From Manifold to Manifold: Geometry-Aware Dimensionality Reduction for SPD Matrices

Mehrtash T. Harandi, Mathieu Salzmann, and Richard Hartley

Australian National University, Canberra, ACT 0200, Australia
NICTA, Locked Bag 8001, Canberra, ACT 2601, Australia*

1 Proof of Length Equivalence

Here, we prove Theorem 1 from Section 3, *i.e.*, the equivalence between the length of any given curve under the geodesic distance δ_g and the Stein metric δ_S up to scale of $2\sqrt{2}$. The proof of this theorem follows several steps. We start with the definition of curve length and intrinsic metric. Without any assumption on differentiability, let $(\mathcal{M}, d)$ be a metric space. A curve in $\mathcal{M}$ is a continuous function $\gamma : [0, 1] \to \mathcal{M}$ and joins the starting point $\gamma(0) = x$ to the end point $\gamma(1) = y$.

Definition 1. *The length of a curve γ is the supremum of $l(\gamma; \{t_i\})$ over all possible partitions $\{t_i\}$, where $0 = t_0 < t_1 < \cdots < t_{n-1} < t_n = 1$ and $l(\gamma; \{t_i\}) = \sum_i d\left(\gamma(t_i), \gamma(t_{i-1})\right)$.*

Definition 2. *The intrinsic metric $\widehat{\delta}(x, y)$ on $\mathcal{M}$ is defined as the infimum of the lengths of all paths from x to y.*

Theorem 1 ([2]). *If the intrinsic metrics induced by two metrics d_1 and d_2 are identical up to a scale ξ, then the length of any given curve is the same under both metrics up to ξ.*

Theorem 2 ([2]). *If $d_1(x, y)$ and $d_2(x, y)$ are two metrics defined on a space $\mathcal{M}$ such that*

$$\lim_{d_1(x,y)\to 0} \frac{d_2(x, y)}{d_1(x, y)} = 1. \tag{1}$$

uniformly (with respect to x and y), then their intrinsic metrics are identical.

Therefore, here, we need to study the behavior of

$$\lim_{\delta_S^2(\mathbf{X},\mathbf{Y})\to 0} \frac{\delta_g^2(\mathbf{X}, \mathbf{Y})}{\delta_S^2(\mathbf{X}, \mathbf{Y})}$$

to prove our theorem on curve length equivalence.

* NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the ARC through the ICT Centre of Excellence program.

Proof. Let us first note that for an affine invariant metric δ on $\mathcal{S}_{++}^d$,

$$\delta^2(\mathbf{X},\mathbf{Y}) = \delta^2(\mathbf{I}_d, \mathbf{D}^{-1/2}\mathbf{L}^T\mathbf{Y}\mathbf{L}\mathbf{D}^{-1/2}) \triangleq \delta^2(\mathbf{I}_d,\mathbf{M}) \ ,$$

where $\mathbf{X} = \mathbf{L}\mathbf{D}\mathbf{L}^T$ and $\mathbf{L}\mathbf{L}^T = \mathbf{I}_d$. Similarly, we can decompose $\mathbf{M}$ as $\mathbf{M} = \tilde{\mathbf{L}}\tilde{\mathbf{D}}\tilde{\mathbf{L}}^T$, with $\tilde{\mathbf{L}}\tilde{\mathbf{L}}^T = \tilde{\mathbf{L}}^T\tilde{\mathbf{L}} = \mathbf{I}_d$, which yields

$$\delta^2(\mathbf{X},\mathbf{Y}) = \delta^2(\mathbf{I}_d, \tilde{\mathbf{D}}) \ .$$

