On the Global Optima of Kernelized Adversarial Representation Learning (Supplementary Material)

Bashir Sadeghi
Michigan State University
sadeghib@msu.edu

Runyi Yu
Eastern Mediterranean University
yu@ieee.org

Vishnu Boddeti
Michigan State University
vishnu@msu.edu

In this supplementary material we include; (1) Section 1.1: Proof of Lemma 1, (2) Section 1.2: Proof of relation between constrained optimization problem in (8) and its Lagrangian formulation in (9), (3) Section 1.3: Proof of Theorem 2, (4) Section 1.4: Proof of Theorem 3, (5) Section 2: Empirical moments based solution to linear encoder, (6) Section 3: A detailed description of the Kernel-ARL extension, including derivation of its solution, (7) Section 3.2: Proof of Lemma 4, (8) Section 4: Additional analysis of experimental results, and (9) Section 5: Discussion on computational complexity of the Spectral-ARL solutions.

1. Proofs

We recall that for any square matrix $\mathbf{M}$, its trace, denoted by $\mathrm{Tr}[\mathbf{M}]$ is defined as the sum of all its diagonal elements. The Frobenius norm of $\mathbf{M}$ can be obtained as $\|\mathbf{M}\|_F^2 = \mathrm{Tr}(\mathbf{M}\mathbf{M}^T)$. This allows us to express the MSE of a centered random vector in terms of its covariance matrix:

$$\mathbb{E}\Big\{\big\|\mathbf{y}-\mathbf{b}_y\big\|^2\Big\} = \mathrm{Tr}\Big[\mathbb{E}\big\{(\mathbf{y}-\mathbf{b}_y)(\mathbf{y}-\mathbf{b}_y)^T\big\}\Big] = \mathrm{Tr}[\mathbf{C}_y].$$

Let $\mathbf{A}$ and $\mathbf{B}$ be two arbitrary matrices with the same dimension. Further, assume that the subspace $\mathcal{R}(\mathbf{A})$ is orthogonal to $\mathcal{R}(\mathbf{B})$. Then, using orthogonal decomposition (i.e., Pythagoras theorem), we have

$$\big\|\mathbf{A}+\mathbf{B}\big\|_F^2 = \big\|\mathbf{A}\big\|_F^2 + \big\|\mathbf{B}\big\|_F^2.$$

We provide the statements of the lemmas and theorems for sake of convenience, along with their proofs.

1.1. Proof of Lemma 1

Lemma 1. *Let $\mathbf{x}$ and $\boldsymbol{t}$ be two random vectors with $\mathbb{E}[\mathbf{x}] = 0$, $\mathbb{E}[\boldsymbol{t}] = \mathbf{b}$, and $\mathbf{C}_x \succ 0$. Consider a linear regressor, $\hat{\boldsymbol{t}} = \mathbf{W}\mathbf{z} + \mathbf{b}$, where $\mathbf{W} \in \mathbb{R}^{m\times r}$ is the parameter matrix, and $\mathbf{z} \in \mathbb{R}^r$ is an encoded version of $\mathbf{x}$ for a given $\boldsymbol{\Theta}_E$: $\mathbf{x} \mapsto \mathbf{z} = \boldsymbol{\Theta}_E\mathbf{x}$, $\boldsymbol{\Theta}_E \in \mathbb{R}^{r\times d}$. The minimum MSE that can be achieved by designing $\mathbf{W}$ is given as*

$$\min_{\mathbf{W}} \mathbb{E}[\|\boldsymbol{t}-\hat{\boldsymbol{t}}\|^2] = \mathrm{Tr}\big[\mathbf{C}_t\big] - \big\|P_{\mathcal{M}}\mathbf{Q}_x^{-T}\mathbf{C}_{xt}\big\|_F^2$$

where $\mathbf{M} = \mathbf{Q}_x\boldsymbol{\Theta}_E^T \in \mathbb{R}^{d\times r}$, and $\mathbf{Q}_x \in \mathbb{R}^{d\times d}$ is a Cholesky factor of $\mathbf{C}_x$ as shown in (1).

Proof. Direct calculations yield:

$$\begin{aligned}
J_t &= \mathbb{E}\Big\{\|\boldsymbol{t}-\hat{\boldsymbol{t}}\|^2\Big\} \\
&= \mathrm{Tr}\Big[\mathbb{E}\Big\{(\boldsymbol{t}-\mathbf{b}-\mathbf{W}\mathbf{z})(\boldsymbol{t}-\mathbf{b}-\mathbf{W}\mathbf{z})^T\Big\}\Big] \\
&= \mathrm{Tr}\Big[\mathbb{E}\Big\{(\boldsymbol{t}-\mathbf{b})(\boldsymbol{t}-\mathbf{b})^T+(\mathbf{W}\boldsymbol{\Theta}_E\mathbf{x})(\mathbf{W}\boldsymbol{\Theta}_E\mathbf{x})^T-(\boldsymbol{t}-\mathbf{b})(\mathbf{W}\boldsymbol{\Theta}_E\mathbf{x})^T-(\mathbf{W}\boldsymbol{\Theta}_E\mathbf{x})(\boldsymbol{t}-\mathbf{b})^T\Big\}\Big] \\
&= \mathrm{Tr}\Big[\mathbf{C}_t+(\mathbf{W}\boldsymbol{\Theta}_E)\mathbf{C}_x(\mathbf{W}\boldsymbol{\Theta}_E)^T-\mathbf{C}_{tx}(\mathbf{W}\boldsymbol{\Theta}_E)^T-(\mathbf{W}\boldsymbol{\Theta}_E)\mathbf{C}_{tx}^T\Big] \\
&= \mathrm{Tr}\Big[\mathbf{C}_t+(\mathbf{W}\boldsymbol{\Theta}_E\mathbf{Q}_x^T)(\mathbf{W}\boldsymbol{\Theta}_E\mathbf{Q}_x^T)^T-\mathbf{C}_{tx}(\mathbf{W}\boldsymbol{\Theta}_E)^T-(\mathbf{W}\boldsymbol{\Theta}_E)\mathbf{C}_{tx}^T\Big] \\
&= \mathrm{Tr}\Big[(\mathbf{W}\boldsymbol{\Theta}_E\mathbf{Q}_x^T-\mathbf{C}_{tx}\mathbf{Q}_x^{-1})(\mathbf{W}\boldsymbol{\Theta}_E\mathbf{Q}_x^T-\mathbf{C}_{tx}\mathbf{Q}_x^{-1})^T+\mathbf{C}_t-(\mathbf{C}_{tx}\mathbf{Q}_x^{-1})(\mathbf{C}_{tx}\mathbf{Q}_x^{-1})^T\Big] \\
&= \big\|\mathbf{Q}_x\boldsymbol{\Theta}_E^T\mathbf{W}^T-\mathbf{Q}_x^{-T}\mathbf{C}_{xy}\big\|_F^2-\big\|\mathbf{Q}_x^{-T}\mathbf{C}_{xt}\big\|_F^2+\mathrm{Tr}[\mathbf{C}_t]
\end{aligned}$$

