Supplementary Material: Large-Scale Sparse Kernel Canonical Correlation Analysis

Viivi Uurtio [1 2] **Sahely Bhadra** [3] **Juho Rousu** [1 2]

1. The Gradients with Respect to $\mathbf{u}$

The objective function, when applying the polynomial kernel, is

$$\rho_{\text{poly}} = \frac{((\mathbf{X}\cdot\mathbf{u}+\mathbf{r}_x)^{d_x})^\top\cdot\mathbf{k}}{||(\mathbf{X}\cdot\mathbf{u}+\mathbf{r}_x)^{d_x})||\cdot||\mathbf{k}||} \tag{1}$$

where $\mathbf{k}$ denotes the polynomial kernel on $\mathbf{Yv}$. The gradient of the objective function, with respect to $\mathbf{u}$:

$$\begin{aligned}
\frac{\partial\rho_{\text{poly}}}{\partial\mathbf{u}} \quad = \quad & d_x\cdot(((\mathbf{r}_x+\mathbf{X}\cdot\mathbf{u})^{d_x})^\top\cdot(\mathbf{r}_x+\mathbf{X}\cdot\mathbf{u})^{d_x})^{-\frac{1}{2}} \\
& \cdot(\mathbf{k}^\top\cdot\mathbf{k})^{-\frac{1}{2}}\cdot(\mathbf{k}\odot(\mathbf{r}_x+\mathbf{X}\cdot\mathbf{u})^{(d_x-1)})^\top \\
& \cdot\mathbf{X}-((d_x\cdot(((\mathbf{r}_x+\mathbf{X}\cdot\mathbf{u})^{d_x})^\top \\
& \cdot(\mathbf{r}_x+\mathbf{X}\cdot\mathbf{u})^{d_x})^{-\frac{3}{2}}\cdot(\mathbf{k}^\top\cdot\mathbf{k})^{-\frac{1}{2}})/2 \\
& \cdot\mathbf{k}^\top\cdot(\mathbf{r}_x+\mathbf{X}\cdot\mathbf{u})^{d_x}\cdot((\mathbf{r}_x+\mathbf{X}\cdot\mathbf{u})^{d_x} \\
& \odot(\mathbf{r}_x+\mathbf{X}\cdot\mathbf{u})^{(d_x-1)})^\top\cdot\mathbf{X} \\
& +(d_x\cdot(((\mathbf{r}_x+\mathbf{X}\cdot\mathbf{u})^{d_x})^\top \\
& \cdot(\mathbf{r}_x+\mathbf{X}\cdot\mathbf{u})^{d_x})^{-\frac{3}{2}}\cdot(\mathbf{k}^\top\cdot\mathbf{k})^{-\frac{1}{2}})/2 \\
& \cdot((\mathbf{r}_x+\mathbf{X}\cdot\mathbf{u})^{d_x})^\top\cdot\mathbf{k}\cdot((\mathbf{r}_x+\mathbf{X}\cdot\mathbf{u})^{d_x} \\
& \odot(\mathbf{r}_x+\mathbf{X}\cdot\mathbf{u})^{(d_x-1)})^\top\cdot\mathbf{X}).
\end{aligned}$$

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

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

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

where $\mathbf{x}$ is the vector of norms of examples of $\mathbf{X}$ and $\mathbf{k}$ denotes the Gaussian kernel on $\mathbf{Yv}$. The gradient of the objective function, with respect to $\mathbf{u}$:

