A Appendix Overview

First, we present the details of the various proofs of Section 4 in Appendix B. Next, in Appendix C, we describe the network layers for the building GrCNF, and the detailed model architectures, hyperparameters, and implementation on the experiments. Finally, in Appendix D, we provide a summary of the fundamentals of a Grassmann manifold, which is the core concept of this study.

B Proofs

B.1 Proposition 1

First, we invoked the following two corollaries.

Corollary 1 (Diffeomorphism Invariance of Flows)**.** *Let $F : \mathcal{M} \to \mathcal{N}$ be a diffeomorphism. If X is a smooth vector field over $\mathcal{M}$ and θ is the flow of X, then the flow of $F_* X$[4] is $\eta_t = F \circ \theta_t \circ F^{-1}$, with domain $N_t = F(M_t)$ for each $t \in \mathbb{R}$.*

Proof. See Lee (2003, Corollary 9.14). □

Corollary 2 (Homogeneity Property)**.** *The horizontal lift $\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}}$ at representative $\boldsymbol{Y} \in \mathrm{St}(k, D)$ relative to $\boldsymbol{\xi}_{[\boldsymbol{Y}]} \in T_{[\boldsymbol{Y}]}\,\mathrm{Gr}(k, D)$ satisfies the following homogeneity (equivariance) property (4) with regard to $^{\forall}\boldsymbol{Q} \in \mathcal{O}(k)$.*

$$\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{YQ}} = \overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}}\boldsymbol{Q}.$$

Proof. $\pi(\boldsymbol{Y}) = \pi(\boldsymbol{YQ})$ is true for $^{\forall}\boldsymbol{Y} \in \mathrm{St}(k, D), \boldsymbol{Q} \in \mathcal{O}(k)$. Therefore, $\pi(\boldsymbol{Y}) = (\pi \circ q)(\boldsymbol{Y})$ is true when defined as $q(\boldsymbol{Y}) = \boldsymbol{YQ}$. When the derivative $d\pi(\cdot)[\cdot]$ of both sides is applied to the horizontal lift $\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}}$ of $\boldsymbol{\xi}_{[\boldsymbol{Y}]}$, the following is obtained:

$$d\pi(\boldsymbol{Y})\left[\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}}\right] = d(\pi \circ q)(\boldsymbol{Y})\left[\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}}\right] = d\pi(q(\boldsymbol{Y}))\left[dq(\boldsymbol{Y})\left[\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}}\right]\right] = d\pi(\boldsymbol{YQ})\left[\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}}\boldsymbol{Q}\right]. \tag{10}$$

Moreover, from (98) which is definition of horizontal lift, the following equation is true.

$$\boldsymbol{\xi}_{[\boldsymbol{Y}]} = d\pi(\boldsymbol{Y})\left[\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}}\right] = d\pi(\boldsymbol{YQ})\left[\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{YQ}}\right]. \tag{11}$$

Subsequently, we obtain the following equation.

$$\boldsymbol{\xi}_{[\boldsymbol{Y}]} = d\pi(\boldsymbol{YQ})\left[\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{YQ}}\right] = d\pi(\boldsymbol{YQ})\left[\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}}\boldsymbol{Q}\right]. \tag{12}$$

Finally, the uniqueness of the horizontal lift yields $\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{YQ}} = \overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}}\boldsymbol{Q}$. □

Proposition 1. *Let $\mathrm{Gr}(k, D)$ be a Grassmann manifold, X be any time-dependent vector field on $\mathrm{Gr}(k, D)$, and $F_{\mathsf{X},T}$ be a flow on a X. Let $\overline{\mathsf{X}}$ be any time-dependent horizontal lift and $\overline{F}_{\overline{\mathsf{X}},T}$ be a flow of $\overline{\mathsf{X}}$. $\overline{\mathsf{X}}$ is a vector field on $\mathrm{St}(k, D)$ if and only if $\overline{F}_{\overline{\mathsf{X}},T}$ is a flow on $\mathrm{St}(k, D)$ and satisfies invariance condition $\overline{\mathsf{X}} \sim \overline{\mathsf{X}}'$ for all $\overline{F}_{\overline{\mathsf{X}},T} \sim \overline{F}_{\overline{\mathsf{X}}',T}$. Therefore, X is a vector field on $\mathrm{Gr}(k, D)$ if and only if $F_{\mathsf{X},T} := \left[\overline{F}_{\overline{\mathsf{X}},T}\right]$ is a flow on $\mathrm{Gr}(k, D)$, and vice versa.*

Proof. **Flow $F_{\mathsf{X},T}$ on $\mathrm{Gr}(k, D)$ ⇒ Vector Field X on $\mathrm{Gr}(k, D)$.** Let $\theta : \mathrm{St}(k, D) \times \mathcal{O}(k) \to \mathrm{St}(k, D), (\boldsymbol{Y}, \boldsymbol{Q}) \mapsto \boldsymbol{YQ}$ be a map representing the right action of the orthogonal group. In addition, let $F_{\mathsf{X},T}$ be a flow on $\mathrm{Gr}(k, D)$ and $\overline{F}_{\overline{\mathsf{X}},T}$ be a flow on $\mathrm{St}(k, D)$. These satisfy $\overline{F}_{\overline{\mathsf{X}}\boldsymbol{Q},T} \sim$

[4] F_* denotes the pushforward, that is, another notation for the differential of F.

$\overline{F}_{\overline{\mathsf{X}},T}, \overline{F}_{\overline{\mathsf{X}}\boldsymbol{Q},T} \in F_{\mathsf{X},T}, \overline{F}_{\overline{\mathsf{X}},T} \in F_{\mathsf{X},T}$.

$$
\begin{aligned}
\overline{\mathsf{X}}\left(t, \overline{F}_{\overline{\mathsf{X}}\boldsymbol{Q},t}\left(\boldsymbol{Y}\boldsymbol{Q}\right)\right) &= \overline{\mathsf{X}}\left(t, \overline{F}_{\overline{\mathsf{X}},t}\left(\boldsymbol{Y}\right)\boldsymbol{Q}\right) && (13)\\
&= \frac{d}{dt}\left\{\overline{F}_{\overline{\mathsf{X}},t}\left(\boldsymbol{Y}\right)\boldsymbol{Q}\right\} && (14)\\
&= \frac{d}{dt}\left(\theta \circ \overline{F}_{\overline{\mathsf{X}},t}\right)\left(\boldsymbol{Y}\right) && (15)\\
&= d(\theta)_{\boldsymbol{Y}}\left\{\frac{d}{dt}\overline{F}_{\overline{\mathsf{X}},t}\left(\boldsymbol{Y}\right)\right\} && (16)\\
&= d(\theta)_{\boldsymbol{Y}}\left\{\overline{\mathsf{X}}\left(t, \overline{F}_{\overline{\mathsf{X}},t}\left(\boldsymbol{Y}\right)\right)\right\} && (17)\\
&= \overline{\mathsf{X}}\left(t, \overline{F}_{\overline{\mathsf{X}},t}\left(\boldsymbol{Y}\right)\right)\boldsymbol{Q}. && (18)
\end{aligned}
$$

Thus, $\overline{\mathsf{X}} \sim \overline{\mathsf{X}}\boldsymbol{Q}$ is true. Therefore, $\overline{\mathsf{X}}$ is the horizontal lift of the vector field X on $\mathrm{Gr}(k, D)$ and is unique for X.

Flow $\boldsymbol{F_{\mathsf{X},T}}$ on $\mathbf{Gr}(\boldsymbol{k}, \boldsymbol{D})$ $\Leftarrow$ Vector Field X on $\mathbf{Gr}(\boldsymbol{k}, \boldsymbol{D})$. Let $\theta : \mathrm{St}(k, D) \times \mathcal{O}(k) \to \mathrm{St}(k, D), (\boldsymbol{Y}, \boldsymbol{Q}) \mapsto \boldsymbol{Y}\boldsymbol{Q}$ be a map representing the right action of the orthogonal group. In addition, let $\overline{\mathsf{X}}$ be a vector field over a horizontal bundle $T^{\mathrm{h}}\,\mathrm{St}(k, D)$ on $\mathrm{St}(k, D)$ and $\overline{F}_{\overline{\mathsf{X}},T}$ be its flow. From the Corollary 1,

$$
\begin{aligned}
\overline{F}_{\theta_* \circ \overline{\mathsf{X}},T} &= \theta \circ \overline{F}_{\overline{\mathsf{X}},T} \circ \theta^{-1} && (19)\\
\overline{F}_{\theta_* \circ \overline{\mathsf{X}},T} \circ \theta &= \theta \circ \overline{F}_{\overline{\mathsf{X}},T} && (20)\\
\overline{F}_{d(\theta)_{\boldsymbol{Y}}\overline{\mathsf{X}},T} \circ \theta &= \theta \circ \overline{F}_{\overline{\mathsf{X}},T} && (21)\\
\overline{F}_{\overline{\mathsf{X}}\boldsymbol{Q},T}\left(\boldsymbol{Y}\boldsymbol{Q}\right) &= \overline{F}_{\overline{\mathsf{X}},T}\left(\boldsymbol{Y}\right)\boldsymbol{Q}. && (22)
\end{aligned}
$$

Note that $d(\theta)_{\boldsymbol{Y}}\overline{\mathsf{X}} = \overline{\mathsf{X}}\boldsymbol{Q}$ is derived from the Corollary 2 and (4) in Zhu & Sato (2021). This indicates that $\overline{F}_{\overline{\mathsf{X}},T} \sim \overline{F}_{\overline{\mathsf{X}}',T}$ is true for any $\overline{\mathsf{X}}, \overline{\mathsf{X}}' \in T^{\mathrm{h}}\,\mathrm{St}(k, D)$ that satisfies $\overline{\mathsf{X}} \sim \overline{\mathsf{X}}'$. Thus, a new flow can be defined as $F_{\mathsf{X},T} := \left[\overline{F}_{\overline{\mathsf{X}},T}\right]$. This is a flow on a $\mathrm{Gr}(k, D)$. Because $\overline{\mathsf{X}}$ is a vector field in a horizontal bundle $T^{\mathrm{h}}\,\mathrm{St}(k, D)$ on $\mathrm{St}(k, D)$, it is a horizontal lift of the vector field X on $\mathrm{Gr}(k, D)$ and is therefore unique for X.

Thus, the proof is complete. □

B.2 Proposition 2

Proposition 2. *Let $\mathrm{Gr}(k, D)$ be a Grassmann manifold. Let p be the probability density on $\mathrm{Gr}(k, D)$ and F be the flow on $\mathrm{Gr}(k, D)$. Suppose $\overline{p}$ is a density on $\mathrm{St}(k, D)$ and $\overline{F}$ is a flow on $\mathrm{St}(k, D)$. Then, the distribution $\overline{p}_{\overline{F}}$ after transformations by $\overline{F}$ is also a density on $\mathrm{St}(k, D)$. Further, the invariance condition $\overline{p}_{\overline{F}} \sim \overline{p}_{\overline{F}'}$ is satisfied for all $\overline{F} \sim \overline{F}'$. Therefore, $p_F := [\overline{p}_{\overline{F}}]$ is a distribution on $\mathrm{Gr}(k, D)$.*

Proof. Let $\theta : \mathrm{St}(k, D) \times \mathcal{O}(k) \to \mathrm{St}(k, D) : (\boldsymbol{Y}, \boldsymbol{Q}) \mapsto \boldsymbol{YQ}$ be a map representing the right action of the orthogonal group.