Hence, the minimizer of J_t is obtained by minimizing the first term in the last equation, which is a standard least square error problem. Let $\mathbf{M}=\mathbf{Q}_x\boldsymbol{\Theta}_E^T$, then the minimizer is given by

$$\mathbf{W}^T=\mathbf{M}^\dagger\mathbf{Q}_x^{-T}\mathbf{C}_{xt}$$

Using the orthogonal decomposition

$$\big\|\mathbf{Q}_x^{-T}\mathbf{C}_{xt}\big\|_F^2=\big\|P_{\mathcal{M}}\mathbf{Q}_x^{-T}\mathbf{C}_{xt}\big\|_F^2+\big\|P_{\mathcal{M}^\perp}\mathbf{Q}_x^{-T}\mathbf{C}_{xt}\big\|_F^2$$

and

$$\begin{aligned}
\big\|\mathbf{Q}_x\boldsymbol{\Theta}_E^T\mathbf{W}^T-\mathbf{Q}_x^{-T}\mathbf{C}_{xt}\big\|_F^2 &= \big\|\mathbf{M}\mathbf{W}^T-P_{\mathcal{M}}\mathbf{Q}_x^{-T}\mathbf{C}_{xt}\big\|_F^2+\big\|P_{\mathcal{M}^\perp}\mathbf{Q}_x^{-T}\mathbf{C}_{xt}\big\|_F^2 \\
&= \big\|\underbrace{\mathbf{M}\mathbf{M}^\dagger}_{P_{\mathcal{M}}}\mathbf{Q}_x^{-T}\mathbf{C}_{xt}-P_{\mathcal{M}}\mathbf{Q}_x^{-T}\mathbf{C}_{xt}\big\|_F^2+\big\|P_{\mathcal{M}^\perp}\mathbf{Q}_x^{-T}\mathbf{C}_{xt}\big\|_F^2 \\
&= \big\|P_{\mathcal{M}^\perp}\mathbf{Q}_x^{-T}\mathbf{C}_{xt}\big\|_F^2,
\end{aligned}$$

we obtain the minimum value as

$$\mathrm{Tr}\big[\mathbf{C}_t\big]-\big\|P_{\mathcal{M}}\mathbf{Q}_x^{-T}\mathbf{C}_{xt}\big\|_F^2$$

□

1.2. Relation Between Constrained Optimization Problem in (8) and its Lagrangian Formulation in (9)

Consider the optimization problem in (8)

$$\mathbf{G}_\alpha=\arg\min_{\mathbf{G}} J_y(\mathbf{G}),\quad \text{s.t.}\quad J_s(\mathbf{G})\geq\alpha. \tag{A}$$

and the optimization problem in (9)

$$\mathbf{G}_\lambda=\arg\min_{G} J_\lambda(\mathbf{G}) \tag{B}$$

where

$$J_\lambda(\mathbf{G})=(1-\lambda)J_y(\mathbf{G})-\lambda J_s(\mathbf{G}),\quad \lambda\in[0,1]$$

Claim 1. For each $\lambda\in[0,1)$, solution $\mathbf{G}_\lambda$ of (B) is also a solution of (A) with

$$\alpha=J_s(\mathbf{G}_\lambda). \tag{C}$$

Proof. Let us consider (A) while assuming that (B) is satisfied. For each λ and $\mathbf{G}_\lambda$, let α be given as in (C). For an arbitrary $\mathbf{G}$ satisfying $J_s(\mathbf{G})\geq\alpha$, we have

$$\begin{aligned}
(1-\lambda)J_y(\mathbf{G}_\lambda)-\lambda\alpha &= (1-\lambda)J_y(\mathbf{G}_\lambda)-\lambda J_s(\mathbf{G}_\lambda) \\
&\leq (1-\lambda)J_y(\mathbf{G})-\lambda J_s(\mathbf{G}),
\end{aligned}$$

where the second step is from the assumption that B is satisfied. Consequently, we have,

$$(1-\lambda)\big[J_y(\mathbf{G})-J_y(\mathbf{G}_\lambda)\big]\geq\lambda\big[J_s(\mathbf{G})-\alpha\big]\geq 0.$$

Since $J_s(\mathbf{G})\geq\alpha$, this implies that $J_y(\mathbf{G})\geq J_y(\mathbf{G}_\lambda)$ and consequently $\mathbf{G}_\lambda$ is a possible minimizer of problem (A). □

1.3. Proof of Theorem 2

Theorem 2. *As a function of $\mathbf{G}_E \in \mathbb{R}^{d\times r}$, the objective function in equation (9) is neither convex nor differentiable.*

Proof. Recall that $P_{\mathcal{G}}$ is equal to $\mathbf{G}_E(\mathbf{G}_E^T\mathbf{G}_E)^\dagger\mathbf{G}_E^T$. Therefore, due to the involvement of the pseudo inverse, (9) is not differentiable (see [2]).

For non-convexity consider the theorem that $f(\mathbf{G}_E)$ is convex in $\mathbf{G}_E \in \mathbb{R}^{d\times r}$ if and only if $h(t) = f(t\,\mathbf{G}_1 + \mathbf{G}_2)$ is convex in $t \in \mathbb{R}$ for any constants $\mathbf{G}_1,\ \mathbf{G}_2 \in \mathbb{R}^{d\times r}$ (see [1]).

