Multi-view Metric Learning in Vector-valued Kernel Spaces Supplementary Material

Riikka Huusari **Hachem Kadri** **Cécile Capponi**
Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France

A Appendix

A.1 MVML optimization

Here we go through the derivations of the solutions $\mathbf{A}$, $\mathbf{D}$ and $\mathbf{w}$ for our optimization problem. The presented derivations are for the case without Nyström approximation; however the derivations with Nyström approximation are done exactly the same way.

Solving for g and w

Let us first focus on the case where $\mathbf{A}$ and $\mathbf{w}$ are fixed and we solve for $\mathbf{g}$. We calculate the derivative of the expression in Equation (7):

$$\begin{aligned}
&\frac{d}{d\mathbf{g}} \ \|\mathbf{y}-(\mathbf{w}^T\otimes\mathbf{I}_n)\mathbf{H}\mathbf{g}\|^2+\lambda\left\langle\mathbf{g},\mathbf{A}^\dagger\mathbf{g}\right\rangle\\
&=\frac{d}{d\mathbf{g}} \ \langle\mathbf{y},\mathbf{y}\rangle-2\langle\mathbf{y},(\mathbf{w}^T\otimes\mathbf{I}_n)\mathbf{H}\mathbf{g}\rangle\\
&\quad+\langle(\mathbf{w}^T\otimes\mathbf{I}_n)\mathbf{H}\mathbf{g},(\mathbf{w}^T\otimes\mathbf{I}_n)\mathbf{H}\mathbf{D}\rangle+\lambda\langle\mathbf{g},\mathbf{A}^\dagger\mathbf{g}\rangle\\
&=-2\mathbf{H}(\mathbf{w}^T\otimes\mathbf{I}_n)^T\mathbf{y}\\
&\quad+2\mathbf{H}(\mathbf{w}^T\otimes\mathbf{I}_n)^T(\mathbf{w}^T\otimes\mathbf{I}_n)\mathbf{H}\mathbf{g}+2\lambda\mathbf{A}^\dagger\mathbf{g}
\end{aligned}$$

By setting this to zero we obtain the solution

$$\mathbf{g}=(\mathbf{H}(\mathbf{w}^T\otimes\mathbf{I}_n)^T(\mathbf{w}^T\otimes\mathbf{I}_n)\mathbf{H}+\lambda\mathbf{A}^\dagger)^{-1}\mathbf{H}(\mathbf{w}^T\otimes\mathbf{I}_n)^T\mathbf{y}.$$

As for $\mathbf{w}$ when $\mathbf{A}$ and $\mathbf{g}$ are fixed, we need only to consider optimizing

$$\min_{\mathbf{w}} \ \|\mathbf{y}-(\mathbf{w}^T\otimes\mathbf{I}_n)\mathbf{H}\mathbf{g}\|^2. \tag{19}$$

If we denote that $\mathbf{Z}\in\mathbb{R}^{n\times v}$ is equal to reshaping $\mathbf{Hg}$ by taking the elements of the vector and arranging them onto the columns of $\mathbf{Z}$, we obtain a following form:

$$\min_{\mathbf{w}} \ \|\mathbf{y}-\mathbf{Z}\mathbf{w}\|^2. \tag{20}$$

One can easily see by taking the derivative and setting it to zero that the solution for this is

$$\mathbf{w}=\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{y}. \tag{21}$$

Solving for A in (6)

When we consider $\mathbf{g}$ (and $\mathbf{w}$) to be fixed in the MVML framwork (6), for $\mathbf{A}$ we have the following minimization problem:

$$\min_{\mathbf{A}} \ \lambda\left\langle\mathbf{g},\mathbf{A}^\dagger\mathbf{g}\right\rangle+\eta\|\mathbf{A}\|_F^2$$

Derivating this with respect to $\mathbf{A}$ gives us

$$\begin{aligned}
&\frac{d}{d\mathbf{A}} \ \lambda\left\langle\mathbf{g},\mathbf{A}^\dagger\mathbf{g}\right\rangle+\eta\|\mathbf{A}\|_F^2\\
&=\frac{d}{d\mathbf{A}} \ \lambda\left\langle\mathbf{g},\mathbf{A}^\dagger\mathbf{g}\right\rangle+\eta \ tr(\mathbf{A}\mathbf{A})\\
&=-\lambda\mathbf{A}^\dagger\mathbf{g}\mathbf{g}^T\mathbf{A}^\dagger \ ^7+2\eta\mathbf{A}
\end{aligned}$$

Thus the gradient descent step will be

$$\mathbf{A}^{k+1}=(1-2\mu\eta)\,\mathbf{A}^k+\mu\lambda\left(\mathbf{A}^k\right)^\dagger\mathbf{g}\mathbf{g}^T\left(\mathbf{A}^k\right)^\dagger$$

when moving to the direction of negative gradient with step size μ.