$$\begin{aligned}
\frac{\partial\rho_{\text{RBF}}}{\partial\mathbf{u}} \quad = \quad & ((\exp(1/(2\cdot\sigma^2)\cdot\mathbf{X}\cdot\mathbf{u}-1/(2\cdot\sigma^2) \\
& \cdot\mathbf{x}-\|\mathbf{u}\|_2^2/(2\cdot\sigma^2)\cdot\mathbf{1})^\top \\
& \cdot\exp(1/(2\cdot\sigma^2)\cdot\mathbf{X}\cdot\mathbf{u}-1/(2\cdot\sigma^2) \\
& \cdot\mathbf{x}-\|\mathbf{u}\|_2^2/(2\cdot\sigma^2)\cdot\mathbf{1}))^{(-(1+1/2))}\cdot \\
& (\mathbf{k}^\top\cdot\mathbf{k})^{(-1/2)})/(2\cdot\sigma^2)\cdot\text{sum}(\exp(1/(2\cdot\sigma^2) \\
& \cdot\mathbf{X}\cdot\mathbf{u}-1/(2\cdot\sigma^2)\cdot\mathbf{x}-\|\mathbf{u}\|_2^2/(2\cdot\sigma^2)\cdot\mathbf{1})\odot \\
& \exp(1/(2\cdot\sigma^2)\cdot\mathbf{X}\cdot\mathbf{u}-1/(2\cdot\sigma^2) \\
& \cdot\mathbf{x}-\|\mathbf{u}\|_2^2/(2\cdot\sigma^2)\cdot\mathbf{1}))\cdot\exp(1/(2\cdot\sigma^2)\cdot\mathbf{X}\cdot\mathbf{u} \\
& -1/(2\cdot\sigma^2)\cdot\mathbf{x}-\|\mathbf{u}\|_2^2/(2\cdot\sigma^2)\cdot\mathbf{1})^\top \\
& \cdot\mathbf{k}\cdot\mathbf{u}-((\exp(1/(2\cdot\sigma^2)\cdot\mathbf{X}\cdot\mathbf{u}-1/(2\cdot\sigma^2) \\
& \cdot\mathbf{x}-\|\mathbf{u}\|_2^2/(2\cdot\sigma^2)\cdot\mathbf{1})^\top\cdot\exp(1/(2\cdot\sigma^2)\cdot\mathbf{X}\cdot\mathbf{u} \\
& -1/(2\cdot\sigma^2)\cdot\mathbf{x}-\|\mathbf{u}\|_2^2/(2\cdot\sigma^2)\cdot\mathbf{1}))^{(-(1+1/2))} \\
& \cdot(\mathbf{k}^\top\cdot\mathbf{k})^{(-1/2)})/(4\cdot\sigma^2)\cdot\exp(1/(2\cdot\sigma^2)\cdot\mathbf{X}\cdot \\
& \mathbf{u}-1/(2\cdot\sigma^2)\cdot\mathbf{x}-\|\mathbf{u}\|_2^2/(2\cdot\sigma^2)\cdot\mathbf{1})^\top \\
& \cdot\mathbf{k}\cdot\mathbf{X}^\top\cdot(\exp(1/(2\cdot\sigma^2)\cdot\mathbf{X}\cdot u-1/(2\cdot\sigma^2)\cdot\mathbf{x}- \\
& \|\mathbf{u}\|_2^2/(2\cdot\sigma^2)\cdot\mathbf{1})\odot\exp(1/(2\cdot\sigma^2)\cdot\mathbf{X}\cdot \\
& \mathbf{u}-1/(2\cdot\sigma^2)\cdot\mathbf{x}-\|\mathbf{u}\|_2^2/(2\cdot\sigma^2)\cdot\mathbf{1})) \\
& -((\exp(1/(2\cdot\sigma^2)\cdot\mathbf{X}\cdot\mathbf{u}-1/(2\cdot\sigma^2) \\
& \cdot\mathbf{x}-\|\mathbf{u}\|_2^2/(2\cdot\sigma^2)\cdot\mathbf{1})^\top\cdot\exp(1/(2\cdot\sigma^2)\cdot\mathbf{X}\cdot\mathbf{u} \\
& -1/(2\cdot\sigma^2)\cdot\mathbf{x}-\|\mathbf{u}\|_2^2/(2\cdot\sigma^2)\cdot\mathbf{1}))^{(-(1+1/2))} \\
& \cdot(\mathbf{k}^\top\cdot\mathbf{k})^{(-1/2)})/(4\cdot\sigma^2)\cdot\mathbf{k}^\top\cdot\exp(1/(2\cdot\sigma^2) \\
& \cdot\mathbf{X}\cdot\mathbf{u}-1/(2\cdot\sigma^2)\cdot\mathbf{x}-\|\mathbf{u}\|_2^2/(2\cdot\sigma^2)\cdot\mathbf{1})\cdot\mathbf{X}^\top\cdot \\