In order to use the above theorem, consider rank one matrices

$$\mathbf{G}_1 = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \dots & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{G}_2 = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \dots & 0 \end{bmatrix}.$$

Define $\mathbf{G}_E = (t\,\mathbf{G}_1 + \mathbf{G}_2)$. Then

$$P_{\mathcal{G}}(t) = \mathbf{G}_E(\mathbf{G}_E^T\mathbf{G}_E)^\dagger\mathbf{G}_E^T = \frac{1}{(t+1)^2+1}\begin{bmatrix} (t+1)^2 & (t+1) & 0 & \dots & 0 \\ (t+1) & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}.$$

Using basic properties of trace we get,

$$(1-\lambda)J_y(\mathbf{G}_E) - \lambda J_s(\mathbf{G}_E) = \mathrm{Tr}\big[P_{\mathcal{G}}(t)\mathbf{B}\big],$$

where the matrix $\mathbf{B}$ is given in (14) and we used Lemma 1. Now, represent $\mathbf{B}$ as

$$\mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1d} \\ b_{12} & b_{22} & \dots & b_{2d} \\ \vdots & \vdots & \ddots & \\ b_{1d} & b_{2d} & \dots & b_{dd} \end{bmatrix}.$$

Thus,

$$\mathrm{Tr}\big[P_{\mathcal{G}}(t)\mathbf{B}\big] = b_{11} + \frac{2b_{12}(t+1) + b_{22} - b_{11}}{(t+1)^2+1}.$$

It can be shown that the above function of t is convex only if $b_{12} = 0$ and $b_{11} = b_{22}$. On the other hand, if these two conditions hold, it can be similarly shown that $(1-\lambda)J_y(\mathbf{G}_E) - \lambda J_s(\mathbf{G}_E)$ is non-convex by considering a different pair of matrices $\mathbf{G}_1$ and $\mathbf{G}_2$. This implies that $(1-\lambda)J_y(\mathbf{G}_E) - \lambda J_s(\mathbf{G}_E)$ is not convex. □

1.4. Proof of Theorem 3

Theorem 3. *Assume that the number of negative eigenvalues (β) of $\mathbf{B}$ in (13) is j. Denote $\gamma = \min\{r, j\}$. Then, the minimum value in (10) is given as,*

$$\beta_1 + \beta_2 + \cdots + \beta_\gamma \tag{D}$$

where $\beta_1 \le \beta_2 \le \ldots \le \beta_\gamma < 0$ are the γ least eigenvalues of $\mathbf{B}$. And the minimum can be attained by $\mathbf{G}_E = \mathbf{V}$, where the columns of $\mathbf{V}$ are eigenvectors corresponding to all the γ negative eigenvalues of $\mathbf{B}$.

Proof. Consider the inner optimization problem of (10) in (11). Using the trace optimization problems and their solutions in [3], we get

$$\min_{\mathbf{G}_E^T\mathbf{G}_E=\mathbf{I}_i} J_\lambda(\mathbf{G}_E) \;=\; \min_{\mathbf{G}_E^T\mathbf{G}_E=\mathbf{I}_i} \mathrm{Tr}\big[\mathbf{G}_E^T\mathbf{B}\mathbf{G}_E\big] \;= \beta_1+\beta_2+\cdots+\beta_i,$$

where $\beta_1,\ \beta_2,\ \ldots,\ \beta_i$ are i smallest eigenvalues of $\mathbf{B}$ and minimum value can be achieved by the matrix $\mathbf{V}$ whose columns are corresponding eigenvectors. If the number of negative eigenvalues of $\mathbf{B}$ is less than r, then the optimum i in (10) is j, otherwise the optimum i is r. □

2. Empirical Moments Based Solution to Linear Encoder

In many practical scenarios, we only have access to data samples but not to the true mean vectors and covariance matrices. Therefore, the solution in Section 3.2 might not be feasible in such as case. In this Section, we provide an approach to solve the optimization problem in Section 3.2 which relies on empirical moments and is valid even if the covariance matrix $\mathbf{C}_x$ is not full-rank.

Firstly, for a given $\mathbf{\Theta}_E$, we find

$$J_y = \min_{\mathbf{W}_y,\mathbf{b}_y} \mathrm{MSE}\,(\hat{\mathbf{y}}-\mathbf{y}).$$

Note that the above optimization problem can be separated over $\mathbf{W}_y$, $\mathbf{b}_y$. Therefore, for a given $\mathbf{W}_y$, we first minimize over $\mathbf{b}_y$:

$$\begin{aligned}
&\min_{\mathbf{b}_y} \mathbb{E}\Big\{\big\|\mathbf{W}_y\mathbf{\Theta}_E\mathbf{x}+\mathbf{b}_y-\mathbf{y}\big\|^2\Big\}\\
=\;&\min_{\mathbf{b}_y} \frac{1}{n}\sum_{k=1}^{n}\big\|\mathbf{W}_y\mathbf{\Theta}_E\mathbf{x}_k+\mathbf{b}_y-\mathbf{y}_k\big\|^2\\
=\;&\frac{1}{n}\sum_{k=1}^{n}\big\|\mathbf{W}_y\mathbf{\Theta}_E\mathbf{x}_k+\mathbf{c}-\mathbf{y}_k\big\|^2
\end{aligned}$$

where we used empirical expectation in the second stage and the minimizer $\mathbf{c}$ is

$$\begin{aligned}
\mathbf{c} \;&=\; \frac{1}{n}\sum_{k=1}^{n}\Big(\mathbf{y}_k-\mathbf{W}_y\mathbf{\Theta}_E\mathbf{x}_k\Big)\\
&=\; \frac{1}{n}\sum_{k=1}^{n}\mathbf{y}_k-\mathbf{W}_y\mathbf{\Theta}_E\frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k\\
&=\; \mathbb{E}\big\{\mathbf{y}\big\}-\mathbf{W}_y\mathbf{\Theta}_E\,\mathbb{E}\big\{\mathbf{x}\big\}
\end{aligned}\tag{E}$$