Since all our matrices are positive definite, $\tilde{\mathbf{D}}$ is a diagonal matrix with strictly positive values on its diagonal, and can be written as

$$\tilde{\mathbf{D}} \triangleq \mathrm{Diag}(\exp(t\boldsymbol{\nu})) \ ,$$

with $\boldsymbol{\nu} \in \mathbb{R}^d$ and $t \in \mathbb{R}$. This definition can also be motivated by noting that the tangent vectors at $\mathbf{I}_d$ are symmetric matrices of the form $\tilde{\mathbf{L}}\mathrm{Diag}(t\boldsymbol{\nu})\tilde{\mathbf{L}}^T$. Applying the exponential map yields points on the manifold of the form $\tilde{\mathbf{L}}\mathrm{Diag}(\exp(t\boldsymbol{\nu}))\tilde{\mathbf{L}}^T$. As mentioned before, with an affine invariant metric, the dependency on $\tilde{\mathbf{L}}$ and $\tilde{\mathbf{L}}^T$ can be dropped.

The previous discussion implies that we just need to study the behavior of the Stein metric around $\mathbf{I}_d$ using a diagonal matrix to draw any conclusion. We note that $\tilde{\mathbf{D}} \to \mathbf{I}_d$ iff $t \to 0$. Therefore, given the definitions of δ_g and δ_S from Section 3 of the paper, we have

$$\begin{aligned}
\lim_{\mathbf{X}\to\mathbf{Y}} \frac{\delta_g^2(\mathbf{X},\mathbf{Y})}{\delta_S^2(\mathbf{X},\mathbf{Y})} &= \lim_{t\to 0} \frac{\delta_g^2\Big(\mathbf{I}_d, \mathrm{Diag}\big(\exp(t\boldsymbol{\nu})\big)\Big)}{\delta_S^2\Big(\mathbf{I}_d, \mathrm{Diag}\big(\exp(t\boldsymbol{\nu})\big)\Big)} \\
&= \lim_{t\to 0} \frac{\Big\| \log\Big(\mathrm{Diag}\big(\exp(t\boldsymbol{\nu})\big)\Big)\Big\|_F^2}{\ln\Big|\frac{1}{2}\mathrm{Diag}\big(1+\exp(t\boldsymbol{\nu})\big)\Big| - \frac{1}{2}\ln\Big|\mathrm{Diag}\big(\exp(t\boldsymbol{\nu})\big)\Big|} \\
&= \lim_{t\to 0} \frac{t^2\sum_{i=1}^d \nu_i^2}{\sum_{i=1}^d \ln\Big(1+\exp(t\nu_i)\Big) - t\sum_{i=1}^d \frac{\nu_i}{2} - d\ln(2)} \qquad (2) \\
&= \lim_{t\to 0} \frac{2\sum_{i=1}^d \nu_i^2}{\sum_{i=1}^d \frac{\nu_i^2\exp(t\nu_i)}{\left(1+\exp(t\nu_i)\right)^2}} = 8 \ , \qquad (3)
\end{aligned}$$

where L'Hôpital's rule was used twice from (2) to (3) since the limit in (2) is indefinite. Therefore,

$$\lim_{\mathbf{X}\to\mathbf{Y}} \frac{\delta_g(\mathbf{X},\mathbf{Y})}{\delta_S(\mathbf{X},\mathbf{Y})} = 2\sqrt{2},$$

which concludes the proof.

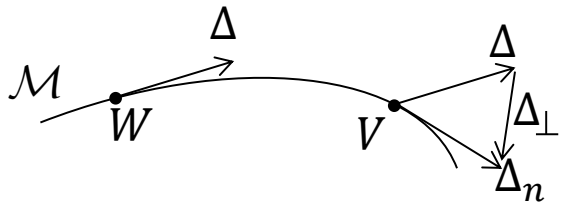

Fig. 1. Parallel transport of a tangent vector Δ from a point $\mathbf{W}$ to another point $\mathbf{V}$ on the manifold.