& (\exp(1/(2\cdot\sigma^2)\cdot\mathbf{X}\cdot\mathbf{u}-1/(2\cdot\sigma^2)\cdot\mathbf{x}- \\
& \|\mathbf{u}\|_2^2/(2\cdot\sigma^2)\cdot\mathbf{1})\odot\exp(1/(2\cdot\sigma^2)\cdot\mathbf{X}\cdot\mathbf{u}-1/(2\cdot\sigma^2)\cdot\mathbf{x} \\
& -\|\mathbf{u}\|_2^2/(2\cdot\sigma^2)\cdot\mathbf{1}))+((\exp(1/(2\cdot\sigma^2)\cdot\mathbf{X}\cdot\mathbf{u}-1/(2\cdot\sigma^2)\cdot\mathbf{x} \\
& -\|\mathbf{u}\|_2^2/(2\cdot\sigma^2)\cdot\mathbf{1})^\top\cdot\exp(1/(2\cdot\sigma^2)\cdot\mathbf{X}\cdot\mathbf{u} \\
& -1/(2\cdot\sigma^2)\cdot\mathbf{x}-\|\mathbf{u}\|_2^2/(2\cdot\sigma^2)\cdot\mathbf{1}))^{(-(1+1/2))}\cdot \\
& (\mathbf{k}^\top\cdot\mathbf{k})^{(-1/2)})/(2\cdot\sigma^2)\cdot\text{sum}(\exp(1/(2\cdot\sigma^2)\cdot \\
& \mathbf{X}\cdot\mathbf{u}-1/(2\cdot\sigma^2)\cdot\mathbf{x}-\|\mathbf{u}\|_2^2/(2\cdot\sigma^2)\cdot\mathbf{1})\odot \\
& \exp(1/(2\cdot\sigma^2)\cdot\mathbf{X}\cdot\mathbf{u}-1/(2\cdot\sigma^2)\cdot\mathbf{x} \\
& -\|\mathbf{u}\|_2^2/(2\cdot\sigma^2)\cdot\mathbf{1}))\cdot\mathbf{k}^\top\cdot\exp(1/(2\cdot\sigma^2) \\
& \cdot\mathbf{X}\cdot\mathbf{u}-1/(2\cdot\sigma^2)\cdot\mathbf{x}-\|\mathbf{u}\|_2^2/(2\cdot\sigma^2)\cdot\mathbf{1})\cdot \\
& \mathbf{u}+((\exp(1/(2\cdot\sigma^2)\cdot\mathbf{X}\cdot\mathbf{u}-1/(2\cdot\sigma^2)\cdot\mathbf{x}- \\
& \|\mathbf{u}\|_2^2/(2\cdot\sigma^2)\cdot\mathbf{1})^\top\cdot\exp(1/(2\cdot\sigma^2)\cdot\mathbf{X}\cdot\mathbf{x}- \\
& 1/(2\cdot\sigma^2)\cdot\mathbf{u}-\|\mathbf{u}\|_2^2/(2\cdot\sigma^2)\cdot\mathbf{1}))^{(-1/2)}\cdot \\
& (\mathbf{k}^\top\cdot\mathbf{k})^{(-1/2)})/(2\cdot\sigma^2)\cdot\mathbf{X}^\top\cdot(\mathbf{k}\odot\exp(1/(2\cdot\sigma^2) \\
& \cdot\mathbf{X}\cdot\mathbf{u}-1/(2\cdot\sigma^2)\cdot\mathbf{x}-\|\mathbf{u}\|_2^2/(2\cdot\sigma^2)\cdot\mathbf{1})) \\
& -((\exp(1/(2\cdot\sigma^2)\cdot\mathbf{X}\cdot\mathbf{u}-1/(2\cdot\sigma^2)\cdot\mathbf{x} \\
& -\|\mathbf{u}\|_2^2/(2\cdot\sigma^2)\cdot\mathbf{1})^\top\cdot\exp(1/(2\cdot\sigma^2) \\
& \cdot\mathbf{X}\cdot\mathbf{u}-1/(2\cdot\sigma^2)\cdot\mathbf{x}-\|\mathbf{u}\|_2^2/(2\cdot\sigma^2)\cdot\mathbf{1}))^{(-1/2)} \\
& \cdot(\mathbf{k}^\top\cdot\mathbf{k})^{(-1/2)})/\sigma^2\cdot\text{sum}(\mathbf{k}\odot\exp(1/(2\cdot\sigma^2)\cdot\mathbf{X}\cdot \\
& \mathbf{u}-1/(2\cdot\sigma^2)\cdot\mathbf{x}-\|\mathbf{u}\|_2^2/(2\cdot\sigma^2)\cdot\mathbf{1}))\cdot\mathbf{u}.
\end{aligned}$$