Let all the columns of matrix $\mathbf{C}$ be equal to $\mathbf{c}$. We now have,

$$\begin{aligned}
J_y \;&=\; \min_{\mathbf{W}_y,\mathbf{b}_y} \mathrm{MSE}\,(\hat{\mathbf{y}}-\mathbf{y})\\
&=\; \min_{\mathbf{W}_y} \frac{1}{n}\big\|\mathbf{W}_y\mathbf{\Theta}_E\mathbf{X}+\mathbf{C}-\mathbf{Y}\big\|_F^2\\
&=\; \min_{\mathbf{W}_y} \frac{1}{n}\big\|\mathbf{W}_y\mathbf{\Theta}_E\tilde{\mathbf{X}}-\tilde{\mathbf{Y}}\big\|_F^2\\
&=\; \min_{\mathbf{W}_y} \frac{1}{n}\big\|\tilde{\mathbf{X}}^T\mathbf{\Theta}_E^T\mathbf{W}_y^T-\tilde{\mathbf{Y}}^T\big\|_F^2\\
&=\; \min_{\mathbf{W}_y} \frac{1}{n}\big\|\mathbf{M}\mathbf{W}_y^T-P_{\mathcal{M}}\tilde{\mathbf{Y}}^T\big\|_F^2+\frac{1}{n}\big\|P_{\mathcal{M}^\perp}\tilde{\mathbf{Y}}^T\big\|_F^2\\
&=\; \frac{1}{n}\big\|\underbrace{\mathbf{M}\mathbf{M}^\dagger}_{P_{\mathcal{M}}}P_{\mathcal{M}}\tilde{\mathbf{Y}}^T-P_{\mathcal{M}}\tilde{\mathbf{Y}}^T\big\|_F^2+\frac{1}{n}\big\|P_{\mathcal{M}^\perp}\tilde{\mathbf{Y}}^T\big\|_F^2\\
&=\; \frac{1}{n}\big\|P_{\mathcal{M}^\perp}\tilde{\mathbf{Y}}^T\big\|_F^2\\
&=\; \frac{1}{n}\big\|\tilde{\mathbf{Y}}^T\big\|_F^2-\frac{1}{n}\big\|P_{\mathcal{M}}\tilde{\mathbf{Y}}^T\big\|_F^2
\end{aligned}$$

Algorithm 1 Spectral Adversarial Representation Learning

Input: data $\mathbf{X}$, target labels $\mathbf{Y}$, sensitive labels $\mathbf{S}$, tolerable leakage $\alpha_{\min} \leq \alpha_{\text{tol}} \leq \alpha_{\max}$, ϵ
Output: linear encoder parameters $\mathbf{\Theta}_E$
$\mathbf{L}_x \leftarrow$ orthonormalize basis of $\tilde{\mathbf{X}}^T$
Initiate $\lambda = 1/2$, $\lambda_{\min} = 0$ and $\lambda_{\max} = 1$
do
    Calculate $\mathbf{B}$ in (G)
    $\mathbf{G}_E \leftarrow$ eigenvectors of negative eigenvalues of $\mathbf{B}$
    $\mathbf{\Theta}_E \leftarrow \mathbf{G}_E^T \mathbf{L}_x^T (\tilde{\mathbf{X}})^\dagger$
    Calculate α using (F)
    if $\alpha < (\alpha_{\text{tol}} - \epsilon)$ **then** $\lambda_{\min} = \lambda$ and $\lambda \leftarrow (\lambda + \lambda_{\max})/2$
    else if $\alpha > (\alpha_{\text{tol}} + \epsilon)$ **then** $\lambda_{\max} = \lambda$ and $\lambda \leftarrow (\lambda + \lambda_{\min})/2$
    end if
while $\left|\alpha - \alpha_{\text{tol}}\right| \geq \epsilon$

where in the third step we used (E), $\mathbf{M} = \tilde{\mathbf{X}}^T \mathbf{\Theta}_E^T$ and the fifth step is due to orthogonal decomposition. Using the same approach, we get

$$J_s \quad = \quad \frac{1}{n}\left\|\tilde{\mathbf{S}}^T\right\|_F^2 - \frac{1}{n}\left\|P_{\mathcal{M}}\tilde{\mathbf{S}}^T\right\|_F^2 \tag{F}$$

Now, assume that the columns of $\mathbf{L}_x$ are orthogonal basis for the column space of $\tilde{\mathbf{X}}^T$. Therefore, for any $\mathbf{M}$, there exist a $\mathbf{G}_E$ such that $\mathbf{L}_x \mathbf{G}_E = \mathbf{M}$. In general, there is no bijection between $\mathbf{\Theta}_E$ and $\mathbf{G}_E$ in the equality $\tilde{\mathbf{X}}^T \mathbf{\Theta}_E^T = \mathbf{L}_x \mathbf{G}_E$. But, there is a bijection between $\mathbf{G}$ and $\mathbf{\Theta}_E$ restricted to $\mathbf{\Theta}_E$'s in which $\mathcal{R}(\mathbf{\Theta}_E^T) \subseteq \mathcal{N}(\tilde{\mathbf{X}}^T)^\perp$. This restricted bijection is sufficient to be considered, since for any $\mathbf{\Theta}_E^T \in \mathcal{N}(\tilde{\mathbf{X}}^T)$ we have $\mathbf{M} = \mathbf{0}$. Once $\mathbf{G}$ is determined, $\mathbf{\Theta}_E^T$ can be obtained as,

$$\mathbf{\Theta}_E^T = (\tilde{\mathbf{X}}^T)^\dagger \mathbf{L}_x \mathbf{G}_E + \mathbf{\Theta}_0, \;\; \mathbf{\Theta}_0 \subseteq \mathcal{N}(\tilde{\mathbf{X}}^T).$$

However, since

$$\left\|\mathbf{\Theta}_E\right\|_F^2 = \left\|\mathbf{\Theta}_E^T\right\|_F^2 = \left\|(\tilde{\mathbf{X}}^T)^\dagger \mathbf{L}_x \mathbf{G}_E\right\|_F^2 + \left\|\mathbf{\Theta}_0\right\|_F^2,$$

choosing $\mathbf{\Theta}_0 = \mathbf{0}$ results in minimum $\left\|\mathbf{\Theta}_E\right\|_F$, which is favorable in terms of robustness to noise. By choosing $\mathbf{\Theta}_0 = \mathbf{0}$, determining the encoder $\mathbf{\Theta}_E$ would be equivalent to determining $\mathbf{G}_E$. Similar to (7), we have $P_{\mathcal{M}} = \mathbf{L}_x P_{\mathcal{G}} \mathbf{L}_x^T$. If we assume that the rank of $P_{\mathcal{G}}$ is i, $J_\lambda(\mathbf{G}_E)$ in (12) can be expressed as,