$$\begin{aligned}
\overline{p}_F\left(\theta \circ \boldsymbol{Y}\right) &= \overline{p}_F\left(\theta \circ \boldsymbol{Y}\right) \frac{\left|\det\left\{J_\theta\left(\boldsymbol{Y}\right)\right\}\right|}{\left|\det\left\{J_\theta\left(\boldsymbol{Y}\right)\right\}\right|} = \frac{\overline{p}_{\theta^{-1}\circ F}\left(\boldsymbol{Y}\right)}{\left|\det\left\{J_\theta\left(\boldsymbol{Y}\right)\right\}\right|} && (23)\\
&= \overline{p}\left(\left(F^{-1} \circ \theta\right)\left(\boldsymbol{Y}\right)\right) \frac{\left|\det\left\{J_{F^{-1}\circ\theta}\left(\boldsymbol{Y}\right)\right\}\right|}{\left|\det\left\{J_\theta\left(\boldsymbol{Y}\right)\right\}\right|} && (24)\\
&= \left(\overline{p} \circ F^{-1}\right) \circ \theta\left(\boldsymbol{Y}\right) \frac{\left|\det\left\{J_{\theta\circ F^{-1}}\left(\boldsymbol{Y}\right)\right\}\right|}{\left|\det\left\{J_\theta\left(\boldsymbol{Y}\right)\right\}\right|} && (25)\\
&= \theta \circ \left(\overline{p} \circ F^{-1}\right)\left(\boldsymbol{Y}\right) \frac{\left|\det\left\{J_\theta\left(F^{-1}\left(\boldsymbol{Y}\right)\right) J_{F^{-1}}\left(\boldsymbol{Y}\right)\right\}\right|}{\left|\det\left\{J_\theta\left(\boldsymbol{Y}\right)\right\}\right|} && (26)\\
&= \theta \circ \left(\overline{p} \circ F^{-1}\right)\left(\boldsymbol{Y}\right) \frac{\left|\det\left\{J_\theta\left(F^{-1}\left(\boldsymbol{Y}\right)\right)\right\}\right| \left|\det\left\{J_{F^{-1}}\left(\boldsymbol{Y}\right)\right\}\right|}{\left|\det\left\{J_\theta\left(\boldsymbol{Y}\right)\right\}\right|} && (27)\\
&= \theta \circ \overline{p}\left(F^{-1}\right)\left(\boldsymbol{Y}\right) \left|\det\left\{J_{F^{-1}}\left(\boldsymbol{Y}\right)\right\}\right| \frac{\left|\det\left\{J_\theta\left(F^{-1}\left(\boldsymbol{Y}\right)\right)\right\}\right|}{\left|\det\left\{J_\theta\left(\boldsymbol{Y}\right)\right\}\right|} && (28)\\
&= \theta \circ \overline{p}_F\left(\boldsymbol{Y}\right) \frac{\left|\det\left\{J_\theta\left(F^{-1}\left(\boldsymbol{Y}\right)\right)\right\}\right|}{\left|\det\left\{J_\theta\left(\boldsymbol{Y}\right)\right\}\right|} && (29)\\
&= \theta \circ \overline{p}_F\left(\boldsymbol{Y}\right), && (30)
\end{aligned}$$

where $\left|\det\left\{J_\theta\left(\boldsymbol{X}\right)\right\}\right| = 1$ is true because θ is the action of the orthogonal group. Therefore, as $\overline{p}_F\left(\theta \circ \boldsymbol{Y}\right) = \theta \circ \overline{p}_F\left(\boldsymbol{Y}\right)$ is true, $\overline{p}_F \sim \overline{p}_F \boldsymbol{Q} \in p_F$ is true. Based on this, it can be concluded that the subject is satisfied. □

B.3 Proposition 3

Proposition 3. *The distribution $p_{\mathrm{Gr}(k,D)}$ on a Grassmann manifold $\mathrm{Gr}(k, D)$ based on the matrix-variate Gaussian distribution $\mathcal{MN}$ can be expressed as follows.*

$$p_{\mathrm{Gr}(k,D)}\left([\boldsymbol{X}]; [\boldsymbol{M}], \boldsymbol{U}, \boldsymbol{V}\right) = V_{\mathrm{Gr}(k,D)} \mathcal{MN}\left(\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}}; \boldsymbol{0}, \boldsymbol{U}, \boldsymbol{V}\right) \left|\det\left(\nabla_{\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}}} \overline{R}_{\boldsymbol{M}}\right)\right|, \tag{31}$$

where $\boldsymbol{M}$ is an orthonormal basis matrix denoting the mean of the distribution, $\boldsymbol{U}$ is a positive definite matrix denoting the row directional variance, $\boldsymbol{V}$ is a positive definite matrix denoting the column directional variance, and $\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}}$ is a random sample from $\mathcal{MN}$ in an $(D-k) \times k$-dimensional horizontal space $T^{\mathrm{h}}_{\boldsymbol{M}} \mathrm{St}(k, D)$. $V_{\mathrm{Gr}(k,D)}$ denotes the total volume of $\mathrm{Gr}(k, D)$ defined by (113), $\overline{R}_{\boldsymbol{M}}$ denotes the horizontal retraction at $\boldsymbol{M}$, and $\left|\det\left(\nabla_{\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}}} \overline{R}_{\boldsymbol{M}}\right)\right|$ denotes the Jacobian.

Proof. Let $p_{\mathrm{Gr}}\left([\boldsymbol{X}]\right)$ be a probability density function on a $\mathrm{Gr}(k, D)$. From (101), let $(d\boldsymbol{X})$ be the invariant measure on $\mathrm{Gr}(k, D)$ and $d\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}}$ be the Lebesgue measure on $T^{\mathrm{h}}_{\boldsymbol{M}} \mathrm{St}(k, D)$. Subsequently, a change of variables was performed according to the following:

$$\begin{aligned}
p_{\mathrm{Gr}(k,D)}\left([\boldsymbol{X}]\right)(d\boldsymbol{X}) &= p_{\mathrm{Gr}(k,D)}\left(\left[\overline{R}_{\boldsymbol{M}}\left(\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}}\right)\right]\right) d\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}} && (32)\\
p_{\mathrm{Gr}(k,D)}\left([\boldsymbol{X}]\right) &= p_{\mathrm{Gr}(k,D)}\left(\left[\overline{R}_{\boldsymbol{M}}\left(\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}}\right)\right]\right) \left|\det\left(\frac{d\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}}}{d\overline{R}_{\boldsymbol{M}}}\right)\right| && (33)\\
p_{\mathrm{Gr}(k,D)}\left([\boldsymbol{X}]\right) &= p_{\mathrm{Gr}(k,D)}\left(\left[\overline{R}_{\boldsymbol{M}}\left(\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}}\right)\right]\right) \left|\det\left(\nabla_{\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}}} \overline{R}_{\boldsymbol{M}}\right)\right|^{-1}. && (34)
\end{aligned}$$

Suppose $p_{\mathrm{Gr}}\left([\boldsymbol{X}]\right)$ is integrable with the probability measure $[d\boldsymbol{X}]$ on $\mathrm{Gr}(k, D)$ defined by (114). Then, we obtain the following relation.

$$
\begin{aligned}
&\int_{\mathrm{Gr}(k,D)} p_{\mathrm{Gr}(k,D)}\left([\boldsymbol{X}]\right)[d\boldsymbol{X}] && (35)\\
&\qquad= \frac{1}{V_{\mathrm{Gr}(k,D)}} \int_{\mathrm{Gr}(k,D)} p_{\mathrm{Gr}(k,D)}\left([\boldsymbol{X}]\right)(d\boldsymbol{X}) && (36)\\
&\qquad= \frac{1}{V_{\mathrm{Gr}(k,D)}} \int_{T^{\mathrm{h}}_{\boldsymbol{M}}\,\mathrm{St}(k,D)} p_{\mathrm{Gr}(k,D)}\left(\left[\overline{R}_{\boldsymbol{M}}\left(\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}}\right)\right]\right)\left|\det\left(\nabla_{\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}}}\overline{R}_{\boldsymbol{M}}\right)\right|^{-1} d\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}}. && (37)
\end{aligned}
$$

In addition, we obtain the following equation based on $[d\boldsymbol{X}]$.

$$
\int_{\mathrm{Gr}(k,D)} p_{\mathrm{Gr}(k,D)}\left([\boldsymbol{X}]\right)[d\boldsymbol{X}] = \int_{T^{\mathrm{h}}_{\boldsymbol{M}}\,\mathrm{St}(k,D)} \mathcal{MN}\left(\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}}\right) d\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}} = 1, \tag{38}
$$

where $\mathcal{MN}\left(\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}}\right)$ denotes the matrix-variate Gaussian distribution (Mathai et al. (2022)). Thus, the probability density function on $\mathrm{Gr}(k, D)$ can be expressed as follows:

$$
p_{\mathrm{Gr}(k,D)}\left(\left[\overline{R}_{\boldsymbol{M}}\left(\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}}\right)\right]\right) = V_{\mathrm{Gr}(k,D)}\mathcal{MN}\left(\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}}\right)\left|\det\left(\nabla_{\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}}}\overline{R}_{\boldsymbol{M}}\right)\right|. \tag{39}
$$

The Jacobian can be represented as $\left|\det\left(\nabla_{\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}}}\overline{R}_{\boldsymbol{M}}\right)\right| = \left|\det\left\{\left(\partial\overline{R}_{\boldsymbol{M}}\right)^{\top}\left(\partial\overline{R}_{\boldsymbol{M}}\right)\right\}\right|^{\frac{1}{2}}$ from Evans & Ronald (2015), where $\partial\overline{R}_{\boldsymbol{M}} = \frac{\partial\overline{R}_{\boldsymbol{M}}}{\partial\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}}}$. Further, $\partial\overline{R}_{\boldsymbol{M}}$ can be computed as follows. First, we define the horizontal retraction $\overline{R}_{\boldsymbol{Y}} : T^{\mathrm{h}}_{\boldsymbol{Y}}\,\mathrm{St}(k, D) \to \mathrm{St}(k, D)$ based on the Cayley transform from Zhu & Sato (2021).

$$
\boldsymbol{X} = \overline{R}_{\boldsymbol{M}}\left(\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}}\right) = \boldsymbol{M} + \overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}} - \left(\frac{1}{2}\boldsymbol{M} + \frac{1}{4}\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}}\right)\left(\boldsymbol{I}_k + \frac{1}{4}\overline{\boldsymbol{\xi}}^{\mathrm{h}\ \top}_{\boldsymbol{M}}\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}}\right)^{-1}\overline{\boldsymbol{\xi}}^{\mathrm{h}\ \top}_{\boldsymbol{M}}\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}}. \tag{40}
$$

This is a fixed time ($t = 1$) version of (122). Next, for improved visibility in subsequent calculations, let $\boldsymbol{E} = \overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}},\ \boldsymbol{F} = \frac{1}{2}\boldsymbol{M} + \frac{1}{4}\boldsymbol{E},\ \boldsymbol{G} = \left(\boldsymbol{I}_k + \frac{1}{4}\boldsymbol{H}\right)^{-1},\ \boldsymbol{H} = \boldsymbol{E}^{\top}\boldsymbol{E}$. In addition, let D be defined as the operator for the derivative of a matrix by a matrix. Then, the derivative $\nabla_{\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}}}\overline{R}_{\boldsymbol{M}} = \mathsf{D}\boldsymbol{X}$ by $\boldsymbol{E}$ is as follows:

$$
\nabla_{\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{M}}}\overline{R}_{\boldsymbol{M}} = \mathsf{D}\boldsymbol{X} = \mathsf{D}\boldsymbol{M} + \mathsf{D}\boldsymbol{E} - \mathsf{D}\left(\boldsymbol{F}\boldsymbol{G}\boldsymbol{H}\right). \tag{41}
$$

Finally, each derivative can be calculated as follows:

$$
\begin{aligned}
\mathsf{D}\boldsymbol{M} =& \boldsymbol{0}, && (42)\\
\mathsf{D}\boldsymbol{E} =& \boldsymbol{I}_{Dk}, && (43)\\
\mathsf{D}\left(\boldsymbol{F}\boldsymbol{G}\boldsymbol{H}\right) =& \mathsf{D}\left(\boldsymbol{F}\left(\boldsymbol{G}\boldsymbol{H}\right)\right) && (44)\\
=& \left\{\left(\boldsymbol{G}\boldsymbol{H}\right)^{\top} \otimes \boldsymbol{I}_D\right\}\mathsf{D}\boldsymbol{F} + \left(\boldsymbol{I}_k \otimes \boldsymbol{F}\right)\mathsf{D}\left(\boldsymbol{G}\boldsymbol{H}\right) && (45)\\
=& \left\{\left(\boldsymbol{G}\boldsymbol{H}\right)^{\top} \otimes \boldsymbol{I}_D\right\}\mathsf{D}\boldsymbol{F} + \left(\boldsymbol{I}_k \otimes \boldsymbol{F}\right)\left\{\left(\boldsymbol{H}^{\top} \otimes \boldsymbol{I}_k\right)\mathsf{D}\boldsymbol{G} + \left(\boldsymbol{I}_k \otimes \boldsymbol{G}\right)\mathsf{D}\boldsymbol{H}\right\} && (46)\\
=& \left\{\left(\boldsymbol{G}\boldsymbol{H}\right)^{\top} \otimes \boldsymbol{I}_D\right\}\mathsf{D}\boldsymbol{F}\\
&+ \left(\boldsymbol{I}_k \otimes \boldsymbol{F}\right)\left(\boldsymbol{H}^{\top} \otimes \boldsymbol{I}_k\right)\mathsf{D}\boldsymbol{G} + \left(\boldsymbol{I}_k \otimes \boldsymbol{F}\right)\left(\boldsymbol{I}_k \otimes \boldsymbol{G}\right)\mathsf{D}\boldsymbol{H} && (47)\\
=& \left(\boldsymbol{G}^{\top}\boldsymbol{H}^{\top} \otimes \boldsymbol{I}_D\right)\mathsf{D}\boldsymbol{F} + \left(\boldsymbol{H}^{\top} \otimes \boldsymbol{F}\right)\mathsf{D}\boldsymbol{G} + \left(\boldsymbol{I}_k \otimes \boldsymbol{F}\boldsymbol{G}\right)\mathsf{D}\boldsymbol{H}, && (48)\\
\mathsf{D}\boldsymbol{F} =& \mathsf{D}\left(\frac{1}{2}\boldsymbol{M}\right) + \mathsf{D}\left(\frac{1}{4}\boldsymbol{E}\right) = \frac{1}{4}\mathsf{D}\boldsymbol{E}, && (49)\\
\mathsf{D}\boldsymbol{G} =& -\left(\boldsymbol{G}^{\top} \otimes \boldsymbol{G}\right)\mathsf{D}\boldsymbol{H}, && (50)\\
\mathsf{D}\boldsymbol{H} =& \left(\boldsymbol{I}_{k^2} + \boldsymbol{K}_{k,k}\right)\left(\boldsymbol{I}_k \otimes \boldsymbol{E}^{\top}\right)\mathsf{D}\boldsymbol{E}, && (51)
\end{aligned}
$$

where $\otimes$ denotes the Kronecker product. $\boldsymbol{K}_{D,k}$ is a $Dk \times Dk$ matrix $\boldsymbol{K}_{D,k} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\boldsymbol{L}_{i,j} \otimes \boldsymbol{L}_{i,j}^{\top}\right)$ referred to as the commutation matrix, which denotes the transposition operation of $D \times k$. Further, $\boldsymbol{L}_{i,j}$ is a $D \times k$ matrix whose (i, j) component is 1 whereas all other components are 0. □