Solving for A in (11)

To solve $\mathbf{A}$ from equation (11) we use proximal minimization. Let us recall the optimization problem after the change of the variable:

$$\begin{aligned}
\min_{\mathbf{A},\mathbf{g},\mathbf{w}} \ &\|\mathbf{y}-(\mathbf{w}^T\otimes\mathbf{I}_n)\mathbf{H}\mathbf{g}\|^2+\lambda\langle\mathbf{g},\mathbf{A}^\dagger\mathbf{g}\rangle\\
&+\eta\sum_{\gamma\in\mathcal{G}}\|\mathbf{A}_\gamma\|_F,
\end{aligned}$$

and denote

$$h(\mathbf{A})=\lambda\left\langle\mathbf{g},\mathbf{A}^\dagger\mathbf{g}\right\rangle$$

and

$$\Omega(\mathbf{A})=\eta\sum_{\gamma\in\mathcal{G}}\|\mathbf{A}_\gamma\|_F$$

for the two terms in our optimization problem that contain the matrix $\mathbf{A}$.

[7]Matrix cookbook (Equation 61): `https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`.

Without going into detailed theory of proximal operators and proximal minimization, we remark that the proximal minimization algorithm update takes the form

$$\mathbf{A}^{k+1} = \mathbf{prox}_{\mu^k\Omega}(\mathbf{A}^k - \mu^k \nabla h(\mathbf{A}^k)).$$

It is well-known that in traditional group-lasso situation the proximal operator is

$$[\mathbf{prox}_{\mu^k\Omega}(\mathbf{z})]_\gamma = \left(1 - \frac{\eta}{\|\mathbf{z}_\gamma\|_2}\right)_+ \mathbf{z}_\gamma,$$

where $\mathbf{z}$ is a vector and $+$ denotes the maximum of zero and the value inside the brackets. In our case we are solving for a matrix, but due to the equivalence of Frobenious norm to vector 2-norm we can use this exact same operator. Thus we get as the proximal update

$$[\mathbf{A}^{k+1}]_\gamma = \left(1 - \frac{\eta}{\|[\mathbf{A}^k - \mu^k\nabla h(\mathbf{A}^k)]_\gamma\|_F}\right)_+ [\mathbf{A}^k - \mu^k \nabla h(\mathbf{A}^k)]_\gamma,$$

where

$$\nabla h(\mathbf{A}^k) = -\lambda(\mathbf{A}^k)^{-1}\mathbf{g}\mathbf{g}^T(\mathbf{A}^k)^{-1}.$$

We can see from the update fromula and the derivative that if $\mathbf{A}^k$ is a positive matrix, the update without block-multiplication, $\mathbf{A}^k - \mu^k\nabla h(\mathbf{A}^k)$, will be positive, too. This is unfortunately not enough to guarantee the general positivity of $\mathbf{A}^{k+1}$. However we note that it is, indeed, positive if it is block-diagonal, and in general whenever a matrix of the multipliers α

$$\alpha_{st} = \left(1 - \frac{\eta}{\|[\mathbf{A}^k - \mu^k\nabla h(\mathbf{A}^k)]_{st}\|_2}\right)_+$$

is positive, then $\mathbf{A}^{k+1}$ is, too (see [12] for reference - this is a blockwise Hadamard product where the blocks commute).

A.2 Proof of Theorem 1

Theorem 1. *Let $\mathcal{H}$ be a vector-valued RKHS associated with the the multi-view kernel K defined by Equation 4. Consider the hypothesis class $\mathcal{H}_\lambda = \{x \mapsto f_{u,\mathbf{A}}(x) = \Gamma_\mathbf{A}(x)^* u : \mathbf{A} \in \Delta, \|u\|_\mathcal{H} \le \beta\}$, with $\Delta = \{\mathbf{A} : \mathbf{A} \succ 0, \|\mathbf{A}\|_F \le \alpha\}$. The empirical Rademacher complexity of $\mathcal{H}_\lambda$ can be upper bounded as follows:*

$$\hat{\mathcal{R}}_n(\mathcal{H}_\lambda) \le \frac{\beta\sqrt{\alpha\|q\|_1}}{n},$$

where $q = \left(tr(\mathbf{K}_l^2)\right)_{l=1}^v$, and $\mathbf{K}_l$ is the Gram matrix computed from the training set $\{x_1, \dots, x_n\}$ with the kernel k_l defined on the view l. For kernels k_l such that $tr(\mathbf{K}_l^2) \le \tau n$, we have

$$\hat{\mathcal{R}}_n(\mathcal{H}_\lambda) \le \beta\sqrt{\frac{\alpha\tau v}{n}}.$$