*Equal contribution [1]Department of Computer Science, Aalto University, Espoo, Finland [2]Helsinki Institute for Information Technology, Helsinki, Finland [3]Computer Science and Engineering, Indian Institute of Technology Palakkad, Palakkad, India. Correspondence to: Viivi Uurtio <viivi.uurtio@aalto.fi>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

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

2. Proof of Theorems 4.3

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

$$\begin{aligned} \tilde{\mathbf{u}}(\mathbf{w}_x) &= \arg\min_{\mathbf{u}\in\mathbb{R}^p} \|\mathbf{w}_x - \phi_x(\mathbf{u})\|_2 \\ \tilde{\mathbf{v}}(\mathbf{w}_y) &= \arg\min_{\mathbf{v}\in\mathbb{R}^q} \|\mathbf{w}_y - \phi_y(\mathbf{v})\|_2 \end{aligned} \quad (3)$$

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

$$\begin{aligned} B_{\mathbf{X}} &= \max_{\mathbf{w}_x\in\mathcal{L}(\phi_x,\mathbf{X})} \|\mathbf{w}_x - \phi_x(\tilde{\mathbf{u}}(\mathbf{w}_x))\|_2 \\ B_{\mathbf{Y}} &= \max_{\mathbf{w}_y\in\mathcal{L}(\phi_y,\mathbf{Y})} \|\mathbf{w}_y - \phi_x(\tilde{\mathbf{v}}(\mathbf{w}_y))\|_2. \end{aligned} \quad (4)$$

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.,*

$$\begin{aligned} &\forall\, \phi_x(\mathbf{u}) \in \mathcal{H}_\S,\ \|\phi_x(\mathbf{u})\| \le M_x \\ \textit{and}\quad &\forall\, \phi_y(\mathbf{v}) \in \mathcal{H}_\dagger,\ \|\phi_y(\mathbf{v})\| \le M_y. \end{aligned} \quad (5)$$

Then,

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

Proof. Form bound of norm in Hilbert spaces:

$$\begin{aligned} &\forall(\mathbf{u} \text{ and } \mathbf{x}_i), \\ &\quad -M_x^2 \le \langle \phi_x(\mathbf{u}), \phi_x(\mathbf{x}_i)\rangle = k^x(\mathbf{x}_i,\mathbf{u}) \le M_x^2 \\ &\text{and } \forall\,(\mathbf{v} \text{ and } \mathbf{y}_i), \\ &\quad -M_y^2 \le \langle \phi_x(\mathbf{v}), \phi_y(\mathbf{y}_i)\rangle k^y(\mathbf{y}_i,\mathbf{v}) \le M_y^2. \end{aligned} \quad (6)$$

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

$$\begin{aligned} \rho_{\text{preimage}} &= \frac{\mathbf{k}^x(\tilde{\mathbf{u}}(\mathbf{w}_x))^T \mathbf{k}^y(\tilde{\mathbf{v}}(\mathbf{w}_y))}{||\mathbf{k}^x(\tilde{\mathbf{u}}(\mathbf{w}_x))||||\mathbf{k}^y(\tilde{\mathbf{v}}(\mathbf{w}_y)))||} \\ &= \frac{\mathbf{k}^x(\tilde{\mathbf{u}}(\mathbf{w}_x))^T \mathbf{k}^y(\tilde{\mathbf{v}}(\mathbf{w}_y))}{\sqrt{\sum_i k^x(\mathbf{x}_i,\tilde{\mathbf{u}}(\mathbf{w}_x))^2}\sqrt{\sum_i k^y(\mathbf{y}_i,\tilde{\mathbf{u}}(\mathbf{w}_y))^2}} \\ &\ge \frac{\mathbf{k}^x(\tilde{\mathbf{u}}(\mathbf{w}_x))^T \mathbf{k}^y(\tilde{\mathbf{v}}(\mathbf{w}_y))}{nM_x^2 M_y^2} \end{aligned} \quad (7)$$

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

The difference between the correlation found by KCCA and by its pre-image $\mathbf{u}(\tilde{\mathbf{w}_x^*})$ and $\mathbf{v}(\tilde{\mathbf{w}_y^*})$ is

