1. The Gradients with Respect to \(u\)

The objective function, when applying the polynomial kernel, is

\[
\rho_{\text{poly}} = \frac{((X \cdot u + r_x)^{d_x})^\top \cdot k}{\| (X \cdot u + r_x)^{d_x} \| \cdot \| k \|} \tag{1}
\]

where \(k\) denotes the polynomial kernel on \(Yv\). The gradient of the objective function, with respect to \(u\):

\[
\frac{\partial \rho_{\text{poly}}}{\partial u} = \frac{d_x \cdot ((r_x + X \cdot u)^{d_x})^\top \cdot (r_x + X \cdot u)^{d_x} - \frac{1}{2}}{\| (r_x + X \cdot u)^{d_x} \| \cdot \| k \|}
\]

\[
\cdot (k^\top \cdot k)^{-\frac{1}{2}} \cdot (k \odot (r_x + X \cdot u)^{(d_x - 1)})^\top
\]

\[
\cdot X - ((d_x \cdot ((r_x + X \cdot u)^{d_x}))^\top
\]

\[
\cdot (r_x + X \cdot u)^{d_x} - \frac{1}{2} \cdot (k^\top \cdot k) - \frac{1}{2}) / 2
\]

\[
\cdot k^\top \cdot (r_x + X \cdot u)^{d_x} \cdot (r_x + X \cdot u)^{d_x}
\]

\[
\odot (r_x + X \cdot u)^{(d_x - 1)}^\top \cdot X.
\]

The gradient with respect to \(v\) is obtained similarly. The dominating computational cost arises from the matrix-vector product \(X \cdot u\) which lead to \(O((p + q)n)\) time-complexity per update where \(X \in \mathbb{R}^{n \times p}\).

The objective function, when applying the Gaussian (RBF) kernel, is

\[
\rho_{\text{RBF}} = \frac{\exp(-\frac{x}{\sigma^2}) + \frac{X \cdot u - \|u\|_2^2 \cdot 1}{\|u\|_2^2 \cdot 1} \cdot k}{\exp(-\frac{x}{\sigma^2}) + \frac{\|u\|_2^2 \cdot 1}{\|u\|_2^2 \cdot 1}} \cdot \| k \| \tag{2}
\]

where \(x\) is the vector of norms of examples of \(X\) and \(k\) denotes the Gaussian kernel on \(Yv\). The gradient of the

**Supplementary Material:**

Large-Scale Sparse Kernel Canonical Correlation Analysis

Viivi Uurtio 1,2 Sahely Bhadra 3 Juho Rousu 1,2

---

1 Equal contribution
2 Department of Computer Science, Aalto University, Espoo, Finland
3 Helsinki Institute for Information Technology, Helsinki, Finland
4 Computer Science and Engineering, Indian Institute of Technology Palakkad, Palakkad, India.

Correspondence to: Viivi Uurtio <viivi.uurtio@aalto.fi>.

The gradient with respect to $v$ is obtained similarly. The dominating computational cost arises from the matrix-vector product $X \cdot u$ which lead to $O((p + q)n)$ time-complexity per update where $X \in \mathbb{R}^{n \times p}$.

2. Proof of Theorems 4.3

In general cases, pre-image for $x$ is chosen such that the squared distance of $w_x$ and $\phi_x(u)$ is minimized.

$$\hat{u}(w_x) = \arg \min_{u \in \mathbb{R}^p} \|w_x - \phi_x(u)\|_2$$

$$\hat{v}(w_y) = \arg \min_{v \in \mathbb{R}^q} \|w_y - \phi_y(v)\|_2 \tag{3}$$

And hence the bounds on pre-image errors are then defined as

$$B_X = \max_{w_x \in L(\phi_x, X)} \|w_x - \phi_x(\hat{u}(w_x))\|_2$$

$$B_Y = \max_{w_y \in L(\phi_y, Y)} \|w_y - \phi_y(\hat{v}(w_y))\|_2 \tag{4}$$

**Theorem 2.1.** Let us assume that norm in $\mathcal{H}_x$ and $\mathcal{H}_y$ are upper bounded by $M_x$ and $M_y$, i.e.,

$$\forall \phi_x(u) \in \mathcal{H}_x, \|\phi_x(u)\| \leq M_x$$

and

$$\forall \phi_y(v) \in \mathcal{H}_y, \|\phi_y(v)\| \leq M_y \tag{5}$$

Then,

$$\rho_{\text{gradKCCA}} \geq \rho_{\text{preimage}} \geq \rho_{\text{kcca}} - \left(\frac{B_y}{M_y} + \frac{B_x}{M_x}\right)$$

**Proof.** Form bound of norm in Hilbert spaces:

$$\forall (u \text{ and } x_i),$$

$$-M_x^2 \leq \langle \phi_x(u), \phi_x(x_i) \rangle = k^x(x_i, u) \leq M_x^2$$

and

$$\forall (v \text{ and } y_i),$$

$$-M_y^2 \leq \langle \phi_y(v), \phi_y(y_i) \rangle k^y(y_i, v) \leq M_y^2 \tag{6}$$