$$J_\lambda(\mathbf{G}_E) = \lambda\left\|\mathbf{L}_x \mathbf{G}_E \mathbf{G}_E^T \mathbf{L}_x^T \tilde{\mathbf{S}}^T\right\|_F^2 - (1-\lambda)\left\|\mathbf{L}_x \mathbf{G}_E \mathbf{G}_E^T \mathbf{L}_x^T \tilde{\mathbf{Y}}^T\right\|_F^2$$

where $\mathbf{G}_E \mathbf{G}_E^T = P_{\mathcal{G}}$ for some orthogonal matrix $\mathbf{G}_E \in \mathbb{R}^{d \times i}$. This resembles the optimization problem in (10) and therefore it has the same solution as Theorem 3 with modified $\mathbf{B}$ given by

$$\mathbf{B} \quad = \quad \mathbf{L}_x^T \Big(\lambda \tilde{\mathbf{S}}^T \tilde{\mathbf{S}} - (1-\lambda)\tilde{\mathbf{Y}}^T \tilde{\mathbf{Y}}\Big) \mathbf{L}_x. \tag{G}$$

Once $\mathbf{G}_E$ is determined, $\mathbf{\Theta}_E$ can be obtained as $\mathbf{G}_E^T \mathbf{L}_x^T (\tilde{\mathbf{X}})^\dagger$. Algorithm 1 summarizes our entire solution for the case if one wishes to consider the constrained optimization problem in (8) instead of Lagrangian version of it in (9).

3. Non-linear Extension Through Kernelization

We assume that $\mathbf{x}$ is non-linearly mapped to $\phi_x(\mathbf{x})$ as illustrated in Figure 1. From the representer theorem (see[4]), we note that $\mathbf{\Theta}_E$ can be expressed as $\mathbf{\Theta}_E = \mathbf{\Lambda} \tilde{\mathbf{\Phi}}_x^T$. Consequently the embedded representation $\boldsymbol{z}$ can be computed as,

$$\mathbf{z} = \mathbf{\Theta}_E \phi_x(\mathbf{x}) = \mathbf{\Lambda} \tilde{\mathbf{\Phi}}_x^T \phi_x(\mathbf{x}) = \mathbf{\Lambda} \mathbf{D}^T [k_x(\mathbf{x}_1, \mathbf{x}), \cdots, k_x(\mathbf{x}_n, \mathbf{x})]^T$$

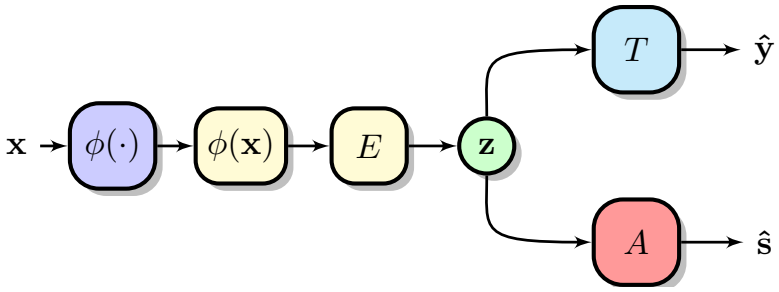


Figure 1: **Kernelized Adversarial Representation Learning** consists of four entities, a kernel $\phi_x(\cdot)$, an encoder E that obtains a compact representation $\boldsymbol{z}$ of the mapped input data $\phi_x(\boldsymbol{x})$, a predictor T that predicts a desired target attribute $\boldsymbol{y}$ and an adversary that seeks to extract a sensitive attribute $\boldsymbol{s}$, both from the embedding $\boldsymbol{z}$.

3.1. Learning

First, for a given fixed $\boldsymbol{\Theta}_E$, we find

$$J_y = \min_{\mathbf{W}_y, \mathbf{b}_y} \text{MSE}\,(\hat{\mathbf{y}} - \mathbf{y}).$$

Note that the above optimization problem can be separated over $\mathbf{W}_y$, $\mathbf{b}_y$. Therefore, for a given $\mathbf{W}_y$, we first minimize over $\mathbf{b}_y$:

$$\begin{aligned} & \min_{\mathbf{b}_y} \mathbb{E}\Big\{ \big\| \mathbf{W}_y \boldsymbol{\Theta}_E \phi_x(\mathbf{x}) + \mathbf{b}_y - \mathbf{y} \big\|^2 \Big\} \\ = \; & \min_{\mathbf{b}_y} \frac{1}{n} \sum_{k=1}^{n} \big\| \mathbf{W}_y \boldsymbol{\Theta}_E \phi_x(\mathbf{x}_k) + \mathbf{b}_y - \mathbf{y}_k \big\|^2 \\ = \; & \frac{1}{n} \sum_{k=1}^{n} \big\| \mathbf{W}_y \boldsymbol{\Theta}_E \phi_x(\mathbf{x}_k) + \mathbf{c} - \mathbf{y}_k \big\|^2 \end{aligned}$$

where the minimizer $\mathbf{c}$ is,

$$\begin{aligned} \mathbf{c} \; &= \; \frac{1}{n} \sum_{k=1}^{n} \Big(\mathbf{y}_k - \mathbf{W}_y \boldsymbol{\Theta}_E \phi_x(\mathbf{x}_k) \Big) \\ &= \; \frac{1}{n} \sum_{k=1}^{n} \mathbf{y}_k - \mathbf{W}_y \boldsymbol{\Theta}_E \frac{1}{n} \sum_{k=1}^{n} \phi_x(\mathbf{x}_k) \\ &= \; \mathbb{E}\big\{ \mathbf{y} \big\} - \mathbf{W}_y \boldsymbol{\Theta}_E \, \mathbb{E}\big\{ \phi_x(\mathbf{x}) \big\}. \end{aligned} \tag{H}$$