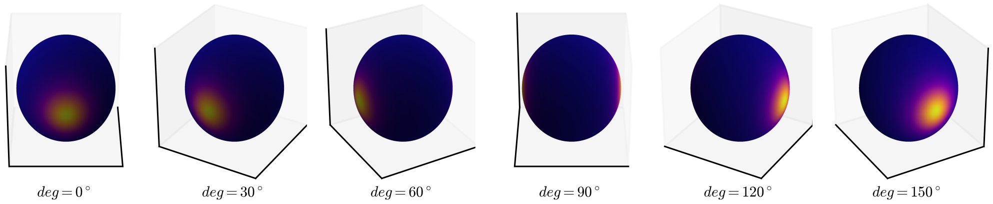


Figure 6: $p_{\mathrm{Gr}(1,3)}\left([\boldsymbol{X}]\right)$ with $\boldsymbol{M} = \left(1.0, 0.0, 0.0\right)^\top, \boldsymbol{U} = \sigma^2 \boldsymbol{I}_3, \boldsymbol{V} = \boldsymbol{I}_1, \sigma^2 = 0.5$. Each sphere in the figure indicates $\mathrm{Gr}(1,3)$, with brighter spheres representing higher densities.

For details on the formulae for matrix derivatives used in this proof, please refer to Magnus & Neudecker (2019).

$p_{\mathrm{Gr}(k,D)}\left([\boldsymbol{X}]\right) = p_{\mathrm{Gr}(k,D)}\left([\boldsymbol{X}]; [\boldsymbol{M}], \boldsymbol{U}, \boldsymbol{V}\right)$ is a probability distribution following mean $[\boldsymbol{M}]$ and matrix variance $\boldsymbol{U}, \boldsymbol{V}$. Using $\mathrm{Gr}(1,3)$ as an example, we qualitatively confirmed through visualization that $p_{\mathrm{Gr}(1,3)}\left([\boldsymbol{X}]\right)$ is a density on $\mathrm{Gr}(1,3)$. $\mathrm{Gr}(1,3)$ is a 1-dimensional subspace in a 3-dimensional space; that is, a space whose elements are lines passing through the origin in 3-dimensional space. For the visualization, we expressed $\mathrm{Gr}(1,3)$ by mapping a 1-dimensional subspace to two points on a sphere (one point on the sphere and its antipodal point) of radius 1 centered at the origin.

Figure 6 shows the density of $p_{\mathrm{Gr}(1,3)}\left([\boldsymbol{X}]\right)$ with $\boldsymbol{M} = \left(1.0, 0.0, 0.0\right)^\top, \boldsymbol{U} = \sigma^2 \boldsymbol{I}_3, \boldsymbol{V} = \boldsymbol{I}_1, \sigma^2 = 0.5$. Each sphere in the figure indicates $\mathrm{Gr}(1,3)$, with brighter spheres representing higher densities. The leftmost figure shows $\boldsymbol{M}$ as viewed from the front diagonally above, and the other figures present the views when the viewpoint is rotated clockwise around the z-axis by 30° to 150° with movement to the right. In the leftmost figure, the density is highly spread around $\boldsymbol{M}$. In the other figures (particularly the rightmost one), the antipodal point ($-\boldsymbol{M} = \left(-1.0, 0.0, 0.0\right)^\top$) is densely spread out. This implies that when only one $\boldsymbol{M}$ is specified as the representative of the equivalence class $[\boldsymbol{M}]$, the density around the other elements in the equivalence class $[\boldsymbol{M}]$ is as high as that around the representative. Thus, we can confirm that $p_{\mathrm{Gr}(k,D)}\left([\boldsymbol{X}]\right)$ has a density of $\mathrm{Gr}(1,3)$.

C Experimental Details for Learning GrCNF

C.1 Details on ODE Solver with Orthogonal Integration

We explain in detail an ordinary differential equation (ODE) solver on $\mathrm{Gr}(k, D)$. Several studies on ODE solvers employed on a manifold $\mathcal{M}$ have been reported (Munthe-Kaas (1999); Iserles et al. (2000); Hairer (2006)). Hairer (2006) proposed a simple projection method that projected onto the manifold at each step and the symmetric projection method suitable for long-time integration. However, this method requires $\mathcal{M}$ to be a submanifold in Euclidean space, and thus cannot be applied on $\mathrm{Gr}(k, D)$. In contrast, Celledoni & Owren (2002) proposed an intrinsic ODE solver that was applicable to a Stiefel manifold and did not assume an outer Euclidean space. This solver works on a Stiefel manifold, thus it must be reformulated into a solver suitable for $\mathrm{Gr}(k, D)$, which is our problem setting. We introduce below a solver on $\mathrm{Gr}(k, D)$ via ODE operating on the horizontal space $T^{\mathrm{h}}_{\boldsymbol{Y}} \mathrm{St}(k, D)$, based on results from Celledoni & Owren (2002) and Section 4. This is an intrinsic approach regardless of whether $\mathcal{M}$ has been embedded in a larger space with a corresponding extension of the vector field.

First, consider an ODE expressed using a vector field X_θ on a curve $\gamma(t) : [0, \infty) \to \mathrm{St}(k, D)$ such that for each time t.

$$\frac{d\gamma(t)}{dt} = \mathsf{X}_\theta(t, \gamma(t)), \quad \gamma(0) = \boldsymbol{Y}, \tag{52}$$

where X_θ is constructed by a neural network with parameter θ, as described in Section 5.3. Horizontal retraction $\overline{R}_{\boldsymbol{Y}}$ which is defined as (122) and is described in Appendix D.6, defines local coordinates of $\mathrm{Gr}(k, D)$ in a neighborhood of the point $[\boldsymbol{Y}]$. Thus, the solution of ODE can be expressed as:

$$\gamma(t) = \overline{R}_{\boldsymbol{Y}}(\epsilon(t)), \tag{53}$$

where $\epsilon : [0, \infty) \to T^{\mathrm{h}}_{\boldsymbol{Y}} \mathrm{St}(k, D)$ is the curve on $T^{\mathrm{h}}_{\boldsymbol{Y}} \mathrm{St}(k, D)$. By differentiating (53) with t, the following equation is obtained:

$$\frac{d\gamma(t)}{dt} = \frac{d}{dt}\overline{R}_{\boldsymbol{Y}}(\epsilon(t)) = \mathsf{X}_\theta(t, \gamma(t)). \tag{54}$$

Therefore, the ODE defined on $T^{\mathrm{h}}_{\boldsymbol{Y}} \mathrm{St}(k, D)$ is obtained as (8):

$$\begin{aligned} \frac{d\mathsf{Vec}(\epsilon(t))}{dt} &= \left(\nabla_\epsilon \overline{R}_{\boldsymbol{Y}}\right)^{-1} \mathsf{Vec}\left(\mathsf{X}_\theta\left(t, \overline{R}_{\boldsymbol{Y}}(\epsilon(t))\right)\right) \\ &= \nabla_\gamma \overline{R}_{\boldsymbol{Y}}^{-1} \mathsf{Vec}\left(\mathsf{X}_\theta\left(t, \overline{R}_{\boldsymbol{Y}}(\epsilon(t))\right)\right), \end{aligned}$$

where Vec denotes the map of vertically concatenating matrices and converting them into a single vector. $\nabla_\gamma \overline{R}_{\boldsymbol{Y}}^{-1}$ can be calculated using the derivative of (123):

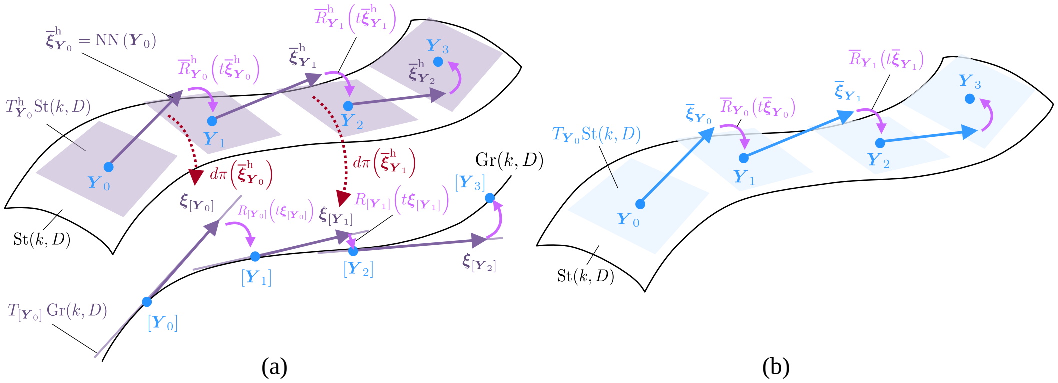
$$\nabla_\gamma \overline{R}_{\boldsymbol{Y}}^{-1} = 2\left(\boldsymbol{N}^{-\top} \otimes \boldsymbol{I}_D\right) \nabla \boldsymbol{M} + 2\left(\boldsymbol{I}_k \otimes \boldsymbol{M}\right) \nabla \boldsymbol{N}^{-1}, \tag{55}$$

where $\boldsymbol{X} = \gamma(t)$, $\boldsymbol{M} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{Y}$, $\boldsymbol{N} = \boldsymbol{I}_k + \boldsymbol{X}^\top \boldsymbol{Y}$, $\nabla \boldsymbol{M} = \boldsymbol{I}_{Dk} - \left(\boldsymbol{I}_k \otimes \boldsymbol{X}\boldsymbol{X}^\top\right)$ and $\nabla \boldsymbol{N}^{-1} = -\left(\boldsymbol{N}^{-\top} \otimes \boldsymbol{N}^{-1}\right)\left(\boldsymbol{I}_k \otimes \boldsymbol{X}^\top\right)$. Because (8) is an ODE on $T^{\mathrm{h}}_{\boldsymbol{Y}} \mathrm{St}(k, D) \cong \mathbb{R}^{(D-k)\times k}$ from Absil et al. (2008), it can be solved using an ODE solver such as Runge-Kutta methods that operate on Euclidean space. This study used Algorithm 5.1 presented in Celledoni & Owren (2002). In each step, first, (8) was solved using the Runge-Kutta method of order 5 as in Dormand & Prince (1980). Subsequently, the solution of ODE (52) was obtained by applying the solution ϵ to (53).