As $\rho_{\text{kcca}} \geq 0$, the correlation achieved by pre-images is also positive. For positive $\rho_{\text{preimage}}$:

$$\rho_{\text{preimage}} = \frac{k^x(\hat{u}(w_x))^T k^y(\hat{v}(w_y))}{\|k^x(\hat{u}(w_x))\| \|k^y(\hat{v}(w_y))\|}$$

$$= \frac{k^x(\hat{u}(w_x))^T k^y(\hat{v}(w_y))}{\sqrt{\sum_i k^x(x_i, u(w_x))^2} \sqrt{\sum_i k^y(y_i, v(w_y))^2}}$$

$$\geq \frac{k^x(\hat{u}(w_x))^T k^y(\hat{v}(w_y))}{n M_x^2 M_y^2} \tag{7}$$

Note that, there always exist a pair of solution $\alpha^*$ and $\beta^*$ which gives optimal solution for KCCA and also satisfies $\|K^x \alpha^*\| = \sqrt{n}M_x^2$ and $\|K^y \beta^*\| = \sqrt{n}M_y^2$ (simple scaling of optimal solution). Given such solution $\alpha^*$ and $\beta^*$ for KCCA, the corresponding pre-image solution $\check{u}(w_x^*)$ and $\check{v}(w_y^*)$ is obtained by plugging the KCCA optimum $w_x^* = \sum_i \alpha^*_i \phi_x(x_i)$ and $w_y^* = \sum_i \beta^*_i \phi_y(y_i)$ into (3) (i.e., equation (8) and (9) in main manuscript).

The difference between the correlation found by KCCA and by its pre-image $\check{u}(w_x^*)$ and $\check{v}(w_y^*)$ is

$$\rho_{\text{kcca}} - \rho_{\text{preimage}} = \frac{\alpha^T K^x \beta^*}{\|K^x \alpha^*\| \|K^y \beta^*\|} - \frac{\check{k}^x(\check{u}(w_x^*))^T \check{k}^y(\check{v}(w_y^*))}{\|\check{k}^x(\check{u}(w_x^*))\| \|\check{k}^y(\check{v}(w_y^*))\|}$$

$$\leq \frac{\alpha^T K^x \beta^*}{n M_x^2 M_y^2} - \frac{\check{k}^x(\check{u}(w_x^*))^T \check{k}^y(\check{v}(w_y^*))}{n M_x^2 M_y^2} \tag{8}$$

[using (7) and the fact $\|K^x \alpha^*\| = \sqrt{n}M_x^2$ and $\|K^y \beta^*\| = \sqrt{n}M_y^2$]

$$= \left(\frac{K^x \alpha^*}{n M_x^2 M_y^2}\right)^T \left(\frac{K^y \beta^*}{n M_x^2 M_y^2} - \frac{\check{k}^y(\check{v}(w_y^*))}{n M_x^2 M_y^2}\right)$$

$$+ \left(\frac{K^x \alpha^*}{n M_x^2 M_y^2} - \frac{\check{k}^x(\check{u}(w_x^*))}{n M_x^2 M_y^2}\right)^T \frac{\check{k}^y(\check{v}(w_y^*))}{n M_x^2 M_y^2} \tag{9}$$

Where,

$$\left(\frac{K^x \alpha^*}{n M_x^2 M_y^2}\right)^T \left(\frac{K^y \beta^*}{n M_x^2 M_y^2} - \frac{\check{k}^y(\check{v}(w_y^*))}{n M_x^2 M_y^2}\right)$$

$$\leq \|K^x \alpha^*\| \|K^y \beta^* - \check{k}^y(\check{v}(w_y^*))\|$$

$$= \sqrt{n}M_x^2 \sqrt{\sum_i \langle \phi_y(y_i), (w_y - \phi_y(\check{v}(w_y^*)))^2}$$

$$\leq \sqrt{n}M_x^2 \sqrt{\sum_i \langle \phi_y(y_i) \rangle^2 \|w_y - \phi_y(\check{v}(w_y^*))\|^2}$$

$$\leq \sqrt{n}M_x^2 \sqrt{\sum_i M_y^2 B_y^2}$$

$$= \sqrt{n}M_x^2 M_y B_y$$

$$= n M_x^2 M_y B_y \tag{9}$$

Again similarly,

$$\left(\frac{K^x \alpha^* - \check{k}^x(\check{u}(w_x^*))}{n M_x^2 M_y^2}\right)^T \check{k}^y(\check{v}(w_y^*))$$

$$\leq n M_x M_y^2 B_x \tag{10}$$
Hence from (8) we get,

$$\rho_{kcca} - \rho_{preimage}$$

$$\leq \frac{nM_y^2M_x B_y}{nM_x^2M_y^2} + \frac{nM_x M_y^2 B_x}{nM_x^2M_y^2}$$

$$= \frac{B_y}{M_y} + \frac{B_x}{M_x}$$

(11)

Hence using Lemma 4.2

$$\rho_{gradKCCA} \geq \rho_{preimage} \geq \rho_{kcca} - \left( \frac{B_y}{M_y} + \frac{B_x}{M_x} \right)$$