\eta) \tag{25}$$

and

$$\frac{1}{\sqrt{1+\frac{2\eta\langle f_i,g_i\rangle}{\|f_i\|^2}+\frac{\eta^2\|g_i\|^2}{\|f_i\|^2}}}\frac{1}{\sqrt{1+\frac{2\eta\langle f_j,g_j\rangle}{\|f_j\|^2}+\frac{\eta^2\|g_j\|^2}{\|f_j\|^2}}}=(1-2\eta)^2+2\eta(\frac{\langle f_i,g_i''\rangle}{\|f_i\|^2}+\frac{\langle f_j,g_j''\rangle}{\|f_j\|^2})+o(\eta) \tag{26}$$

Then

$$\begin{aligned}\hat{s}_{ij} &= \frac{|\langle f_i,f_j\rangle|}{\|f_i\|\|f_j\|}|(1+2\eta)^2+\frac{4\eta^2\langle g_i'',g_j''\rangle}{\langle f_i,f_j\rangle}|((1-2\eta)^2+2\eta(\frac{\langle f_i,g_i''\rangle}{\|f_i\|^2}+\frac{\langle f_j,g_j''\rangle}{\|f_j\|^2})+o(\eta))\\ &\geq \frac{|\langle f_i,f_j\rangle|}{\|f_i\|\|f_j\|}((1+2\eta)^2+\frac{4\eta^2\langle g_i'',g_j''\rangle}{\langle f_i,f_j\rangle})((1-2\eta)^2+2\eta(\frac{\langle f_i,g_i''\rangle}{\|f_i\|^2}+\frac{\langle f_j,g_j''\rangle}{\|f_j\|^2})+o(\eta))\\ &= s_{ij}(1+2\eta(\frac{\langle f_i,g_i''\rangle}{\|f_i\|^2}+\frac{\langle f_j,g_j''\rangle}{\|f_j\|^2})+o(\eta))\\ &> s_{ij}\end{aligned} \tag{27}$$

This holds for all i,j, hence $s(\widehat{\mathcal{F}})>s(\mathcal{F})$. The proof completes.

3 Proof of Theorem 1

For the ease of presentation, we use n to denote the number of data examples. Note that this number is denoted by N in the main paper. Some of the proof techniques draw inspiration from (Xie et al., 2015b). A well established result in learning theory is that the generalization error can be upper bounded by the Rademacher complexity. We start from the Rademacher complexity, seek a further upper bound of it and show how $s(\mathcal{F})$ affects this upper bound. The Rademacher complexity $\mathcal{R}_n(\mathcal{A})$ of the loss function set $\mathcal{A}$ is defined as

$$\mathcal{R}_n(\mathcal{A}) = \mathbb{E}[\sup_{\ell\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^n\sigma_i\ell(u(\mathbf{x}_i,\mathbf{y}_i),t_i)] \tag{28}$$

where σ_i is uniform over $\{-1,1\}$ and $\{(\mathbf{x}_i,\mathbf{y}_i,t_i)\}_{i=1}^n$ are i.i.d samples drawn from p^*. Another form of Rademacher complexity Bartlett and Mendelson (2003) can be written as $\mathcal{R}_{||}(\mathcal{A})=\mathbb{E}[\sup_{\ell\in\mathcal{A}}|\frac{2}{n}\sum_{i=1}^n\sigma_i\ell(u(\mathbf{x}_i,\mathbf{y}_i),t_i)|]$. The Rademacher complexity can be utilized to upper bound the estimation error, as shown in Lemma 3.

Lemma 3. *(Anthony and Bartlett, 1999; Bartlett and Mendelson, 2003; Liang, 2015) With probability at least* $1-\delta$

$$L(u)-\widehat{L}(u)\leq 4\mathcal{R}_n(\mathcal{A})+\gamma\sqrt{\frac{2\log(2/\delta)}{n}} \tag{29}$$

for $\gamma\geq\sup_{\mathbf{x},\mathbf{y},t,u}|\ell(u(\mathbf{x},\mathbf{y}),t)|$

Our analysis starts from this lemma and we seek further upper bound of $\mathcal{R}_n(\mathcal{A})$. The analysis needs an upper bound of the Rademacher complexity of the hypothesis set $\mathcal{F}$, which is given in Lemma 4.