2 Conjugate Gradient on Grassmann Manifolds

In our formulation, we model the projection $\mathbf{W}$ as a point on a Grassmann manifold $\mathcal{G}(m, n)$. The Grassmann manifold $\mathcal{G}(m, n)$ consists of the set of all linear m-dimensional subspaces of $\mathbb{R}^n$. In particular, this lets us handle constraints of the form $\mathbf{W}^T\mathbf{W} = \mathbf{I}_m$. Learning the projection then boils down to solving a non-linear optimization problem on the Grassmann manifold. Here, we employ a conjugate gradient (CG) method on the manifold, which requires some notions of differential geometry reviewed below.

In differential geometry, the shortest path between two points on a manifold is a curve called a *geodesic*. The *tangent space* at a point on a manifold is a vector space that consists of the tangent vectors of all possible curves passing through this point. Unlike flat spaces, on a manifold one cannot transport a tangent vector Δ from one point to another point by simple translation. To get a better intuition, take the case where the manifold is a sphere, and consider two tangent spaces, one located at the pole and one at a point on the equator. Obviously the tangent vectors at the pole do not belong to the tangent space at the equator. Therefore, simple vector translation is not sufficient. As illustrated in Fig. 1, transporting Δ from $\mathbf{W}$ to $\mathbf{V}$ on the manifold $\mathcal{M}$ requires subtracting the normal component $\Delta_\perp$ at $\mathbf{V}$ for the resulting vector to be a tangent vector. Such a transfer of tangent vector is called *parallel transport*. Parallel transport is required by the CG method to compute the new descent direction by combining the gradient direction at the current and previous solutions.

On a Grassmann manifold, the above-mentioned operations have efficient numerical forms and can thus be used to perform optimization on the manifold. CG on a Grassmann manifold can be summarized by the following steps:

(i) Compute the gradient $\nabla_{\mathbf{W}} L$ of the objective function $L(\mathbf{W})$ on the manifold at the current solution using

$$\nabla_{\mathbf{W}} L = D_{\mathbf{W}} L - \mathbf{W}\mathbf{W}^T D_{\mathbf{W}} L \; . \tag{4}$$

(ii) Determine the search direction $\mathbf{H}$ by parallel transporting the previous search direction and combining it with $\nabla_{\mathbf{W}} L$.

(iii) Perform a line search along the geodesic at $\mathbf{W}$ in the direction $\mathbf{H}$. On the Grassmann manifold, the geodesics going from point $\mathbf{X}$ in direction Δ can be represented

by the Geodesic Equation [1]

$$\mathbf{X}(t) = \begin{bmatrix}\mathbf{X}\mathbf{V} & \mathbf{U}\end{bmatrix} \begin{bmatrix}\cos(\Sigma t) \\ \sin(\Sigma t)\end{bmatrix} \mathbf{V}^T \tag{5}$$

where t is the parameter indicating the location along the geodesic, and $\mathbf{U}\Sigma\mathbf{V}^T$ is the compact singular value decomposition of Δ.

These steps are repeated until convergence to a local minimum, or until a maximum number of iterations is reached.

3 Additional Experiments

3.1 Parameter Sensitivity

In all our experiments, the parameters of our approach were set in a principled manner (*i.e.*, ν_w as the minimum number of samples in one class, and ν_b by cross-validation). In this section, we nonetheless study the influence of the number of nearest neighbor from different classes (ν_b) on the overall performance. To this end, we employed the UIUC material dataset and report the accuracy of our NN-Stein-ML method when varying this parameter and fixing the other to the value reported in Section 5 ($\nu_w = 6$). Fig. 2 depicts the recognition accuracy for values of ν_b in the interval $[1, 12]$. Note that for $\nu_b = 1$, which is equivalent to mainly considering the intra-class discrimination, the performance drops. For $\nu_b = 12$, which makes the inter-class discrimination dominant, the performance drops even further. The maximum performance of 58.6% is reached for $\nu_b = 4$, which again shows that balance between the intra-class and inter-class terms is important. Note that our cross-validation procedure led to $\nu_b = 3$, which is not the optimal value on the test data, but still yields good accuracy.