C.2 Loss Function based on Variational Inference

In the setting in Section 6.3, we used orthonormalized data $\boldsymbol{Y}$ as in $\boldsymbol{P}\boldsymbol{P}^\top \simeq \boldsymbol{Y}\boldsymbol{\Lambda}\boldsymbol{Y}^\top$, where $\boldsymbol{\Lambda}$ is diagonal (Huang et al. (2015)), such that the k-dimensional point cloud data $\boldsymbol{P}$ of N points $N \times k$ matrix is a matrix with k orthonormal basis vectors. Thus, generating a complete point cloud requires the estimation of the scale parameters $\sqrt{\boldsymbol{\Lambda}}$ to be $\boldsymbol{P} = \boldsymbol{Y}\sqrt{\boldsymbol{\Lambda}}$ and a loss function that incorporates this. In this study, we approximated by maximizing the evidence lower bound (ELBO), which is the lower bound of the overall log-likelihood $\log p_\psi(\boldsymbol{P})$ of $p_\psi(\boldsymbol{P})$, using a variational inference framework. The loss function is the variational energy $-\,\mathrm{ELBO}(\boldsymbol{P})$ with negative ELBO.

$$\mathrm{NLL} = -\log p_\psi(\boldsymbol{P}) \leq -\,\mathrm{ELBO}(\boldsymbol{P}) = \mathrm{Loss}. \tag{56}$$


$\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}_0} = \mathrm{NN}(\boldsymbol{Y}_0)$
$\overline{R}^{\mathrm{h}}_{\boldsymbol{Y}_1}\left(t\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}_1}\right)$
$\overline{R}^{\mathrm{h}}_{\boldsymbol{Y}_0}\left(t\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}_0}\right)$
$\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}_1}$
$\boldsymbol{Y}_3$
$\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}_2}$
$T^{\mathrm{h}}_{\boldsymbol{Y}_0}\mathrm{St}(k, D)$
$\boldsymbol{Y}_1$
$\boldsymbol{Y}_2$
$\boldsymbol{Y}_0$
$d\pi\left(\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}_0}\right)$
$d\pi\left(\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}_1}\right)$
$\mathrm{Gr}(k, D)$
$[\boldsymbol{Y}_3]$
$\boldsymbol{\xi}_{[\boldsymbol{Y}_0]}$
$R_{[\boldsymbol{Y}_0]}\left(t\boldsymbol{\xi}_{[\boldsymbol{Y}_0]}\right)$
$\boldsymbol{\xi}_{[\boldsymbol{Y}_1]}$
$R_{[\boldsymbol{Y}_1]}\left(t\boldsymbol{\xi}_{[\boldsymbol{Y}_1]}\right)$
$\mathrm{St}(k, D)$
$[\boldsymbol{Y}_1]$
$[\boldsymbol{Y}_2]$
$\boldsymbol{\xi}_{[\boldsymbol{Y}_2]}$
$T_{[\boldsymbol{Y}_0]}\,\mathrm{Gr}(k, D)$
$[\boldsymbol{Y}_0]$
$\overline{R}_{\boldsymbol{Y}_1}(t\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}_1})$
$\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}_1}$
$\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}_0}$
$\overline{R}_{\boldsymbol{Y}_0}(t\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}_0})$
$T_{\boldsymbol{Y}_0}\mathrm{St}(k, D)$
$\boldsymbol{Y}_3$
$\boldsymbol{Y}_1$
$\boldsymbol{Y}_2$
$\boldsymbol{Y}_0$
$\mathrm{St}(k, D)$

(a) (b)

Figure 7: The proposed ODE solver on the Grassmann manifold and the ODE solver on the Stiefel manifold (Celledoni & Owren (2002)). (a) The proposed ODE solver working on a Grassmann manifold. The $\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}}$ obtained by the proposed neural architecture NN is the horizontal lift (98), that is, $d\pi(\boldsymbol{Y})\left[\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}}\right] = \boldsymbol{\xi}_{[\boldsymbol{Y}]}$, of the tangent vector $\boldsymbol{\xi}_{[\boldsymbol{Y}]}$ at the point $[\boldsymbol{Y}]$ on the Grassmann manifold. The proposed ODE solver maps and updates the tangent vector $\boldsymbol{\xi}_{[\boldsymbol{Y}]}$ at the point $[\boldsymbol{Y}]$ onto the Grassmann manifold at each step using horizontal retraction (121). On the other hand, (b) the solver of Celledoni & Owren (2002), working on Stiefel manifolds, updates in each step by mapping the tangent vector $\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}$ at the point $\boldsymbol{Y}$ onto the Stiefel manifold. In other words, the difference between the proposed ODE solver and the ODE solver on Stiefel manifolds is that the ODE solver on Stiefel manifolds works only on Stiefel manifolds, while the proposed ODE solver always updates in each step with the Stiefel manifold and Grassmann manifolds linked together.

$\mathrm{ELBO}(\boldsymbol{P})$ can be decomposed as follows.

$$\begin{aligned} \mathrm{ELBO}(\boldsymbol{P}) &= \log p_\psi(\boldsymbol{P}) - D_{KL}\left(q_\phi(\boldsymbol{Y} \,|\, \boldsymbol{P}) || p_\psi(\boldsymbol{Y} \,|\, \boldsymbol{P})\right) && (57)\\ &= \mathbb{E}_{q_\phi(\boldsymbol{Y} \,|\, \boldsymbol{P})}\left[\log p_\psi(\boldsymbol{P} \,|\, \boldsymbol{Y})\right] - D_{KL}\left(q_\phi(\boldsymbol{Y} \,|\, \boldsymbol{P}) || p_\theta(\boldsymbol{Y})\right), && (58) \end{aligned}$$

where $q_\phi(\boldsymbol{Y} \,|\, \boldsymbol{P})$ is the inference model with parameter ϕ, $p_\psi(\boldsymbol{P} \,|\, \boldsymbol{Y})$ is the decoder model with parameter ψ, $p_\psi(\boldsymbol{Y} \,|\, \boldsymbol{P})$ is the posterior distribution with parameter ψ, and $p_\theta(\boldsymbol{Y})$ is the prior distribution with parameter θ. Further, $D_{KL}\left(q_\phi(\boldsymbol{Y} \,|\, \boldsymbol{P}) || p_\theta(\boldsymbol{Y})\right)$ can be formulated using differential entropy as follows.

$$D_{KL}\left(q_\phi(\boldsymbol{Y} \,|\, \boldsymbol{P}) || p_\theta(\boldsymbol{Y})\right) = -\mathbb{E}_{q_\phi(\boldsymbol{Y} \,|\, \boldsymbol{P})}\left[p_\theta(\boldsymbol{Y})\right] - H\left[q_\phi(\boldsymbol{Y} \,|\, \boldsymbol{P})\right]. \tag{59}$$

Thus, the final loss function is as follows.

$$\begin{aligned} \mathrm{Loss} &= -\,\mathrm{ELBO}(\boldsymbol{P}) && (60)\\ &= -\mathbb{E}_{q_\phi(\boldsymbol{Y} \,|\, \boldsymbol{P})}\left[\log p_\psi(\boldsymbol{P} \,|\, \boldsymbol{Y})\right] - \mathbb{E}_{q_\phi(\boldsymbol{Y} \,|\, \boldsymbol{P})}\left[p_\theta(\boldsymbol{Y})\right] - H\left[q_\phi(\boldsymbol{Y} \,|\, \boldsymbol{P})\right]. && (61) \end{aligned}$$

Each term of the loss function can be calculated as follows.

Expectation of log-likelihood $\mathbb{E}_{q_\phi(\boldsymbol{Y} \,|\, \boldsymbol{P})}\left[\log p_\psi(\boldsymbol{P} \,|\, \boldsymbol{Y})\right]$ is the reconstruction log-likelihood of $\boldsymbol{P}$. The expectation is estimated by Monte Carlo sampling.

Differential entropy In the decomposition of a point cloud $\boldsymbol{P}$, there exists arbitrariness in the choice of $\boldsymbol{Y}$ and $\boldsymbol{\Lambda}$, as in $\boldsymbol{P}\boldsymbol{P}^\top \simeq \boldsymbol{Y}\boldsymbol{\Lambda}\boldsymbol{Y}^\top$. In this study, we assumed that the diagonal components of $\boldsymbol{\Lambda}$ are in descending order, and we restricted the decomposition arbitrariness to be an action $\boldsymbol{Q} \in \mathcal{O}(k)$. If we suppose that the action $\boldsymbol{Q}$ follows a uniform distribution when $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{Q}$ holds, then $\boldsymbol{Y}$ also follows a uniform distribution in the k-dimensional subspace $\mathrm{span}(\boldsymbol{Y})$. Although this is a uniform distribution on $\mathrm{St}(k, k)$, we can consider a uniform distribution on $\mathcal{O}(k)$ because $\mathrm{St}(k, k) = \mathcal{O}(k)$. The probability density function of $\boldsymbol{Q}$ is represented by $U_{\mathcal{O}(k)}(\boldsymbol{Q})$ in (111).

Therefore, the differential entropy of the decoder model can be calculated as follows.

$$
\begin{aligned}
H\left[q_\phi(\boldsymbol{Y} \,|\, \boldsymbol{P})\right] &= \mathbb{E}\left[-\log q_\phi(\boldsymbol{Y} \,|\, \boldsymbol{P})\right] && (62)\\
&= -\int_{\mathcal{O}(k)} U_{\mathcal{O}(k)}\left(\boldsymbol{Q}\right) \log U_{\mathcal{O}(k)}\left(\boldsymbol{Q}\right) [d\boldsymbol{Q}] && (63)\\
&= -\int_{\mathcal{O}(k)} \frac{1}{V_{\mathcal{O}(k)}} \log \frac{1}{V_{\mathcal{O}(k)}} [d\boldsymbol{Q}] && (64)\\
&= \frac{\log V_{\mathcal{O}(k)}}{V_{\mathcal{O}(k)}} \int_{\mathcal{O}(k)} [d\boldsymbol{Q}] && (65)\\
&= \frac{\log V_{\mathcal{O}(k)}}{V_{\mathcal{O}(k)}}. && (66)
\end{aligned}
$$

Expectation of prior distribution We used (9) for the prior distribution $p_\theta(\boldsymbol{Y})$. Further, reparameterization was used to enable differentiable Monte Carlo estimation of expectations.

$$
\mathbb{E}_{q_\phi(\boldsymbol{Y} \,|\, \boldsymbol{P})}\left[p_\theta(\boldsymbol{Y})\right] = \frac{1}{L} \sum_{l=1}^{L} p_\theta(\boldsymbol{X}\boldsymbol{Q}_l) \quad \text{s.t.} \quad \boldsymbol{X} \sim \boldsymbol{Y},\ \boldsymbol{Q}_l \in \mathcal{O}(k)\,, \tag{67}
$$

where it was assumed that $\boldsymbol{Q}$ is sampled from a uniform distribution, which follows Haar measure on $\mathcal{O}(k)$. L is set $L = 1$.

C.3 Implementation Details and Experimental setting

The following sections present more details about the network architectures, training hyperparameters, and experimental conditions for each of the experiments in Section 6.

C.3.1 Artificial Textures

Network Architecture The vector field was constructed with the specific input, intermediate, and output layers described in Section 5.3. The GrCNF architecture is shown on top in Table 3. Layers are denoted as Layer in the table, and were processed from top to bottom. Norm. and Act. denote the normalization and activation functions to be applied immediately after the Layer, and the Norm. and Act. were applied in that order. Further, Out Size denotes the output size after Act. Vec denotes the map of vertically concatenating matrices and converting them into a single vector. Moreover, only row Input denotes the size of the input data, not the input layer (HorP). (9) was used for the loss function.

Hyper-parameters The mean $\boldsymbol{M}$ and covariances $\boldsymbol{U}$ and $\boldsymbol{V}$ in the prior distribution on the Grassmann manifold were set to $\boldsymbol{M} = (1.0, 0.0, 0.0)^\top$, $\boldsymbol{U} = \sigma^2 \boldsymbol{I}_3$, and $\boldsymbol{V} = \boldsymbol{I}_1$, $\sigma = 0.3$, respectively. Other hyperparameters used during the training of GrCNF are shown in Table 4.

Implementation We used PyTorch (Paszke et al. (2019)) to implement the model and run the experiments. The CNF is based on the implementation[5] in Chen et al. (2018) and the framework of the RCNF (Mathieu & Nickel (2020)). Thus, the ODE was solved using the explicit and adaptive Runge–Kutta method (Dormand & Prince (1980)) of order 5, and worked by projecting each step onto a manifold (Hairer (2006)). The autograd in (7) was calculated with torch.autograd.grad (Paszke et al. (2017)) in PyTorch. The experimental hardware was built with an Intel Core i7-9700 CPU and a single NVIDIA GTX 1060 GPU with 6 GB of RAM.

The code used in the experiment to generate the data distributions on $\mathrm{Gr}(1, 3)$ is shown in Listing 1. This implementation of the data distributions is based on the codes in Kim et al. (2020) and Grathwohl et al. (2019)[6].

[5] We used the authors' implementation: `https://github.com/rtqichen/torchdiffeq.git`.

[6] We used the authors' implementations: `https://github.com/ANLGBOY/SoftFlow.git` and `https://github.com/rtqichen/ffjord.git`.