$$\begin{aligned} &\rho_{\text{kcca}} - \rho_{\text{preimage}} \\ &= \frac{\boldsymbol{\alpha}^{*T}\mathbf{K}^x\mathbf{K}^y\boldsymbol{\beta}^*}{\|\mathbf{K}^x\boldsymbol{\alpha}^*\|\|\mathbf{K}^y\boldsymbol{\beta}^*\|} - \frac{\mathbf{k}^x(\tilde{\mathbf{u}}(\mathbf{w}_x^*))^T\mathbf{k}^y(\tilde{\mathbf{v}}(\mathbf{w}_y^*))}{\|\mathbf{k}^x(\tilde{\mathbf{u}}(\mathbf{w}_x^*))\|\|\mathbf{k}^y(\tilde{\mathbf{v}}(\mathbf{w}_y^*))\|} \\ &\le \frac{\boldsymbol{\alpha}^{*T}\mathbf{K}^x\mathbf{K}^y\boldsymbol{\beta}^*}{nM_x^2M_y^2} - \frac{\mathbf{k}^x(\tilde{\mathbf{u}}(\mathbf{w}_x^*))^T\mathbf{k}^y(\tilde{\mathbf{v}}(\mathbf{w}_y^*))}{nM_x^2M_y^2} \\ &\qquad [\text{using (7) and the fact } \|\mathbf{K}^x\boldsymbol{\alpha}^*\| = \sqrt{n}M_x^2 \text{ and } \|\mathbf{K}^y\boldsymbol{\beta}^*\| = \sqrt{n}M_y^2] \\ &= \frac{\boldsymbol{\alpha}^{*T}\mathbf{K}^x\mathbf{K}^y\boldsymbol{\beta}^* - \boldsymbol{\alpha}^{*T}\mathbf{K}^x\mathbf{k}^y(\tilde{\mathbf{v}}(\mathbf{w}_y^*))}{nM_x^2M_y^2} \\ &\quad + \frac{\boldsymbol{\alpha}^{*T}\mathbf{K}^x\mathbf{k}^y(\tilde{\mathbf{v}}(\mathbf{w}_y^*)) - \mathbf{k}^x(\tilde{\mathbf{u}}(\mathbf{w}_x^*))^T\mathbf{k}^y(\tilde{\mathbf{v}}(\mathbf{w}_y^*))}{nM_x^2M_y^2} \\ &\qquad [\text{by adding and subtracting same term}] \\ &= \left(\frac{\mathbf{K}^x\boldsymbol{\alpha}^*}{nM_x^2M_y^2}\right)^T\left(\mathbf{K}^y\boldsymbol{\beta}^* - \mathbf{k}^y(\tilde{\mathbf{v}}(\mathbf{w}_y^*))\right) \\ &\quad + \left(\mathbf{K}^x\boldsymbol{\alpha}^* - \mathbf{k}^x(\tilde{\mathbf{u}}(\mathbf{w}_x^*))\right)^T\frac{\mathbf{k}^y((\tilde{\mathbf{v}}(\mathbf{w}_y^*))}{nM_x^2M_y^2} \end{aligned} \quad (8)$$

Where,

$$\begin{aligned} &\left(\mathbf{K}^x\boldsymbol{\alpha}^*\right)^T\left(\mathbf{K}^y\boldsymbol{\beta}^* - \mathbf{k}^y(\tilde{\mathbf{v}}(\mathbf{w}_y^*))\right) \\ \le\ & \|\mathbf{K}^x\boldsymbol{\alpha}^*\|\|\mathbf{K}^y\boldsymbol{\beta}^* - \mathbf{k}^y(\tilde{\mathbf{v}}(\mathbf{w}_y^*))\| \\ =\ & \sqrt{n}M_x^2\sqrt{\sum_i \langle\phi_y(\mathbf{y}_i), (\mathbf{w}_y - \phi_y(\tilde{\mathbf{v}}(\mathbf{w}_y^*))\rangle^2} \\ \le\ & \sqrt{n}M_x^2\sqrt{\sum_i \|\phi_y(\mathbf{y}_i)\|^2\|(\mathbf{w}_y - \phi_y(\tilde{\mathbf{v}}(\mathbf{w}_y^*))\|^2} \\ \le\ & \sqrt{n}M_x^2\sqrt{\sum_i M_y^2B_y^2)} \\ =\ & \sqrt{n}M_x^2(\sqrt{n}M_yB_y) \\ =\ & nM_x^2M_yB_y \end{aligned} \quad (9)$$

Again similarly,

$$\begin{aligned} &\left(\mathbf{K}^x\boldsymbol{\alpha}^* - \mathbf{k}^x(\tilde{\mathbf{u}}(\mathbf{w}_x^*))\right)^T\mathbf{k}^y((\tilde{\mathbf{v}}(\mathbf{w}_y^*)) \\ \le\ & nM_xM_y^2B_x \end{aligned} \quad (10)$$

Hence from (8) we get,

$$\begin{aligned}
\rho_{\text{kcca}} & -\rho_{\text{preimage}} \\
& \leq \frac{nM_x^2 M_y B_y}{nM_x^2 M_y^2} + \frac{nM_x M_y^2 B_x}{nM_x^2 M_y^2} \\
& = \frac{B_y}{M_y} + \frac{B_x}{M_x} \qquad (11)
\end{aligned}$$

Hence using Lemma 4.2

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

□