3.2 Influence of the Number of Observations

Finally, as discussed in Section 4.3, we studied the sensitivity of our learning method to the number of observations used to build the RCMs. To this end, we employed the UIUC material dataset. For the training images, where computational cost is unimportant, we generated RCMs using all possible observations (our setup provided us with 9600 observations per image). For the test RCMs, we reduced the number of observations on an octave basis, *i.e.*, downsampled the number of observations by a factor of two repetitively. Fig. 3 depicts the performance of CDL, as well as of NN classifiers with both the Stein metric and the AIRM, with and without our learning scheme. The point where the number of observations r matches the size of the RCM n (*i.e.*, minimum number of observations to have a valid SPD matrix) is marked by a vertical dashed line. On the left side of this line, the number of observations is less than n. Therefore, for CDL, NN-Stein and NN-AIRM, a small regularizer of the form $\epsilon\mathbf{I}_n$ has to be added to the RCMs to make them positive definite. Note that no such regularizer was necessary when using our approach. From Fig. 3, we can see that all algorithms have a stable performance when the number of observations is large enough. When reducing the number of observations below n, the performance of CDL, NN-Stein and NN-AIRM drops down by

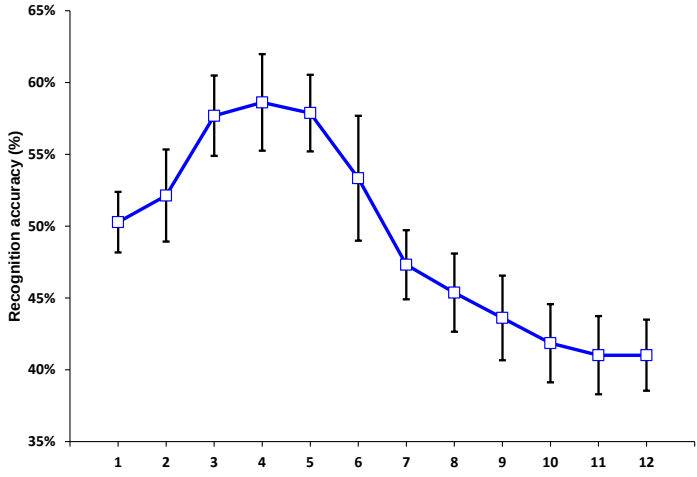

Fig. 2. Accuracy on the UIUC material dataset for varying values of ν_b.

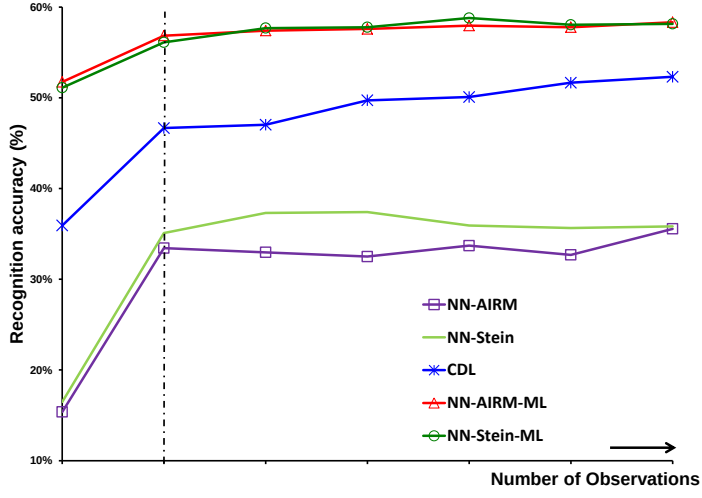

Fig. 3. Sensitivity of different algorithms to the number of observations used to create RCMs.

17%, 19% and 20%, respectively. In contrast, with our learning algorithm, the drop in performance is less than 7%.

References

1. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton, NJ, USA (2008)
2. Hartley, R., Trumpf, J., Dai, Y., Li, H.: Rotation averaging. Int. Journal of Computer Vision (IJCV) (2013)