Listing 1: Code for the data distributions.

```
import numpy as np

def get_data_batch(batch_size, dist):
    rng = np.random.RandomState()

    if dist == "2spirals":
        n = np.sqrt(np.random.rand(batch_size // 2, 1)) * 540 * (2 * np.pi) / 360
        d1x = -np.cos(n) * n + np.random.rand(batch_size // 2, 1) * 0.1
        d1y = np.sin(n) * n + np.random.rand(batch_size // 2, 1) * 0.1
        x = np.vstack((np.hstack((d1x, d1y)), np.hstack((-d1x, -d1y)))) / 3
        sample_2d = x + np.random.randn(*x.shape) * 0.1

    elif dist == "swissroll":
        data = sklearn.datasets.make_swiss_roll(n_samples=batch_size, noise=.3)[0]
        data = data.astype("float32")[:, [0, 2]]
        sample_2d = data / 5

    elif dist == "2circles":
        data = sklearn.datasets.make_circles(n_samples=batch_size, \
               factor=.5, noise=0.05)[0]
        data = data.astype("float32")
        sample_2d = data * 3

    elif dist == "2sines":
        x = (rng.rand(batch_size) - 0.5) * 2 * np.pi
        u = (rng.binomial(1, 0.5, batch_size) - 0.5) * 2
        y = u * np.sin(x) * 2.5
        x += np.random.randn(*x.shape) * 0.1
        y += np.random.randn(*y.shape) * 0.1
        sample_2d = np.stack((x, y), 1)

    elif dist == "target":
        shapes = np.random.randint(7, size=batch_size)
        mask = []
        for i in range(7):
            mask.append((shapes == i) * 1.)

        theta = np.linspace(0, 2 * np.pi, batch_size, endpoint=False)
        x = (mask[0] + mask[1] + mask[2]) * (rng.rand(batch_size) - 0.5) * 4 + \
            (-mask[3] + mask[4] * 0.0 + mask[5]) * 2 * np.ones(batch_size) + \
            mask[6] * np.cos(theta)
        y = (mask[3] + mask[4] + mask[5]) * (rng.rand(batch_size) - 0.5) * 4 + \
            (-mask[0] + mask[1] * 0.0 + mask[2]) * 2 * np.ones(batch_size) + \
            mask[6] * np.sin(theta)
        x += np.random.randn(*x.shape) * 0.1
        y += np.random.randn(*y.shape) * 0.1
        sample_2d = np.stack((x, y), 1)

    norm = sample_2d / np.max(np.linalg.norm(sample_2d, axis=1))
    sample_3d = np.concatenate((np.ones((batch_size, 1)), norm), axis=1)
    return sample_3d / np.linalg.norm(sample_3d, axis=1)[:, np.newaxis]
```

C.3.2 DW4 and LJ13

Network Architecture As in Appendix C.3.1, the vector field was constructed with the specific input, intermediate, and output layers described in Section 5.3. The GrCNF architecture is shown on the bottom left and right in Table 3. The bottom left and right were used for experiments on the DW4 and LJ13 datasets, respectively. The views presented in the table is the same as in Appendix C.3.1. (9) was used for the loss function. In addition, for architectures in methods other than GrCNF, please refer to Garcia Satorras et al. (2021).

Hyperparameters The mean $\boldsymbol{M}$ and covariances $\boldsymbol{U}$ and $\boldsymbol{V}$ in the prior distribution on the Grassmann manifold were set to $\boldsymbol{M} = \boldsymbol{I}_{4\times 2}$, $\boldsymbol{U} = \sigma^2 \boldsymbol{I}_4$, and $\boldsymbol{V} = \sigma^2 \boldsymbol{I}_2$, $\sigma = 0.3$ for DW4 and $\boldsymbol{M} = \boldsymbol{I}_{13\times 3}$, $\boldsymbol{U} = \sigma^2 \boldsymbol{I}_{13}$, and $\boldsymbol{V} = \sigma^2 \boldsymbol{I}_3$, $\sigma = 0.3$ for LJ13, respectively. Other hyperparameters used during the training of GrCNF are shown in Table 4. In addition, for the hyperparameters in methods other than GrCNF, please refer to Garcia Satorras et al. (2021).

Implementation The experimental hardware was built using a single NVIDIA Quadro RTX 8000 GPU with 48 GB of GDDR6 RAM. The other environments were the same as in Appendix C.3.1.

Table 3: Network architectures for each experiment; (left) for the Simple Texture dataset, (middle) for the DW4 dataset, (right) for the LJ13 dataset.

GrCNF for Textures

Layer	Out Size	Norm./Act.
Input	3×1	-
HorP	3×1	-/Tanh
Vec	3	-/-
CS	64	-/Tanh
CS	64	-/Tanh
CS	1	-/Tanh
Grad	3×1	-

GrCNF for DW4

Layer	Out Size	Norm./Act.
Input	4×2	-
HorP	4×2	-/Tanh
Vec	8	-/-
CS	64	-/Tanh
CS	64	-/Tanh
CS	64	-/Tanh
CS	1	-/Tanh
Grad	4×2	-

GrCNF for LJ13

Layer	Out Size	Norm./Act.
Input	13×3	-
HorP	13×3	-/Tanh
Vec	39	-/-
CS	32	-/Tanh
CS	32	-/Tanh
CS	32	-/Tanh
CS	1	-/Tanh
Grad	13×3	-

Full results In Tables 5 and 6, the same DW4 and LJ13 averaged results from Section 6.2 were reported; however, they included the standard deviations over the three runs.

C.3.3 QM9 Positional

Network Architecture On the QM9 Positional, we addressed the task of generating the molecular $\boldsymbol{P}$ by estimating the scale parameter $\sqrt{\boldsymbol{\Lambda}} = \operatorname{diag}\left(\left\{\sqrt{\lambda_i}\right\}_{i=1}^{3}\right)$, in addition to the generation of the orthonormal basis matrix $\boldsymbol{Y}$ with GrCNF. Because the molecular generation task requires a specialized loss function based on variational inference, we used (56), as explained in Appendix C.2. We designed two networks to achieve this. The first is the same GrCNF architecture as in previous experiments, and the second is a scale estimator. Table 7 shows the architectures. The left side of the table shows the GrCNF architecture and the right side shows the scale estimator. The scale estimator estimated one scale parameter from each of the three orthonormal basis vectors $\boldsymbol{Y} = \left\{\boldsymbol{y}_i \in \mathbb{R}^{19}\right\}_{i=1}^{3}$, for a total of three parameters $\left\{\sqrt{\lambda_i} \in \mathbb{R}\right\}_{i=1}^{3}$. With the orthonormal orthogonal basis matrix $\boldsymbol{Y}$ and the estimated scale parameter $\sqrt{\boldsymbol{\Lambda}}$, we generated a point cloud $\boldsymbol{P} = \boldsymbol{Y}\sqrt{\boldsymbol{\Lambda}}$. In this study, the overall architecture that generates $\boldsymbol{P}$ is also named GrCNF.

Hyperparameters The mean $\boldsymbol{M}$ and covariances $\boldsymbol{U}$ and $\boldsymbol{V}$ in the prior distribution on the Grassmann manifold were set to $\boldsymbol{M} = \boldsymbol{I}_{19\times 3}$, $\boldsymbol{U} = \sigma^2 \boldsymbol{I}_{19}$, and $\boldsymbol{V} = \sigma^2 \boldsymbol{I}_3$, $\sigma = 0.3$, respectively. In addition, for the hyperparameters in methods other than GrCNF, please refer to Garcia Satorras et al. (2021).

Implementation The experimental hardware was built using a single NVIDIA A100 GPU with 80GB PCIe of GDDR6 RAM.

Table 4: List of hyperparameters used in various experiments. A “-” indicates that the hyperparameter is unused.

		Textures	**DW4**	**LJ13**	**QM9**
# of Data	Train	∞	10^2/10^3/10^4/10^5	10/10^2/10^3/10^4	$13,831$
	Validation	500	1000	1000	2,501
	Test	500	1000	1000	1,813
Optimizer	Name	Adam	Adam	Adam	Adam
	beta1	0.9	0.9	0.9	0.9
	beta2	0.999	0.999	0.999	0.999
	Weight Decay	-	1.0e-12	1.0e-12	1.0e-12
	Learning Rate	1.0e-3	1.0e-4	1.0e-4	5.0e-4
Schedule	Epoch	72000	1000/300/50/6	500/1000/300/50	160
	LR Step with 0.1	20000	-	-	-
	Batch Size	500	100	10/100/100/100	128
NeuralODE	Integration Time	Training	Training	Training	0.1
	atol	1.0e-5	1.0e-5	1.0e-5	1.0e-5
	rtol	1.0e-5	1.0e-5	1.0e-5	1.0e-5
	Adjoint	✗	✓	✓	✓

Table 5: Negative log-likelihood comparison on the test partition of DW4 dataset for different amounts of training samples; averaged over 3 runs and including standard deviations.

		DW4		
# of Samples	10^2	10^3	10^4	10^5
GNF	-2.30 ± 1.59	-7.04 ± 0.64	-7.19 ± 0.99	-7.93 ± 1.10
GNF-att	-2.02 ± 1.34	-4.13 ± 1.20	-5.25 ± 0.89	-6.74 ± 0.89
GNF-att-aug	-3.11 ± 2.15	-4.04 ± 3.40	-6.51 ± 0.49	-9.42 ± 1.15
Simple dynamics	-1.22 ± 0.05	-1.28 ± 0.01	-1.36 ± 0.02	-1.39 ± 0.04
E-NF	-0.54 ± 0.45	-9.89 ± 2.30	-12.15 ± 1.16	-15.29 ± 0.53
GrCNF	**-12.53 ± 0.92**	**-13.74 ± 0.30**	**-14.09 ± 0.44**	**-16.07 ± 0.46**

Table 6: Negative log-likelihood comparison on the test partition of LJ13 dataset for different amounts of training samples; averaged over 3 runs and including standard deviations.

		LJ13		
# Samples	10	10^2	10^3	10^4
GNF	6.77 ± 0.39	-0.76 ± 1.12	-4.26 ± 2.76	-12.43 ± 1.21
GNF-att	6.91 ± 0.17	1.40 ± 0.79	-6.81 ± 2.09	-12.05 ± 2.28
GNF-att-aug	2.95 ± 0.55	-6.11 ± 1.12	-13.94 ± 0.95	-15.74 ± 0.58
Simple dynamics	-1.10 ± 2.55	-3.87 ± 0.25	-3.72 ± 0.08	-3.59 ± 0.52
E-NF	-12.86 ± 3.67	-15.75 ± 5.02	-31.51 ± 1.19	-32.83 ± 1.98
GrCNF	**-23.64 ± 2.23**	**-44.24 ± 4.26**	**-58.02 ± 5.43**	**-58.71 ± 4.71**

Table 7: Network architectures for QM9 Positional. (Left) GrCNF architecture, (Right) scale estimation architecture. The scale estimator estimates one scale parameter from each of the three orthonormal basis vectors $\boldsymbol{Y} = \left\{\boldsymbol{y}_i \in \mathbb{R}^{19}\right\}_{i=1}^{3}$, for $\sqrt{\boldsymbol{\Lambda}} = \mathrm{diag}\left(\left\{\sqrt{\lambda_i}\right\}_{i=1}^{3}\right)$ with three parameters $\left\{\sqrt{\lambda_i} \in \mathbb{R}\right\}_{i=1}^{3}$. Using the orthonormal orthogonal basis matrix $\boldsymbol{Y}$ and the estimated scale parameter $\sqrt{\boldsymbol{\Lambda}}$, we generate a point cloud $\boldsymbol{P} = \boldsymbol{Y}\sqrt{\boldsymbol{\Lambda}}$. SiLU is an activation function proposed in (Ramachandran et al. (2017)) and BatchNorm. is a batch normalization layer in (Ioffe & Szegedy (2015)).

GrCNF for QM9 Positional

Layer	Out Size	Norm./Act.
Input	19×3	-
HorP	19×3	-/Tanh
Vec	57	-/-
CS	32	-/Tanh
CS	32	-/Tanh
CS	32	-/Tanh
CS	1	-/Tanh
Grad	19×3	-

Scale Estimator for QM9 Positional

Layer	Out Size	Norm./Act.
Input	19×3	-
FC	128	BatchNorm./SiLU
FC	256	BatchNorm./SiLU
FC	256	BatchNorm./SiLU
FC	128	BatchNorm./SiLU
FC	3	-/ReLU

D Fundamentals of Concepts Associated with Grassmann Manifold

D.1 Definition of Stiefel Manifold

Definition 2. *An (orthogonal or compact) Stiefel manifold* $\mathrm{St}(k, D)$ *is defined as the set of orthonormal bases of* k*-dimensional subspaces in the Euclidean space* $\mathbb{R}^D$ *as in (68).*

$$\mathrm{St}(k, D) := \left\{ \boldsymbol{Y} \in \mathbb{R}^{D \times k} \;\middle|\; \boldsymbol{Y}^\top \boldsymbol{Y} = \boldsymbol{I}_k \right\}. \tag{68}$$

For $\boldsymbol{Y} \in \mathrm{St}(k, D)$, the space $\mathrm{span}(\boldsymbol{Y})$ spanned by its column vectors is the element of $\mathrm{Gr}(k, D)$.

$$f : \mathrm{St}(k, D) \to \mathrm{Gr}(k, D) : \boldsymbol{Y} \mapsto \mathrm{span}(\boldsymbol{Y}). \tag{69}$$

$\mathrm{St}(k, D)$ is a $Dk - \frac{k(k+1)}{2}$-dimensional compact manifold (Absil et al. (2008)).

D.2 Equivalence Relation

To define the equivalence relation $\sim$ [7] on a Stiefel manifold $\mathrm{St}(k, D)$, we introduce the following two lemmas.