Let all the columns of $\mathbf{C}$ be equal to $\mathbf{c}$. Therefore we now have,

$$\begin{aligned}
&\min_{\mathbf{W}_y,\mathbf{b}_y} \text{MSE}\,(\hat{\mathbf{y}}-\mathbf{y}) \\
=\;& \min_{\mathbf{W}_y} \frac{1}{n}\big\|\mathbf{W}_y\boldsymbol{\Theta}_E\boldsymbol{\Phi}_x+\mathbf{C}-\mathbf{Y}\big\|_F^2 \\
=\;& \min_{\mathbf{W}_y} \frac{1}{n}\big\|\mathbf{W}_y\boldsymbol{\Theta}_E\tilde{\boldsymbol{\Phi}}_x-\tilde{\mathbf{Y}}\big\|_F^2 \\
=\;& \min_{\mathbf{W}_y} \frac{1}{n}\big\|\tilde{\boldsymbol{\Phi}}_x^T\boldsymbol{\Theta}_E^T\mathbf{W}_y^T-\tilde{\mathbf{Y}}^T\big\|_F^2 \\
=\;& \min_{\mathbf{W}_y} \frac{1}{n}\big\|\mathbf{M}\mathbf{W}_y^T-P_{\mathcal{M}}\tilde{\mathbf{Y}}^T\big\|_F^2+\frac{1}{n}\big\|P_{\mathcal{M}^\perp}\tilde{\mathbf{Y}}^T\big\|_F^2 \\
=\;& \frac{1}{n}\big\|\underbrace{\mathbf{M}\mathbf{M}^\dagger}_{P_{\mathcal{M}}}P_{\mathcal{M}}\tilde{\mathbf{Y}}^T-P_{\mathcal{M}}\tilde{\mathbf{Y}}^T\big\|_F^2+\frac{1}{n}\big\|P_{\mathcal{M}^\perp}\tilde{\mathbf{Y}}^T\big\|_F^2 \\
=\;& \frac{1}{n}\big\|P_{\mathcal{M}^\perp}\tilde{\mathbf{Y}}^T\big\|_F^2 \\
=\;& \frac{1}{n}\big\|\tilde{\mathbf{Y}}^T\big\|_F^2-\frac{1}{n}\big\|P_{\mathcal{M}}\tilde{\mathbf{Y}}^T\big\|_F^2 \qquad \text{(I)}
\end{aligned}$$

where the third step is due to (H), $\mathbf{M}=\tilde{\boldsymbol{\Phi}}_x^T\boldsymbol{\Theta}_E^T$ and the fifth step is the orthogonal decomposition w.r.t. $\mathbf{M}$. Using the same approach, we get

$$J_s \;=\; \frac{1}{n}\big\|\tilde{\mathbf{S}}^T\big\|_F^2-\frac{1}{n}\big\|P_{\mathcal{M}}\tilde{\mathbf{S}}^T\big\|_F^2 \qquad \text{(J)}$$

Finding optimal $\boldsymbol{\Theta}_E$ is equivalent to finding optimal $\boldsymbol{\Lambda}$ (since $\boldsymbol{\Theta}_E=\boldsymbol{\Lambda}\tilde{\boldsymbol{\Phi}}_x^T$) where we would have $\mathbf{M}=\tilde{\boldsymbol{\Phi}}_x^T\tilde{\boldsymbol{\Phi}}_x\boldsymbol{\Lambda}^T=\tilde{\mathbf{K}}_x\boldsymbol{\Lambda}^T$. Now, assume that the columns of $\mathbf{L}_x$ are orthogonal basis for the column space of $\tilde{\mathbf{K}}_x$. As a result, for any $\mathbf{M}$, there exist $\mathbf{G}_E$ such that $\mathbf{L}_x\mathbf{G}_E=\mathbf{M}$. In general, there is no bijection between $\boldsymbol{\Lambda}$ and $\mathbf{G}_E$ in the equality $\tilde{\mathbf{K}}_x\boldsymbol{\Lambda}^T=\mathbf{L}_x\mathbf{G}_E$. But, there is a bijection between $\mathbf{G}_E$ and $\boldsymbol{\Lambda}$ restricted to $\boldsymbol{\Lambda}$'s in which $\mathcal{R}(\boldsymbol{\Lambda}^T)\subseteq\mathcal{N}(\tilde{\mathbf{K}}_x)^\perp$. This restricted bijection is sufficient, since for any $\boldsymbol{\Lambda}^T\in\mathcal{N}(\tilde{\mathbf{K}}_x)$ we have $\mathbf{M}=\mathbf{0}$. Once $\mathbf{G}_E$ is determined, $\boldsymbol{\Lambda}^T$ can be obtained as,

$$\boldsymbol{\Lambda}^T=(\tilde{\mathbf{K}}_x)^\dagger\mathbf{L}_x\mathbf{G}_E+\boldsymbol{\Lambda}_0,\;\;\boldsymbol{\Lambda}_0\subseteq\mathcal{N}(\tilde{\mathbf{K}}_x)$$

However, since

$$\big\|\boldsymbol{\Lambda}\big\|_F^2=\big\|\boldsymbol{\Lambda}^T\big\|_F^2=\big\|(\tilde{\mathbf{K}}_x)^\dagger\mathbf{L}_x\mathbf{G}_E\big\|_F^2+\big\|\boldsymbol{\Lambda}_0\big\|_F^2,$$

choosing $\boldsymbol{\Lambda}_0=\mathbf{0}$ results in minimum $\big\|\boldsymbol{\Lambda}\big\|_F$, which is favorable in terms of robustness to the noise. Similar to (7), we have $P_{\mathcal{M}}=\mathbf{L}_xP_{\mathcal{G}}\mathbf{L}_x^T$. If we assume that the rank of $P_{\mathcal{G}}$ is i, $J_\lambda(\mathbf{G}_E)$ in (12) can be expressed as,

$$J_\lambda(\mathbf{G}_E)=\lambda\big\|\mathbf{L}_x\mathbf{G}_E\mathbf{G}_E^T\mathbf{L}_x^T\tilde{\mathbf{S}}^T\big\|_F^2-(1-\lambda)\big\|\mathbf{L}_x\mathbf{G}_E\mathbf{G}_E^T\mathbf{L}_x^T\tilde{\mathbf{Y}}^T\big\|_F^2$$

where $P_{\mathcal{G}}=\mathbf{G}_E\mathbf{G}_E^T$ for some orthogonal matrix $\mathbf{G}_E\in\mathbb{R}^{d\times i}$. This resembles the optimization problem in (10) and therefore have the same solution as Theorem 3 with modified $\mathbf{B}$ as,

$$\mathbf{B} \;=\; \mathbf{L}_x^T\Big(\lambda\tilde{\mathbf{S}}^T\tilde{\mathbf{S}}-(1-\lambda)\tilde{\mathbf{Y}}^T\tilde{\mathbf{Y}}\Big)\mathbf{L}_x \qquad \text{(K)}$$

Once $\mathbf{G}_E$ is determined, $\boldsymbol{\Lambda}$ can be computed as $\mathbf{G}_E^T\mathbf{L}_x^T(\tilde{\mathbf{K}}_x^T)^\dagger$. Algorithm 1 summarizes our entire solution (replacing $\tilde{\mathbf{X}}$ by $\tilde{\mathbf{K}}_x^T$ in steps 3 and 8) if one wishes to consider the constrained optimization problem in (8) instead of unconstrained Lagrangian version in (9). It is worth of mentioning that the objective function $J_\lambda(\mathbf{G}_E)$ is neither convex nor differentiable. The proof is exactly the same as Theorem 3.