Lemma 4. *Let* $\mathcal{R}_n(\mathcal{F})$ *denote the Rademacher complexity of the hypothesis set* $\mathcal{F}$*, then*

$$\mathcal{R}_n(\mathcal{F})\leq\frac{8B(k)^2B'(k,C)^2K}{\sqrt{n}} \tag{30}$$

Proof. Let $\mathcal{V}=\{v:(\mathbf{x},\mathbf{y})\mapsto(f(\mathbf{x})-f(\mathbf{y}))^2, f\in\mathcal{H}\}$ denote the set of hypothesis $v(\mathbf{x},\mathbf{y})=(f(\mathbf{x})-f(\mathbf{y}))^2$, we have

$$\begin{aligned}\mathcal{R}_n(\mathcal{U}) &= \mathbb{E}[\sup_{u\in\mathcal{U}}\frac{1}{n}\sum_{i=1}^n\sigma_i\sum_{j=1}^K(f_j(\mathbf{x}_i)-f_j(\mathbf{y}_i))^2]\\ &= \mathbb{E}[\sup_{u\in\mathcal{U}}\frac{1}{n}\sum_{j=1}^K\sum_{i=1}^n\sigma_i(f_j(\mathbf{x}_i)-f_j(\mathbf{y}_i))^2]\\ &\leq K\mathbb{E}[\sup_{v\in\mathcal{V}}\frac{1}{n}\text{max}_j|\sum_{i=1}^n\sigma_i(f_j(\mathbf{x}_i)-f_j(\mathbf{y}_i))^2|]\\ &= K\mathbb{E}[\sup_{v\in\mathcal{V}}\frac{1}{n}|\sum_{i=1}^n\sigma_i(f(\mathbf{x}_i)-f(\mathbf{y}_i))^2|]\\ &= \frac{K}{2}\mathcal{R}_{||}(\mathcal{V})\end{aligned} \tag{31}$$

Let $\mathcal{G} = \{g : (\mathbf{x}, \mathbf{y}) \mapsto f(\mathbf{x}) - f(\mathbf{y}), f \in \mathcal{H}\}$ denote the set of hypothesis $g(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) - f(\mathbf{y})$ and $h(x) = x^2$, then $\mathcal{R}_{||}(\mathcal{V}) = \mathcal{R}_{||}(h \circ g)$. $h(0) = 0$ and h is Lipschitz continuous with Lipschitz constant L, which can be bounded as follows

$$\begin{aligned} L &= \sup_{g\in\mathcal{G}} |h'(g)| \\ &= \sup_{f,\mathbf{x},\mathbf{y}} 2|f(\mathbf{x}) - f(\mathbf{y})| \\ &\leq 4 \sup_{f,\mathbf{x}} |f(\mathbf{x})| \\ &\leq 4 \sup_{f,\mathbf{x}} |\langle f, k(\mathbf{x}, \cdot)\rangle| \\ &\leq 4 \sup_{f,\mathbf{x}} \|f\| \|k(\mathbf{x}, \cdot)\| \\ &\leq 4B(k)B'(k, C) \end{aligned} \tag{32}$$

According to the composition property of Rademacher complexity (Theorem 12 in Bartlett and Mendelson (2003)), we have

$$\mathcal{R}_{||}(h \circ g) \leq 4B(k)B'(k, C)\mathcal{R}_{||}(g) \tag{33}$$

Now we bound $\mathcal{R}_{||}(g)$:

$$\begin{aligned} \mathcal{R}_{||}(g) &= \mathbb{E}[\sup_{g\in\mathcal{G}} |\tfrac{2}{n} \textstyle\sum_{i=1}^n \sigma_i (f(\mathbf{x}_i) - f(\mathbf{y}_i))|] \\ &= \mathbb{E}[\sup_{g\in\mathcal{G}} |\tfrac{2}{n} \textstyle\sum_{i=1}^n \sigma_i (\langle f, k(\mathbf{x}_i, \cdot)\rangle - \langle f, k(\mathbf{y}_i, \cdot)\rangle)|] \\ &\leq \tfrac{2}{n}\mathbb{E}[\sup_{g\in\mathcal{G}} \|f\| \|\textstyle\sum_{i=1}^n \sigma_i (k(\mathbf{x}_i, \cdot) - k(\mathbf{y}_i, \cdot))\|] \\ &\leq \tfrac{2B(k)}{n}\mathbb{E}[\|\textstyle\sum_{i=1}^n \sigma_i (k(\mathbf{x}_i, \cdot) - k(\mathbf{y}_i, \cdot))\|] \\ &= \tfrac{2B(k)}{n}\mathbb{E}_{(\mathbf{x},\mathbf{y})}[\mathbb{E}_\sigma[\|\textstyle\sum_{i=1}^n \sigma_i (k(\mathbf{x}_i, \cdot) - k(\mathbf{y}_i, \cdot))\|]] \\ &\leq \tfrac{2B(k)}{n}\mathbb{E}_{(\mathbf{x},\mathbf{y})}[\sqrt{\mathbb{E}_\sigma[\|\textstyle\sum_{i=1}^n \sigma_i (k(\mathbf{x}_i, \cdot) - k(\mathbf{y}_i, \cdot))\|^2]}] \text{ (concavity of } \sqrt{\cdot}) \\ &= \tfrac{2B(k)}{n}\mathbb{E}_{(\mathbf{x},\mathbf{y})}[\sqrt{\mathbb{E}_\sigma[\textstyle\sum_{i=1}^n \sigma_i^2 \|k(\mathbf{x}_i, \cdot) - k(\mathbf{y}_i, \cdot)\|^2]}] \ (\forall i \neq j\ \sigma_i \perp\!\!\!\perp \sigma_j) \\ &= \tfrac{2B(k)}{n}\mathbb{E}_{(\mathbf{x},\mathbf{y})}[\sqrt{\textstyle\sum_{i=1}^n \|k(\mathbf{x}_i, \cdot) - k(\mathbf{y}_i, \cdot)\|^2}] \\ &= \tfrac{2B(k)}{n}\mathbb{E}_{(\mathbf{x},\mathbf{y})}[\sqrt{\textstyle\sum_{i=1}^n (k(\mathbf{x}_i, \mathbf{x}_i) + k(\mathbf{y}_i, \mathbf{y}_i) - 2k(\mathbf{x}_i, \mathbf{y}_i))}] \\ &\leq \tfrac{4B(k)B'(k,C)}{\sqrt{n}} \end{aligned} \tag{34}$$

Putting Eq.(33) and Eq.(34) together, we have

$$\mathcal{R}_{||}(\mathcal{V}) \leq \frac{16B(k)^2 B'(k, C)^2}{\sqrt{n}} \tag{35}$$

Plugging into $\mathcal{R}_n(\mathcal{U}) \leq \frac{K}{2}\mathcal{R}_{||}(\mathcal{V})$ completes the proof. □

In addition, we need the following bound of $u(\mathbf{x}, \mathbf{y})$.

Lemma 5.

$$\sup_{\mathbf{x},\mathbf{y},u} u(\mathbf{x}, \mathbf{y}) \leq \mathcal{J} \tag{36}$$

where $\mathcal{J} = 4B(k)^2 B'(k, C)^2((K - 1)s(\mathcal{F}) + 1)$.

Proof. Let $\mathbf{F} = [f_i, \cdots, f_K]$. We have

$$\begin{aligned} u(\mathbf{x}, \mathbf{y}) &= \textstyle\sum_{j=1}^K (f_j(\mathbf{x}) - f_j(\mathbf{y}))^2 \\ &= \textstyle\sum_{j=1}^K (\langle f_j, k(\mathbf{x}, \cdot)\rangle - \langle f_j, k(\mathbf{y}, \cdot)\rangle)^2 \\ &= \|\mathbf{F}^\top (k(\mathbf{x}, \cdot) - k(\mathbf{y}, \cdot))\|^2 \\ &\leq \|\mathbf{F}^\top\|_{op}^2 \|k(\mathbf{x}, \cdot) - k(\mathbf{y}, \cdot)\|^2 \\ &= \|\mathbf{F}\|_{op}^2 \|k(\mathbf{x}, \cdot) - k(\mathbf{y}, \cdot)\|^2 \\ &\leq (k(\mathbf{x}, \mathbf{x}) + k(\mathbf{y}, \mathbf{y}) - 2k(\mathbf{x}, \mathbf{y}))\|\mathbf{F}\|_{op}^2 \\ &\leq 4B'(k, C)^2 \|\mathbf{F}\|_{op}^2 \end{aligned} \tag{37}$$

where $\|\cdot\|_{op}$ denotes the operator norm.