Lemma 1. *The necessary and sufficient conditions for* $\mathrm{span}(\boldsymbol{Y}_1) = \mathrm{span}(\boldsymbol{Y}_2)$ *to hold for* $\boldsymbol{Y}_1, \boldsymbol{Y}_2 \in \mathrm{St}(k, D)$ *are as follows.*

$${}^{\exists}\boldsymbol{Q} \in \mathcal{O}(k) \quad \textit{s.t.} \quad \boldsymbol{Y}_2 = \boldsymbol{Y}_1 \boldsymbol{Q}. \tag{70}$$

Proof. $\mathrm{span}(\boldsymbol{Y}_1) = \mathrm{span}(\boldsymbol{Y}_2) \Leftarrow \boldsymbol{Y}_2 = \boldsymbol{Y}_1 \boldsymbol{Q}$. From the definition, $\boldsymbol{Y}_1^\top \boldsymbol{Y}_1 = \boldsymbol{Y}_2^\top \boldsymbol{Y}_2 = \boldsymbol{I}_k$, the following is obtained:

$$\boldsymbol{I}_k = \boldsymbol{Y}_1^\top \boldsymbol{Y}_1 = \boldsymbol{Q}^\top \boldsymbol{Y}_2^\top \boldsymbol{Y}_2 \boldsymbol{Q} = \boldsymbol{Q}^\top \boldsymbol{Q}. \tag{71}$$

Thus, there exists a k-dimensional orthogonal matrix $\boldsymbol{Q} \in \mathcal{O}(k)$. As the subspace $\mathrm{span}(\boldsymbol{Y})$ is invariant to coordinate transformations by orthogonal matrices, $\mathrm{span}(\boldsymbol{Y}_1) = \mathrm{span}(\boldsymbol{Y}_2)$ is true.

$\mathrm{span}(\boldsymbol{Y}_1) = \mathrm{span}(\boldsymbol{Y}_2) \Rightarrow \boldsymbol{Y}_2 = \boldsymbol{Y}_1 \boldsymbol{Q}$. From the assumption, we immediately concluded that $\boldsymbol{Y}_2 = \boldsymbol{Y}_1 \boldsymbol{Q}$ for $\boldsymbol{Q} \in \mathcal{O}(k)$. □

Lemma 2. *We define the equivalence relation* $\sim$ *on* $\mathrm{St}(k, D)$ *to be* $\boldsymbol{Y}_1 \sim \boldsymbol{Y}_2$ *whenever (70) is satisfied with respect to* $\boldsymbol{Y}_1, \boldsymbol{Y}_2 \in \mathrm{St}(k, D)$*. The fact that a binary relation is an equivalence relation* $\sim$ *implies that the following three statements hold for* ${}^{\forall}\boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{Y}_3 \in \mathrm{St}(k, D)$*.*

Reflexivity $\boldsymbol{Y}_1 \sim \boldsymbol{Y}_1$.

Symmetry $\boldsymbol{Y}_1 \sim \boldsymbol{Y}_2 \Rightarrow \boldsymbol{Y}_2 \sim \boldsymbol{Y}_1$.

Transitivity $\boldsymbol{Y}_1 \sim \boldsymbol{Y}_2 \wedge \boldsymbol{Y}_2 \sim \boldsymbol{Y}_3 \Rightarrow \boldsymbol{Y}_1 \sim \boldsymbol{Y}_3$.

Proof. With Lemma 1, we can confirm that it is valid as follows:

Reflexivity From $\boldsymbol{Y}_1 = \boldsymbol{Y}_1 \boldsymbol{I},\ \boldsymbol{I} \in \mathcal{O}$, we obtain $\boldsymbol{Y}_1 \sim \boldsymbol{Y}_1$.

Symmetry As $\boldsymbol{Y}_2 = \boldsymbol{Y}_1 \boldsymbol{Q}$ is obtained from $\boldsymbol{Y}_1 \sim \boldsymbol{Y}_2$, and $\boldsymbol{Y}_2 \boldsymbol{Q}^\top = \boldsymbol{Y}_1,\ \boldsymbol{Q}^\top \in \mathcal{O}$ is true, then $\boldsymbol{Y}_2 \sim \boldsymbol{Y}_1$ is obtained.

Transitivity $\boldsymbol{Y}_3 = \boldsymbol{Y}_1 \boldsymbol{Q}_1 \boldsymbol{Q}_2$ with $\boldsymbol{Y}_2 = \boldsymbol{Y}_1 \boldsymbol{Q}_1$ and $\boldsymbol{Y}_3 = \boldsymbol{Y}_2 \boldsymbol{Q}_2$. $\boldsymbol{Q}_1 \boldsymbol{Q}_2$ is $\left(\boldsymbol{Q}_1 \boldsymbol{Q}_2\right)^\top \left(\boldsymbol{Q}_1 \boldsymbol{Q}_2\right) = \boldsymbol{Q}_2^\top \boldsymbol{Q}_1^\top \boldsymbol{Q}_1 \boldsymbol{Q}_2 = \boldsymbol{Q}_2^\top \boldsymbol{Q}_2 = \boldsymbol{I}$. Moreover, as $\left(\boldsymbol{Q}_1 \boldsymbol{Q}_2\right) \left(\boldsymbol{Q}_1 \boldsymbol{Q}_2\right)^\top = \boldsymbol{Q}_1 \boldsymbol{Q}_2 \boldsymbol{Q}_2^\top \boldsymbol{Q}_1^\top = \boldsymbol{I}$ holds, $\boldsymbol{Q}_1 \boldsymbol{Q}_2 \in \mathcal{O}$ is true. Therefore, we concluded $\boldsymbol{Y}_1 \sim \boldsymbol{Y}_3$. □

The equivalence class of $\boldsymbol{Y} \in \mathrm{St}(k, D)$ is denoted by $[\boldsymbol{Y}]$. In other words, $[\boldsymbol{Y}]$ is the set of all elements of $\mathrm{St}(k, D)$ that are equivalent to $\boldsymbol{Y}$, and the equivalence relation on $\mathrm{St}(k, D)$ divides $\mathrm{St}(k, D)$ into equivalence classes with no intersection. Thus, $\boldsymbol{Y}$ is then referred to as the representative of the equivalence class $[\boldsymbol{Y}]$. The set of equivalence classes is denoted $\mathrm{St}(k, D)/\!\sim$ and is referred to as the

[7] Reflexive, symmetric and transitive binary relations. As a consequence of these properties, in a given set, one equivalence relation divides (classifies) the set into equivalence classes. Note that R is a binary relation in the set X if for any $x, y \in X$, only either x is related to y by the relation R, or x is not related to y based on the relation that R occurs. We write x is related to y by relation R" as xRy.

quotient of $\mathrm{St}(k, D)$ by the equivalence relation $\sim$. In addition, $\pi :\to \mathrm{St}(k, D) \to \mathrm{St}(k, D)/\sim$ is the natural projection that maps $\boldsymbol{Y} \in \mathrm{St}(k, D)$ onto its equivalence class $[\boldsymbol{Y}]$. This projection π is surjection.

D.3 Quotient Manifold

Definition 3. *Let $\overline{\mathcal{M}}$ be a manifold with equivalence relation $\sim$. The quotient space $\overline{\mathcal{M}}/\sim$ with $\sim$ of $\overline{\mathcal{M}}$ is the set of all equivalence classes. Thus, $\overline{\mathcal{M}}/\sim := \{\pi(\overline{\boldsymbol{x}}) \mid \overline{\boldsymbol{x}} \in \overline{\mathcal{M}}\}$, where $\pi : \overline{\mathcal{M}} \to \overline{\mathcal{M}}/\sim$ is the natural projection and $\pi(\overline{\boldsymbol{x}}) := \left\{\overline{\boldsymbol{y}} \in \overline{\mathcal{M}} \mid \overline{\boldsymbol{y}} \sim \overline{\boldsymbol{x}}\right\}$. Then, $\overline{\mathcal{M}}$ is referred to as the total space or the total manifold. Moreover, $\overline{\mathcal{M}}/\sim$ is referred to as a quotient manifold of $\overline{\mathcal{M}}$ if $\overline{\mathcal{M}}/\sim$ admits a differentiable structure.*

Let $\mathcal{M} = \overline{\mathcal{M}}/\sim$ be a quotient manifold. Further, suppose that $\overline{\mathcal{M}}$ is endowed with a Riemann metric $\overline{g}$, and let $\boldsymbol{x} = \pi(\boldsymbol{x})$. The horizontal space $T^{\mathrm{h}}_{\overline{\boldsymbol{x}}}\overline{\mathcal{M}}$ is the orthogonal complement of the vertical space $T^{\mathrm{v}}_{\overline{\boldsymbol{x}}}\overline{\mathcal{M}} := T_{\overline{\boldsymbol{x}}}\pi^{-1}(\boldsymbol{x})$ in the tangent space $T_{\overline{\boldsymbol{x}}}\overline{\mathcal{M}}$ and is defined as the follows:

$$T^{\mathrm{h}}_{\overline{\boldsymbol{x}}}\overline{\mathcal{M}} := \left(T^{\mathrm{v}}_{\overline{\boldsymbol{x}}}\overline{\mathcal{M}}\right)^{\perp} = \left\{\overline{\boldsymbol{\eta}}_{\overline{\boldsymbol{x}}} \in T_{\overline{\boldsymbol{x}}}\overline{\mathcal{M}} \mid \overline{g}_{\overline{\boldsymbol{x}}}\left(\overline{\boldsymbol{\xi}}_{\overline{\boldsymbol{x}}}, \overline{\boldsymbol{\eta}}_{\overline{\boldsymbol{x}}}\right) = 0, ^{\forall}\overline{\boldsymbol{\xi}}_{\overline{\boldsymbol{x}}} \in T^{\mathrm{v}}_{\overline{\boldsymbol{x}}}\overline{\mathcal{M}}\right\}. \tag{72}$$

The horizontal lift $\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\overline{\boldsymbol{x}}} \in T^{\mathrm{h}}_{\overline{\boldsymbol{x}}}\overline{\mathcal{M}}$ of the tangent vector $\boldsymbol{\xi}_{\boldsymbol{x}} \in T_{\boldsymbol{x}}\mathcal{M}$ at point $\overline{\boldsymbol{x}} \in \pi^{-1}(\boldsymbol{x})$ is a tangent vector that is uniquely determined as $d\pi_{\overline{\boldsymbol{x}}}\left(\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\overline{\boldsymbol{x}}}\right) = \boldsymbol{\xi}_{\boldsymbol{x}}$ (Absil et al. (2008)).

D.4 Grassmann Manifold Exploiting the Quotient Structure

D.4.1 Tangent Space on a Stiefel Manifold

We describe the relationship between tangent space $T_{[\boldsymbol{Y}]}\mathrm{Gr}(k, D)$ on $\mathrm{Gr}(k, D)$ and tangent space $T_{\boldsymbol{Y}}\mathrm{St}(k, D)$ on $\mathrm{St}(k, D)$ to relate the tangent vectors of a Grassmann manifold $\mathrm{Gr}(k, D)$ to the tangent vectors of a Stiefel manifold $\mathrm{St}(k, D)$ in a matrix representation. We take the derivative on both sides of $\boldsymbol{Y}(t)^{\top}\boldsymbol{Y}(t) = \boldsymbol{I}_p$ in (68) by t and solve for $t = 0$.

$$\frac{d}{dt}\left\{\boldsymbol{Y}(t)^{\top}\boldsymbol{Y}(t)\right\} = \frac{d}{dt}\boldsymbol{I}_p \tag{73}$$

$$\frac{d}{dt}\boldsymbol{Y}(t)^{\top}\boldsymbol{Y}(t) + \boldsymbol{Y}(t)^{\top}\frac{d}{dt}\boldsymbol{Y}(t) = 0 \tag{74}$$

$$\frac{d}{dt}\boldsymbol{Y}(0)^{\top}\boldsymbol{Y}(0) + \boldsymbol{Y}(0)^{\top}\frac{d}{dt}\boldsymbol{Y}(0) = 0 \tag{75}$$

$$\overline{\boldsymbol{\xi}}^{\top}_{\boldsymbol{Y}}\boldsymbol{Y} + \boldsymbol{Y}^{\top}\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}} = 0, \tag{76}$$

where $\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}} = \frac{d}{dt}\boldsymbol{Y}(0)$ is the tangent vector at $\boldsymbol{Y}$ [8].

Definition 4. *Define the tangent space $T_{\boldsymbol{Y}}\mathrm{St}(k, D)$ at $\boldsymbol{Y}$ on the Stiefel manifold as follows:*

$$T_{\boldsymbol{Y}}\mathrm{St}(k, D) = \left\{\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}} \in \mathbb{R}^{D\times k} \;\middle|\; \overline{\boldsymbol{\xi}}^{\top}_{\boldsymbol{Y}}\boldsymbol{Y} + \boldsymbol{Y}^{\top}\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}} = \boldsymbol{0}_k\right\}. \tag{77}$$

Let matrix $\boldsymbol{Y}_{\perp} \in \mathbb{R}^{D\times(D-k)}$ be a matrix satisfying the following:

$$\boldsymbol{Y}_{\perp}^{\top}\boldsymbol{Y}_{\perp} = \boldsymbol{I}_{D-k}, \quad \boldsymbol{Y}^{\top}\boldsymbol{Y}_{\perp} = 0, \quad \boldsymbol{Y}\boldsymbol{Y}^{\top} + \boldsymbol{Y}_{\perp}\boldsymbol{Y}_{\perp}^{\top} = \boldsymbol{I}_D. \tag{78}$$

As $\begin{bmatrix} \boldsymbol{Y} & \boldsymbol{Y}_{\perp} \end{bmatrix}$ is an orthogonal matrix [9], the column vectors of $\boldsymbol{Y}$ and $\boldsymbol{Y}_{\perp}$ form an orthonormal basis in $\mathbb{R}^D$. Thus, any $D \times k$ matrix can be written in terms of the $\boldsymbol{C} \in \mathbb{R}^{k\times k}$ and $\boldsymbol{B} \in \mathbb{R}^{(D-k)\times k}$ coefficient matrices as follows:

$$\boldsymbol{Y}\boldsymbol{C} + \boldsymbol{Y}_{\perp}\boldsymbol{B}, \tag{79}$$

[8] The tangent space is defined independently for each point of the manifold; hence, the subscript $\boldsymbol{Y}$, as in $\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}$, is clearly stated to emphasize that it is a tangent vector at $\boldsymbol{Y}$.