3.2. Proof of Lemma 4

Lemma 4. *Let the columns of $\mathbf{L}_x$ be the orthonormal basis for $\tilde{\mathbf{K}}_x$ (in linear case $\tilde{\mathbf{K}}_x=\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$). Further, assume that the columns of $\mathbf{V}_s$ are the singular vectors corresponding to zero singular values of $\tilde{\mathbf{S}}\mathbf{L}_x$ and the columns of $\mathbf{V}_y$ are the singular*

vectors corresponding to non-zero singular values of $\tilde{\mathbf{Y}}\mathbf{L}_x$*. Then, the MSE for the adversary and the target are bounded on both sides i.e.,* $\alpha_{\min} \leq J_s \leq \alpha_{\max}$ *and* $\gamma_{\min} \leq J_y \leq \gamma_{\max}$*:*

$$\gamma_{\min} = \frac{1}{n}\big\|\tilde{\mathbf{Y}}^T\big\|_F^2 - \frac{1}{n}\|\tilde{\mathbf{Y}}\mathbf{L}_x\|_F^2$$
$$\gamma_{\max} = \frac{1}{n}\big\|\tilde{\mathbf{Y}}^T\big\|_F^2 - \frac{1}{n}\big\|\tilde{\mathbf{Y}}\mathbf{L}_x\mathbf{V}_s\big\|_F^2$$
$$\alpha_{\min} = \frac{1}{n}\big\|\tilde{\mathbf{S}}^T\big\|_F^2 - \frac{1}{n}\big\|\tilde{\mathbf{S}}\mathbf{L}_x\mathbf{V}_y\big\|_F^2$$
$$\alpha_{\max} = \frac{1}{n}\big\|\tilde{\mathbf{S}}^T\big\|_F^2$$

Proof. First, let us ignore the objective corresponding to leakage of the sensitive attribute in (8) or equivalently set $\lambda = 0$ in equation (9). In this scenario, J_y achieves its minimum possible value (denoted by $\gamma_{\min}$) as,

$$\begin{aligned}
\gamma_{\min} &= \frac{1}{n}\big\|\tilde{\mathbf{Y}}^T\big\|_F^2 - \frac{1}{n}\max_{\mathbf{\Theta}_E}\big\|P_{\mathcal{M}}\tilde{\mathbf{Y}}^T\big\|_F^2 \\
&= \frac{1}{n}\big\|\tilde{\mathbf{Y}}^T\big\|_F^2 - \frac{1}{n}\max_{\mathbf{G}_E}\big\|\mathbf{L}_x P_{\mathcal{G}}\mathbf{L}_x^T\tilde{\mathbf{Y}}^T\big\|_F^2 \\
&= \frac{1}{n}\big\|\tilde{\mathbf{Y}}^T\big\|_F^2 - \frac{1}{n}\max_i\Big\{\max_{\mathbf{G}_E^T\mathbf{G}_E=\mathbf{I}_i}\mathrm{Tr}\big[\mathbf{G}_E^T\mathbf{L}_x^T\tilde{\mathbf{Y}}^T\tilde{\mathbf{Y}}\mathbf{L}_x\mathbf{G}_E\big]\Big\} \\
&= \frac{1}{n}\big\|\tilde{\mathbf{Y}}^T\big\|_F^2 - \frac{1}{n}\mathrm{Tr}\big[\mathbf{V}_y^T\mathbf{L}_x^T\tilde{\mathbf{Y}}^T\tilde{\mathbf{Y}}\mathbf{L}_x\mathbf{V}_y\big] \\
&= \frac{1}{n}\big\|\tilde{\mathbf{Y}}^T\big\|_F^2 - \frac{1}{n}\sum_k \sigma_k^2 \\
&= \frac{1}{n}\big\|\tilde{\mathbf{Y}}^T\big\|_F^2 - \frac{1}{n}\big\|\tilde{\mathbf{Y}}\mathbf{L}_x\big\|_F^2 \qquad \text{(L)}
\end{aligned}$$

where the fourth step is borrowed from trace optimization problems studied in [3] and σ_k's are the singular values of $\tilde{\mathbf{Y}}\mathbf{L}_x$. Now, we show how to reduce the amount of leakage without degrading the performance of the target task. For this purpose, assume that columns of matrix $\mathbf{G}_E$ is the concatenation of the columns of $\mathbf{V}_y$ together with at least one singular vector corresponding to a zero singular value of $\tilde{\mathbf{Y}}\mathbf{L}_x$. Since $\mathcal{V}_y \subseteq \mathcal{G}$, therefore $\|\mathbf{L}_x P_{\mathcal{V}_y}\mathbf{L}_x^T\mathbf{U}\|_F^2 \leq \|\mathbf{L}_x P_{\mathcal{G}}\mathbf{L}_x^T\mathbf{U}\|_F^2$ for any arbitrary matrix $\mathbf{U}$. As a result, $J_s(\mathbf{G}_E) \geq J_s(\mathbf{V}_y)$. Reducing $\mathbf{V}_y$ by excluding all singular vectors associated with zero singular values form J_y does not change $\gamma_{\min}$ (step five in (L)), but will increase J_s. As a result, $\alpha_{\min}$ in the constrained optimization problem (8) which is associated to the maximum leakage of sensitive attributes is,

$$\begin{aligned}
\alpha_{\min} &= \frac{1}{n}\big\|\tilde{\mathbf{S}}^T\big\|_F^2 - \frac{1}{n}\big\|\mathbf{L}_x P_{\mathcal{V}_y}\mathbf{L}_x^T\tilde{\mathbf{S}}^T\big\|_F^2 \\
&= \frac{1}{n}\big\|\tilde{\mathbf{S}}^T\big\|_F^2 - \frac{1}{n}\mathrm{Tr}\big[\mathbf{V}_y^T\mathbf{L}_x^T\tilde{\mathbf{S}}^T\tilde{\mathbf{S}}\mathbf{L}_x\mathbf{V}_y\big] \\
&= \frac{1}{n}\big\|\tilde{\mathbf{S}}^T\big\|_F^2 - \frac{1}{n}\big\|\tilde{\mathbf{S}}\mathbf{L}_x\mathbf{V}_y\big\|_F^2.
\end{aligned}$$