$$\begin{aligned} \|\mathbf{F}\|_{op}^2 &= \sup_{\|\mathbf{a}\|_2=1} \|\mathbf{F}\mathbf{a}\|_2^2 \\ &= \sup_{\|\mathbf{a}\|_2=1} \mathbf{a}^\top \mathbf{F}^\top \mathbf{F}\mathbf{a} \\ &= \sup_{\|\mathbf{a}\|_2=1} \textstyle\sum_{i=1}^K \sum_{j=1}^K a_i a_j \langle f_i, f_j\rangle \\ &\leq \sup_{\|\mathbf{a}\|_2=1} \textstyle\sum_{i=1}^K \sum_{j=1}^K |a_i||a_j| \|f_i\| \|f_j\| |\cos\theta_{ij}| \\ &\leq \sup_{\|\mathbf{a}\|_2=1} \textstyle\sum_{i=1}^K \sum_{j=1}^K |a_i||a_j| B(k)^2 |\cos\theta_{ij}| \\ &\leq B(k)^2 \sup_{\|\mathbf{a}\|_2=1} (\textstyle\sum_{i=1}^K \sum_{j\neq i}^K |a_i||a_j| s(\mathcal{F}) + \sum_{i=1}^K a_i^2) \end{aligned} \tag{38}$$

where θ_{ij} is the angle between f_i and f_j. Define $\mathbf{a}' = [|a_1|, \cdots, |a_K|]^T$, $\mathbf{Q} \in \mathbb{R}^{K\times K}$: $Q_{ij} = s(\mathcal{F})$ for $i \neq j$ and $Q_{ii} = 1$, then $\|\mathbf{a}'\|_2 = \|\mathbf{a}\|$ and

$$\begin{array}{rcl} \|\mathbf{F}\|_{op}^2 & \leq & B(k)^2 \sup_{\|\mathbf{a}'\|_2=1} \mathbf{a}'^\top \mathbf{Q}\mathbf{a}' \\ & \leq & B(k)^2 \sup_{\|\mathbf{a}'\|_2=1} \lambda_1(\mathbf{Q})\|\mathbf{a}'\|_2^2 \\ & \leq & B(k)^2\lambda_1(\mathbf{Q}) \end{array} \tag{39}$$

where $\lambda_1(\mathbf{Q})$ is the largest eigenvalue of $\mathbf{Q}$. By simple linear algebra we can get $\lambda_1(\mathbf{Q}) = (K-1)s(\mathcal{F}) + 1$, so

$$\|\mathbf{F}\|_{op}^2 \leq B(k)^2((K-1)s(\mathcal{F}) + 1) \tag{40}$$

Substituting to Eq.(37), we have

$$u(\mathbf{x}, \mathbf{y}) \leq 4B(k)^2 B'(k, C)^2((K-1)s(\mathcal{F}) + 1) \tag{41}$$

Then $\sup_{\mathbf{x},\mathbf{y},u} |u(\mathbf{x}, \mathbf{y})| \leq \mathcal{J}$ with

$$\mathcal{J} = 4B(k)^2 B'(k, C)^2((K-1)s(\mathcal{F}) + 1) \tag{42}$$

The proof completes. □

Given these lemmas, we proceed to prove Theorem 1. Since $|\frac{\partial \ell(u(\mathbf{x},\mathbf{y}),t)}{\partial u(\mathbf{x},\mathbf{y})}| \leq \frac{1}{1+\exp(-u(\mathbf{x},\mathbf{y}))} \leq \frac{1}{1+\exp(-J)}$, $\ell(u(\mathbf{x}, \mathbf{y}), t)$ is Lipschitz continuous with respect to the first argument, and the constant L is $\frac{1}{1+\exp(-J)}$. Applying the composition property of Rademacher complexity, we have

$$\mathcal{R}_n(\mathcal{A}) \leq \frac{1}{1 + \exp(-J)}\mathcal{R}_n(\mathcal{U}) \tag{43}$$

Using Lemma 4, we have

$$\mathcal{R}_n(\mathcal{A}) \leq \frac{8B(k)^2 B'(k,C)^2 K}{(1+\exp(-J))\sqrt{n}} \tag{44}$$

In addition, $\sup_{\mathbf{x},\mathbf{y},t,u} |\ell(u(\mathbf{x}, \mathbf{y}), t)| \leq \log(1 + \exp(J))$ Substituting this inequality and Eq.(44) into Lemma 3 completes the proof.

4 Proof of Theorem 2

First, we derive an upper bound of the Babel function Tropp (2004):

$$\begin{array}{rcll} \mu_K(\{f_i\}_{i=1}^K) & = & \max\limits_{i\in\{1,\cdots,K\}} \max\limits_{\Lambda\subset\{1,\cdots,K\}\setminus\{i\};|\Lambda|=m} & \sum_{j\in\Lambda} |\langle f_j, f_i\rangle| \\ & \leq & \max\limits_{i\in\{1,\cdots,K\}} \max\limits_{\Lambda\subset\{1,\cdots,K\}\setminus\{i\};|\Lambda|=m} & \sum_{j\in\Lambda} \|f_i\|\|f_j\| s(\mathcal{F}) \\ & \leq & mB^2(k)s(\mathcal{F}) & \end{array} \tag{45}$$

In Theorem 14 of Vainsencher et al. (2011), we set the upper bound of $\mu_{K-1}(\{f_i\}_{i=1}^K)$ to $mB^2(k)s(\mathcal{F})$, then get Theorem 2.