[9] $\begin{bmatrix} \boldsymbol{Y} & \boldsymbol{Y}_{\perp} \end{bmatrix}^{-1}\begin{bmatrix} \boldsymbol{Y} & \boldsymbol{Y}_{\perp} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y} & \boldsymbol{Y}_{\perp} \end{bmatrix}^{\top}\begin{bmatrix} \boldsymbol{Y} & \boldsymbol{Y}_{\perp} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y} & \boldsymbol{Y}_{\perp} \end{bmatrix}\begin{bmatrix} \boldsymbol{Y} & \boldsymbol{Y}_{\perp} \end{bmatrix}^{\top} = \boldsymbol{I}_D$.

where $\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}} = \boldsymbol{Y}\boldsymbol{C} + \boldsymbol{Y}_{\perp}\boldsymbol{B}$ is inserted. The following equation is obtained.

$$\begin{aligned}
\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}^{\top}\boldsymbol{Y} + \boldsymbol{Y}^{\top}\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}} &= \left(\boldsymbol{Y}\boldsymbol{C} + \boldsymbol{Y}_{\perp}\boldsymbol{B}\right)^{\top}\boldsymbol{Y} + \boldsymbol{Y}^{\top}\left(\boldsymbol{Y}\boldsymbol{C} + \boldsymbol{Y}_{\perp}\boldsymbol{B}\right) && (80)\\
&= \boldsymbol{B}^{\top}\boldsymbol{Y}_{\perp}^{\top}\boldsymbol{Y} + \boldsymbol{C}^{\top}\boldsymbol{Y}^{\top}\boldsymbol{Y} + \boldsymbol{Y}^{\top}\boldsymbol{Y}\boldsymbol{C} + \boldsymbol{Y}^{\top}\boldsymbol{Y}_{\perp}\boldsymbol{B} && (81)\\
&= \boldsymbol{B}^{\top}\boldsymbol{Y}^{\top}\boldsymbol{Y}_{\perp} + \boldsymbol{C}^{\top} + \boldsymbol{C} && (82)\\
&= \boldsymbol{C}^{\top} + \boldsymbol{C} && (83)\\
&= \boldsymbol{0}_k. && (84)
\end{aligned}$$

Thus, the following equation is derived.

$$\boldsymbol{C}^{\top} + \boldsymbol{C} = \boldsymbol{0}_k. \tag{85}$$

Thus, $\boldsymbol{C}$ is a $k \times k$ skew-symmetric matrix $\mathrm{Skew}\,(k)$. Therefore, we obtain the following as another representation of the tangent space on $\mathrm{St}(k, D)$.

$$T_{\boldsymbol{Y}}\mathrm{St}(k, D) = \left\{ \boldsymbol{Y}\boldsymbol{C} + \boldsymbol{Y}_{\perp}\boldsymbol{B} \;\middle|\; \boldsymbol{C} \in \mathrm{Skew}\,(k)\,, \boldsymbol{B} \in \mathbb{R}^{(D-k)\times k} \right\}. \tag{86}$$

D.4.2 Riemannian Metric on a Stiefel Manifold

In the tangent space $T_{\boldsymbol{x}}\mathcal{M}$ defined at each point $\boldsymbol{x} \in \mathcal{M}$ on the manifold $\mathcal{M}$, the inner product h is endowed as a bilinear map. This h is referred to as a Riemannian metric on a manifold, and the manifold $\mathcal{M}$ on which the Riemannian metric h is endowed is referred to as a Riemannian manifold $(\mathcal{M}, h)$. We define the Riemannian metric $\overline{g}$ on $\mathrm{St}(k, D)$ as follows.

$$\overline{g}_{\boldsymbol{Y}}\left(\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}, \overline{\boldsymbol{\eta}}_{\boldsymbol{Y}}\right) := \mathrm{tr}\left(\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}^{\top}\overline{\boldsymbol{\eta}}_{\boldsymbol{Y}}\right) \quad \text{s.t.} \quad \overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}, \overline{\boldsymbol{\eta}}_{\boldsymbol{Y}} \in T_{\boldsymbol{Y}}\mathrm{St}(k, D),\ \boldsymbol{Y} \in \mathrm{St}(k, D)\,. \tag{87}$$

This is the standard inner product of $\mathbb{R}^{D\times k}$ induced by $T_{\boldsymbol{Y}}\mathrm{St}(k, D)$, with $T_{\boldsymbol{Y}}\mathrm{St}(k, D) \subset \mathbb{R}^{D\times k}$ [10][11].

D.4.3 Tangent Space on a Grassmann Manifold

We describe the relation between tangent spaces $T_{[\boldsymbol{Y}]}\mathrm{Gr}(k, D)$ and $T_{\boldsymbol{Y}}\mathrm{St}(k, D)$ to relate tangent vectors in tangent spaces $T_{[\boldsymbol{Y}]}\mathrm{Gr}(k, D)$ on $\mathrm{Gr}(k, D)$ to tangent vectors $\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}} \in T_{\boldsymbol{Y}}\,\mathrm{St}(k, D)$.

First, we define the vertical space $T_{\boldsymbol{Y}}^{\mathrm{v}}\,\mathrm{St}(k, D)$ as a subspace of $T_{\boldsymbol{Y}}\mathrm{St}(k, D)$ as follows.

$$T_{\boldsymbol{Y}}^{\mathrm{v}}\,\mathrm{St}(k, D) := T_{\boldsymbol{Y}}\pi^{-1}\left([\boldsymbol{Y}]\right), \tag{88}$$

where $\pi : \mathrm{St}(k, D) \to \mathrm{Gr}(k, D)$ is the natural projection defined by $\pi\left(\boldsymbol{Y}\right) = [\boldsymbol{Y}]$ [12]. Thus, π converges all $\boldsymbol{Y}' \in \mathrm{St}(k, D)$ such that $\boldsymbol{Y} \sim \boldsymbol{Y}'$ to a point $[\boldsymbol{Y}]$ on $\mathrm{Gr}(k, D)$. Therefore, using (1), (88) can be transformed as follows.

$$T_{\boldsymbol{Y}}^{\mathrm{v}}\,\mathrm{St}(k, D) = T_{\boldsymbol{Y}}\left\{\boldsymbol{Y}\boldsymbol{Q} \mid \boldsymbol{Q} \in \mathcal{O}(k)\right\}. \tag{89}$$

However, $\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}^{\mathrm{v}} \in T_{\boldsymbol{Y}}^{\mathrm{v}}\,\mathrm{St}(k, D)$ can be written as $\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}^{\mathrm{v}} = \boldsymbol{Y}\boldsymbol{S}$ with $\boldsymbol{S} \in T_{\boldsymbol{I}_k}\mathcal{O}(k)$.

$$\begin{aligned}
T_{\boldsymbol{I}_k}\mathcal{O}(k) &= T_{\boldsymbol{I}_k}\mathrm{St}(k, k) && (90)\\
&= \left\{ \boldsymbol{I}_k\boldsymbol{C} + \left(\boldsymbol{I}_{k\perp}\boldsymbol{B} = \boldsymbol{0}_k\right) \;\middle|\; \boldsymbol{C} \in \mathrm{Skew}\,(k)\,, \boldsymbol{B} \in \mathbb{R}^{(k-k=0)\times k} \right\} && (91)\\
&= \mathrm{Skew}\,(k)\,. && (92)
\end{aligned}$$

[10] The inner product $\boldsymbol{A} \cdot \boldsymbol{C} = \boldsymbol{A}^{\top}\boldsymbol{C}$ of a vector is typically referred to as the standard inner product. Further, matrices are similarly defined with a standard inner product, defined as $\boldsymbol{A} \cdot \boldsymbol{C} = \mathrm{tr}\left(\boldsymbol{A}^{\top}\boldsymbol{C}\right)$. A space $\mathbb{R}^{D\times k}$ such that the $D \times k$ matrix $\boldsymbol{A}$ is an element is referred to as a matrix space. The standard basis of the matrix space can be constructed by a matrix wherein only one element in the matrix is 1 and the remaining are 0. The matrix space is a linear space because it satisfies the linearity that is similar to that in case of a linear vector space.

[11] When $\mathcal{N}$ is a submanifold of a Riemannian manifold $(\mathcal{M}, g)$, we define the Riemannian metric $\overline{g}$ of $\mathcal{N}$ to be:

$$\overline{g}_{\boldsymbol{x}}\left(\boldsymbol{\xi}, \boldsymbol{\eta}\right) := g_{\boldsymbol{x}}\left(\boldsymbol{\xi}, \boldsymbol{\eta}\right), \quad \boldsymbol{x} \in \mathcal{N} \subset \mathcal{M}, \boldsymbol{\xi}, \boldsymbol{\eta} \in T_{\boldsymbol{x}}\mathcal{N} \subset T_{\boldsymbol{x}}\mathcal{M}.$$

$\overline{g}$ is an induced metric and $(\mathcal{N}, \overline{g})$ is a Riemannian submanifold of $(\mathcal{M}, g)$. As $\mathrm{St}(k, D)$ is a submanifold of $\mathbb{R}^{D\times k}$, we can define the standard inner product $\boldsymbol{A} \cdot \boldsymbol{C} = \mathrm{tr}\left(\boldsymbol{A}^{\top}\boldsymbol{C}\right)$of $\mathbb{R}^{D\times k}$ as the induced metric $\overline{g}$. Thus, $\mathrm{St}(k, D)$ is a Riemannian submanifold of $\mathbb{R}^{D\times k}$.

[12] Suppose a set is given a suitable equivalence relation. A natural projection is a map that sends each element of a set to the equivalence class to which it belongs.

Thus, we obtain the following formula.

$$T_{\boldsymbol{Y}}^{\mathrm{v}} \operatorname{St}(k, D) = \{\boldsymbol{Y}\boldsymbol{C} \mid \boldsymbol{C} \in \operatorname{Skew}(k)\}\,. \tag{93}$$

Next, we define the horizontal space $T_{\boldsymbol{Y}}^{\mathrm{h}} \operatorname{St}(k, D)$ as the orthogonal complement of $T_{\boldsymbol{Y}}^{\mathrm{v}} \operatorname{St}(k, D)$ in $T_{\boldsymbol{Y}}\operatorname{St}(k, D)$ endowed with the inner product (87).

$$\begin{align} T_{\boldsymbol{Y}}^{\mathrm{h}} \operatorname{St}(k, D) &: = \left(T_{\boldsymbol{Y}}^{\mathrm{v}} \operatorname{St}(k, D)\right)^{\perp} \tag{94}\\ &= \left\{\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}^{\mathrm{h}} \in T_{\boldsymbol{Y}}\operatorname{St}(k, D) \;\middle|\; \operatorname{tr}\left(\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}^{\mathrm{h}\top} \overline{\boldsymbol{\eta}}_{\boldsymbol{Y}}^{\mathrm{v}}\right) = 0, \overline{\boldsymbol{\eta}}_{\boldsymbol{Y}}^{\mathrm{v}} \in T_{\boldsymbol{Y}}^{\mathrm{v}} \operatorname{St}(k, D)\right\}. \tag{95}\end{align}$$

Based on the fact that $T_{\boldsymbol{Y}}^{\mathrm{v}} \operatorname{St}(k, D)$ is a subspace of $T_{\boldsymbol{Y}}\operatorname{St}(k, D)$ and $T_{\boldsymbol{Y}}^{\mathrm{h}} \operatorname{St}(k, D)$ is defined as its orthogonal complement, the direct sum decomposition is as follows.

$$T_{\boldsymbol{Y}}\operatorname{St}(k, D) = T_{\boldsymbol{Y}}^{\mathrm{v}} \operatorname{St}(k, D) \oplus T_{\boldsymbol{Y}}^{\mathrm{h}} \operatorname{St}(k, D)\,, \tag{96}$$

where $\oplus$ denotes direct sum. Moreover, the tangent space is a linear space (Absil et al. (2008)). From (86), element $\boldsymbol{Y}\boldsymbol{C}$ of $T_{\boldsymbol{Y}}^{\mathrm{v}} \operatorname{St}(k, D)$ corresponds to the first term of (93); thus, (96) is formulated as follows:

$$T_{\boldsymbol{Y}}^{\mathrm{h}} \operatorname{St}(k, D) = \left\{\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}^{\mathrm{h}} = \boldsymbol{Y}_{\perp}\boldsymbol{B} \;\middle|\; \boldsymbol{B} \in \mathbb{R}^{(D-k)\times k}\right\}. \tag{97}$$

Note that the horizontal vector $\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}^{\mathrm{h}}$ is not necessarily an orthogonal matrix.