Proof. We start by recalling that the feature map associated to the operator-valued kernel K is the mapping $\Gamma : \mathcal{X} \to \mathcal{L}(\mathcal{Y}, \mathcal{H})$, where $\mathcal{X}$ is the input space, $\mathcal{Y} = \mathbb{R}^v$, and $\mathcal{L}(\mathcal{Y}, \mathcal{H})$ is the set of bounded linear operators from $\mathcal{Y}$ to $\mathcal{H}$ (see, e.g., [19, 7] for more details). It is known that $K(x, z) = \Gamma(x)^*\Gamma(z)$. We denote by $\Gamma_\mathbf{A}$ the feature map associated to our multi-view kernel (Equation 4). We also define the matrix $\mathbf{\Sigma} = (\boldsymbol{\sigma})_{i=1}^n \in \mathbb{R}^{nv}$

$$\begin{aligned}
\hat{\mathcal{R}}_n(\mathcal{H}_\lambda) &= \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{H}}\sup_{\mathbf{A}\in\Delta}\sum_{i=1}^n \boldsymbol{\sigma}_i^\top f_{u,\mathbf{A}}(x_i)\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sup_u \sup_\mathbf{A} \sum_{i=1}^n \langle\boldsymbol{\sigma}_i, \Gamma_\mathbf{A}(x_i)^* u\rangle_{\mathbb{R}^v}\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sup_u \sup_\mathbf{A} \sum_{i=1}^n \langle\Gamma_\mathbf{A}(x_i)\boldsymbol{\sigma}_i, u\rangle_\mathcal{H}\right] \quad (1) \\
&\le \frac{\beta}{n}\mathbb{E}\left[\sup_\mathbf{A} \|\sum_{i=1}^n \Gamma_\mathbf{A}(x_i)\boldsymbol{\sigma}_i\|_\mathcal{H}\right] \quad (2) \\
&= \frac{\beta}{n}\mathbb{E}\left[\sup_\mathbf{A}\left(\sum_{i,j=1}^n \langle\boldsymbol{\sigma}_i, K_\mathbf{A}(x_i, x_j)\boldsymbol{\sigma}_j\rangle_{\mathbb{R}^v}\right)^{\frac{1}{2}}\right] \quad (3) \\
&= \frac{\beta}{n}\mathbb{E}\left[\sup_\mathbf{A}\left(\langle\mathbf{\Sigma}, \mathbf{K}_\mathbf{A}\mathbf{\Sigma}\rangle_{\mathbb{R}^{nv}}\right)^{1/2}\right] \\
&= \frac{\beta}{n}\mathbb{E}\left[\sup_\mathbf{A}\langle\mathbf{\Sigma}, \mathbf{HAH\Sigma}\rangle^{1/2}\right] \\
&= \frac{\beta}{n}\mathbb{E}\left[\sup_\mathbf{A} tr(\mathbf{H\Sigma\Sigma}^\top\mathbf{HA})^{1/2}\right] \\
&\le \frac{\beta}{n}\mathbb{E}\left[\sup_\mathbf{A} tr([\mathbf{H\Sigma\Sigma}^\top\mathbf{H}]^2)^{1/4} tr(\mathbf{A}^2)^{1/4}\right] \quad (4) \\
&\le \frac{\beta}{n}\mathbb{E}\left[\sup_\mathbf{A} tr(\mathbf{H}^2\mathbf{\Sigma\Sigma}^\top)^{1/2} tr(\mathbf{A}^2)^{1/4}\right] \\
&\le \frac{\beta\sqrt{\alpha}}{n}\mathbb{E}\left[\sup_\mathbf{A} tr(\mathbf{H}^2\mathbf{\Sigma\Sigma}^\top)^{1/2}\right] \\
&= \frac{\beta\sqrt{\alpha}}{n}\mathbb{E}\left[tr(\mathbf{H}^2\mathbf{\Sigma\Sigma}^\top)^{1/2}\right] \\
&\le \frac{\beta\sqrt{\alpha}}{n}\left(\mathbb{E}\left[tr(\mathbf{H}^2\mathbf{\Sigma\Sigma}^\top)\right]\right)^{1/2} \quad (5) \\
&= \frac{\beta\sqrt{\alpha}}{n}\left(tr\left[\mathbf{H}^2\,\mathbb{E}(\mathbf{\Sigma\Sigma}^\top)\right]\right)^{1/2} \\
&= \frac{\beta\sqrt{\alpha}}{n}\sqrt{\|(tr(\mathbf{K_1}^2), \dots, tr(\mathbf{K_v}^2))\|_1}.
\end{aligned}$$

Here (1) and (3) are obtained with reproducing property, (2) and (4) with Cauchy-Schwarz inequality, and (5) with Jensen's inequality. The last equality follows from the fact that $tr(\mathbf{H}^2) = \sum_{l=1}^v tr(\mathbf{K_l}^2)$. For kernels k_l that satisfy $tr(\mathbf{K}_l^2) \le \tau n$, $l = 1, \dots, v$, we obtain that

$$\hat{\mathcal{R}}_n(\mathcal{H}_\lambda) \le \beta\sqrt{\frac{\alpha\tau v}{n}}. \quad \square$$