5 Visualization

Given the learned RKHS functions $\{f_i\}_{i=1}^K$, we compute a $K \times K$ matrix $\mathbf{S}$ where S_{ij} is the absolute value of the cosine similarity between f_i and f_j. Then we visualize this matrix using a heatmap obtained by the `imagesc` function in MATLAB. Figure 1 shows the heatmaps of the matrices learned by different methods on the MIMIC-III dataset. The number of RKHS functions was fixed to 200. For all matrices, the diagonal entries equal to 1. From the visualization, we observed the following. For BMD-KDML methods KDML-(SFN,VND,LDD)-(RTR,RFF), the heatmaps have low energy on the off-diagonal entries, which indicates that these matrices are close to an identity matrix and hence the learned RKHS functions are close to being orthogonal. For the unregularized KDML, the heatmap has high energy on the off-diagonal entries. The matrices of KDML-(DPP,Angle) have lower energy compared with KDML and KDML-SHN, but their energy is higher than BMD-KDML methods.

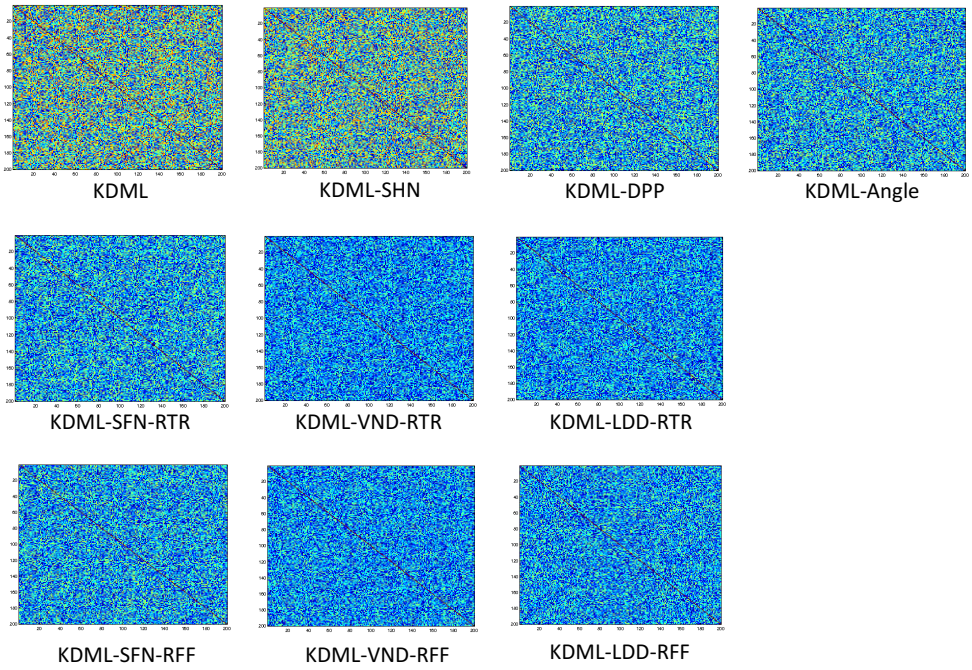

Figure 1: Heatmaps of the Pairwise Cosine Similarities

References

Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 1999.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2003.

Rudolf Gorenflo, Yuri Luchko, and Francesco Mainardi. Analytical properties and applications of the Wright function. *arXiv preprint math-ph/0701069*, 2007.

Nicholas J Higham and Hyun-Min Kim. Solving a quadratic matrix equation by newton's method with exact line searches. *SIAM Journal on Matrix Analysis and Applications*, 23(2):303–316, 2001.

Percy Liang. Lecture notes of statistical learning theory. 2015.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Joel A Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information theory*, 50(10):2231–2242, 2004.

Daniel Vainsencher, Shie Mannor, and Alfred M Bruckstein. The sample complexity of dictionary learning. *Journal of Machine Learning Research*, 12(Nov):3259–3281, 2011.

P. Xie, Y. Deng, and E. Xing. Diversifying restricted Boltzmann machine for document modeling. In *SIGKDD*, 2015a.

P. Xie, Y. Deng, and E. Xing. On the generalization error bounds of neural networks under mutual angular regularization. *arXiv:1511.07110*, 2015b.