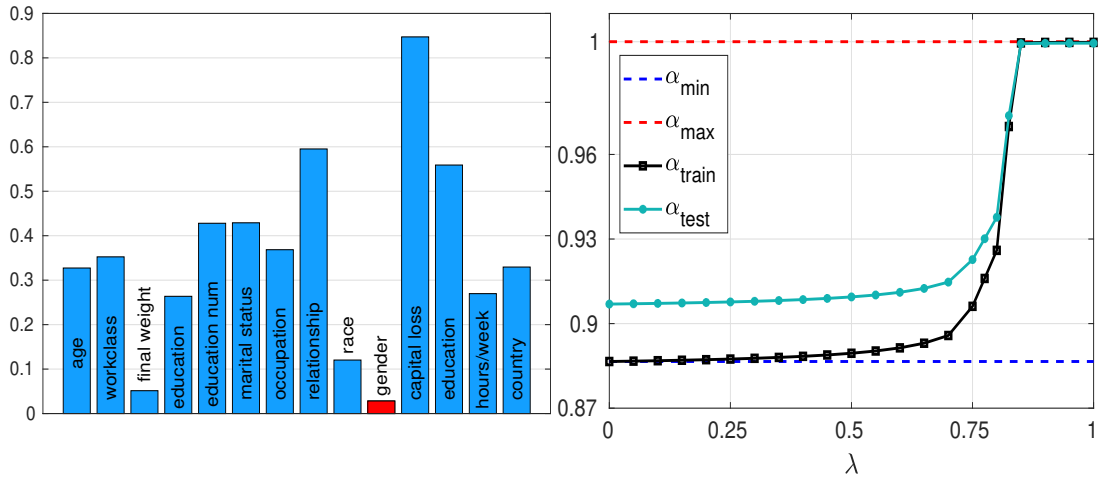
Now, consider the situation where we only seek to prevent leakage of sensitive attributes i.e., the objective of optimization problem in (8) is ignored or equivalently setting $\lambda = 1$ in equation (9). In this case, $\alpha_{\max}$ in the constrained optimization problem (8) which is associated to the minimum leakage of sensitive attributes is,

$$\alpha_{\max} = \frac{1}{n}\big\|\tilde{\mathbf{S}}^T\big\|_F^2$$

which can be achieved via trivial choice of $\mathbf{V}_s = 0$. However, we let the columns of $\mathbf{V}_s$ be the singular vectors corresponding to all zero singular values of $\tilde{\mathbf{S}}\mathbf{L}_x$ to maximize $\big\|P_{\mathcal{M}}\tilde{\mathbf{Y}}^T\big\|_F$ and consequently minimize J_y. As a result, the maximum J_y is,

$$\gamma_{\max} = \frac{1}{n}\big\|\tilde{\mathbf{Y}}^T\big\|_F^2 - \frac{1}{n}\big\|\tilde{\mathbf{Y}}\mathbf{L}_x\mathbf{V}_s\big\|_F^2.$$

□



(a) Adult Dataset Encoder

(b) CIFAR-100: Adversary Bounds

4. Numerical Experiments

For the adult dataset, the linear encoder maps the 14 input features to just one dimension. The weights assigned to each feature is shown in Figure 2a. Notice that the encoder assigns almost zero weight to the gender feature in order to be fair with respect to the gender attribute.

Figure 2b shows the mean squared error (MSE) of the adversary for the CIFAR-100 experiment as a function of the Lagrange multiplier λ. The plot illustrates, (a) the lower and upper bounds $\alpha_{\min}$ and $\alpha_{\max}$ respectively calculated on the training dataset, (b) achievable adversary MSE computed on the training set α_{train}, and finally (c) achievable adversary MSE computed on the test set α_{test}. Observe that on the training dataset all values of $\alpha \in [\alpha_{\min}, \alpha_{\max}]$ are reachable as we sweep through $\lambda \in [0, 1]$. This is however not the case on the test set since the bounds are computed using empirical moments as opposed to the true covariance matrices.

5. Computational Complexity

Solving the optimization problem runs in $\mathcal{O}(d^3)$ since we need to eigendecompose the $d \times d$ matrix $\mathbf{B}$. Both Cholesky factorization $\mathbf{C}_x = \mathbf{Q}_x^T \mathbf{Q}_x$ and obtaining $\mathbf{Q}_x^{-1}$ require $\mathcal{O}(d^3)$. Obtaining the mapping $\mathbf{\Theta}_E$ from $\mathbf{G}$ takes $\mathcal{O}(d^3)$ again. Calculating covariance matrices $\mathbf{C}_x$, $\mathbf{C}_{yx}$ and $\mathbf{C}_{sx}$ can be done in $\mathcal{O}(d^2 n)$, $\mathcal{O}(p^2 n)$ and $\mathcal{O}(q^2 d)$ respectively. In Kernel-SARL, eigendecomposition of $\mathbf{B}$ requires $\mathcal{O}(n^3)$. However, for scalability i.e., large n (e.g., CIFAR-100), the Nyström method (i.e., sampling the data) can be adopted.

References

[1] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004. 3

[2] G. H. Golub and V. Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on numerical analysis*, 10(2):413–432, 1973. 3

[3] E. Kokiopoulou, J. Chen, and Y. Saad. Trace optimization and eigenproblems in dimension reduction methods. *Numerical Linear Algebra with Applications*, 18(3):565–602, 2011. 4, 8

[4] J. Shawe-Taylor, N. Cristianini, et al. *Kernel methods for pattern analysis*. Cambridge university press, 2004. 5