Finally, define the element $\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}^{\mathrm{h}} \in T_{\boldsymbol{Y}}^{\mathrm{h}} \operatorname{St}(k, D)$ of the horizontal space at $\boldsymbol{Y} \in \operatorname{St}(k, D)$ for the tangent vector $\boldsymbol{\xi}_{[\boldsymbol{Y}]} \in T_{[\boldsymbol{Y}]} \operatorname{Gr}(k, D)$ at $[\boldsymbol{Y}] \in \operatorname{Gr}(k, D)$ as satisfying the following formula.

$$d\pi\left(\boldsymbol{Y}\right)\left[\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}^{\mathrm{h}}\right] = \boldsymbol{\xi}_{[\boldsymbol{Y}]}, \tag{98}$$

where $d\pi\left(\boldsymbol{Y}\right) : T_{\boldsymbol{Y}}\operatorname{St}(k, D) \to T_{[\boldsymbol{Y}]} \operatorname{Gr}(k, D)$ is the derivative $\frac{d\pi(\boldsymbol{Y})}{d\boldsymbol{Y}}$ of $\pi : \operatorname{St}(k, D) \to \operatorname{Gr}(k, D)$ at $\boldsymbol{Y} \in \operatorname{St}(k, D)$. The $\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}^{\mathrm{h}} \in T_{\boldsymbol{Y}}^{\mathrm{h}} \operatorname{St}(k, D)$ is referred to as the horizontal lift at $\boldsymbol{Y} \in \operatorname{St}(k, D)$ of $[\boldsymbol{Y}] \in \operatorname{Gr}(k, D)$.

We describe the tangent space of $\operatorname{Gr}(k, D)$ with the concept of horizontal lift.

Definition 5. *Let $T_{\boldsymbol{Y}}^{\mathrm{h}} \operatorname{St}(k, D)$ be a horizontal space on $\operatorname{St}(k, D)$. Then, we define the tangent space $T_{[\boldsymbol{Y}]} \operatorname{Gr}(k, D)$ of the $\operatorname{Gr}(k, D)$ as follows.*

$$T_{[\boldsymbol{Y}]} \operatorname{Gr}(k, D) = \left\{\boldsymbol{\xi}_{[\boldsymbol{Y}]} \;\middle|\; d\pi\left(\boldsymbol{Y}\right)\left[\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}^{\mathrm{h}}\right] = \boldsymbol{\xi}_{[\boldsymbol{Y}]},\, \overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}^{\mathrm{h}} \in T_{\boldsymbol{Y}}^{\mathrm{h}} \operatorname{St}(k, D)\right\}. \tag{99}$$

From the above, $\boldsymbol{\xi}_{[\boldsymbol{Y}]} \in T_{[\boldsymbol{Y}]} \operatorname{Gr}(k, D)$ is obtained from the map $d\pi\left(\boldsymbol{Y}\right)\left[\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}^{\mathrm{h}}\right]$ when $\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}^{\mathrm{h}}$ is obtained. The $\boldsymbol{\xi}_{[\boldsymbol{Y}]}$ is defined by an equivalence class and cannot be treated numerically in matrix form; however, it is sufficient to obtain the $\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}^{\mathrm{h}}$ for actual numerical calculations. For $\boldsymbol{\xi}_{[\boldsymbol{Y}]} \in T_{[\boldsymbol{Y}]} \operatorname{Gr}(k, D)$, there exists a $\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}^{\mathrm{h}} \in T_{\boldsymbol{Y}}^{\mathrm{h}} \operatorname{St}(k, D)$ that uniquely satisfies (98). In other words, we can handle it in matrix form by using elements of the horizontal space of Stiefel manifolds through the concept of horizontal lifting. Figure 1 is a conceptual diagram of the tangent space representation of a Grassmann manifold by horizontal lift.

D.4.4 Riemannian Metric on a Grassmann Manifold

We define the Riemannian metric g of $\operatorname{Gr}(k, D)$ through the concept of horizontal lift.

Definition 6. *Let $\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}^{\mathrm{h}}$ and $\overline{\boldsymbol{\eta}}_{\boldsymbol{Y}}^{\mathrm{h}}$ be the horizontal lifts that become $d\pi\left(\boldsymbol{Y}\right)\left[\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}^{\mathrm{h}}\right] = \boldsymbol{\xi}_{[\boldsymbol{Y}]}$ and $d\pi\left(\boldsymbol{Y}\right)\left[\overline{\boldsymbol{\eta}}_{\boldsymbol{Y}}^{\mathrm{h}}\right] = \boldsymbol{\eta}_{[\boldsymbol{Y}]}$, respectively. Then, we define the Riemannian metric on $\operatorname{Gr}(k, D)$ as follows:*

$$g_{[\boldsymbol{Y}]}\left(\boldsymbol{\xi}_{[\boldsymbol{Y}]}, \boldsymbol{\eta}_{[\boldsymbol{Y}]}\right) := \bar{g}_{\boldsymbol{Y}}\left(\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}^{\mathrm{h}}, \overline{\boldsymbol{\eta}}_{\boldsymbol{Y}}^{\mathrm{h}}\right) = \operatorname{tr}\left(\boldsymbol{B}^{\top}\boldsymbol{D}\right), \tag{100}$$

where $\boldsymbol{B}$ and $\boldsymbol{D}$ are matrices that are $\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}^{\mathrm{h}} = \boldsymbol{Y}_{\perp}\boldsymbol{B}$ and $\overline{\boldsymbol{\eta}}_{\boldsymbol{Y}}^{\mathrm{h}} = \boldsymbol{Y}_{\perp}\boldsymbol{D}$, respectively.

D.5 Invariant Measures

Let the column vectors of matrix $\boldsymbol{Y} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k\} \in \mathbb{R}^{D \times k}$ be the orthonormal basis that span the subspace $\mathrm{span}(\boldsymbol{Y}) \in \mathrm{Gr}(k, D)$ in $\mathbb{R}^D$, and the column vectors of $\boldsymbol{Y}_\perp = \{\boldsymbol{y}_{k+1}, \cdots, \boldsymbol{y}_D\} \in \mathbb{R}^{D \times D-k}$ be the orthogonal complementary space $\mathrm{span}(\boldsymbol{Y}_\perp)$ of $\mathrm{span}(\boldsymbol{Y})$, respectively. Then, the following differential form can be defined.

$$\begin{aligned}(d\boldsymbol{Y}) &= \bigwedge_{j=1}^{D-k} \bigwedge_{i=1}^{k} \boldsymbol{y}_{k+j}^\top d\boldsymbol{y}_i && (101)\\ &= \left(\boldsymbol{y}_{k+1}^\top d\boldsymbol{y}_1 \wedge \cdots \wedge \boldsymbol{y}_{k+1}^\top d\boldsymbol{y}_k\right) \wedge \cdots \wedge \left(\boldsymbol{y}_D^\top d\boldsymbol{y}_1 \wedge \cdots \wedge \boldsymbol{y}_D^\top d\boldsymbol{y}_k\right), && (102)\end{aligned}$$

where $\wedge$ is the wedge product and the relation satisfies $\omega_i \wedge \omega_i = \omega_j \wedge \omega_j = 0$ and $\omega_i \wedge \omega_j = -\omega_j \wedge \omega_i$. The above equation is in $k\,(D-k)$-order differential form, which is an invariant measure on $\mathrm{Gr}(k, D)$ (Chikuse (2003)).

If we define the matrix $\boldsymbol{X}_\perp$ to be $[\ \boldsymbol{X} \quad \boldsymbol{X}_\perp\ ]$ for any point $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k\} \in \mathrm{St}(k, D)$, the differential form for an invariant measure on $\mathrm{St}(k, D)$ is defined as follows.

$$(d\boldsymbol{X}) = \bigwedge_{j=1}^{D-k} \bigwedge_{i=1}^{k} \boldsymbol{x}_{k+j}^\top d\boldsymbol{x}_i \bigwedge_{i<j} {}^{k}_{1}\boldsymbol{x}_j^\top d\boldsymbol{x}_i = (d\boldsymbol{Y})\,(d\boldsymbol{Q}), \tag{103}$$

where $(d\boldsymbol{Q})$ is the invariant measure of $\mathcal{O}(k)$. The integral of (103), that is, the volume of $\mathrm{St}(k, D)$, can be evaluated as follows:

$$V_{\mathrm{St}(k,D)} = \int_{\mathrm{St}(k,D)} (d\boldsymbol{X}). \tag{104}$$

(104) can be computed as follows. First, the surface S_D of the D-dimensional unit sphere can be defined as follows:

$$S_D = \left.\frac{d}{dr}\right|_{r=1} V_D = DV_D = \frac{D\pi^{\frac{D}{2}}}{\Gamma\left(\frac{D}{2}+1\right)} = \frac{2\pi^{\frac{D}{2}}}{\Gamma\left(\frac{D}{2}\right)}, \tag{105}$$

where V_D is the volume of a D-dimensional sphere $\frac{\pi^{\frac{D}{2}}}{\Gamma\left(\frac{D}{2}+1\right)} r^D$ and $\Gamma\left(\frac{D}{2}\right)$ is the gamma function. Then, the following equation is obtained.

$$\int_{\mathrm{St}(k,D)} (d\boldsymbol{X}) = S_D \int_{\mathrm{St}(k-1,D-1)} (d\boldsymbol{X}_1), \tag{106}$$

where $(d\boldsymbol{X}_1)$ is the differential form of $\mathrm{St}(k-1, D-1)$. Thus, (104) can be transformed as follows.

$$V_{\mathrm{St}(k,D)} = \int_{\mathrm{St}(k,D)} (d\boldsymbol{X}) = \prod_{i=1}^{k} S_D = \frac{2^k \pi^{\frac{Dk}{2}}}{\Gamma_k\left(\frac{D}{2}\right)}, \tag{107}$$

where $\Gamma_k\left(\frac{D}{2}\right)$ is the multidimensional gamma function. The invariant measure $(d\boldsymbol{X})$ is an unnormalized measure. A measure normalized to be a probability measure can be formulated as follows:

$$[d\boldsymbol{X}] = \frac{1}{V_{\mathrm{St}(k,D)}}\,(d\boldsymbol{X}). \tag{108}$$

This is a uniform distribution on $\mathrm{St}(k, D)$. As $\mathrm{St}(k, k) = \mathcal{O}(k)$, the volume $V_{\mathcal{O}(k)}$ of $\mathcal{O}(k)$ can be represented using $(d\boldsymbol{Q})$ as follows.

$$V_{\mathcal{O}(k)} = V_{\mathrm{St}(k,k)} = \frac{2^k \pi^{\frac{k^2}{2}}}{\Gamma_k\left(\frac{k}{2}\right)}. \tag{109}$$

Furthermore, a measure normalized to be a probability measure can be represented by the following:

$$[d\boldsymbol{Q}] = \frac{1}{V_{\mathcal{O}(k)}}\,(d\boldsymbol{Q}). \tag{110}$$

From the above, the probability density function $U_{\mathcal{O}(k)}\left(\boldsymbol{Q}\right)$ of the uniform distribution on $\mathcal{O}(k)$ is as follows:

$$U_{\mathcal{O}(k)}\left(\boldsymbol{Q}\right)=\frac{1}{V_{\mathcal{O}(k)}}\quad \text{s.t.}\quad \boldsymbol{Q}\in\mathcal{O}(k)\,, \tag{111}$$

$$\int_{\mathcal{O}(k)} U_{\mathcal{O}(k)}\left(\boldsymbol{Q}\right)\left(d\boldsymbol{Q}\right)=\int_{\mathcal{O}(k)}\frac{1}{V_{\mathcal{O}(k)}}\left(d\boldsymbol{Q}\right)=\int_{\mathcal{O}(k)}\left[d\boldsymbol{Q}\right]=1. \tag{112}$$

As $\mathrm{Gr}(k,D)$ is defined as a quotient manifold $\mathrm{St}(k,D)\,/\mathcal{O}(k)$ as in (2), the volume $V_{\mathrm{Gr}(k,D)}$ of $\mathrm{Gr}(k,D)$ can be defined as follows.

$$V_{\mathrm{Gr}(k,D)}=\int_{\mathrm{Gr}(k,D)}\left(d\boldsymbol{Y}\right)=\frac{V_{\mathrm{St}(k,D)}}{V_{\mathcal{O}(k)}}=\frac{V_{\mathrm{St}(k,D)}}{V_{\mathrm{St}(k,k)}}=\frac{\pi^{\frac{k(D-k)}{2}}\Gamma_k\left(\frac{k}{2}\right)}{\Gamma_k\left(\frac{D}{2}\right)}. \tag{113}$$

The measure normalized to be a probability measure is expressed as:

$$\left[d\boldsymbol{Y}\right]=\frac{1}{V_{\mathrm{Gr}(k,D)}}\left(d\boldsymbol{Y}\right). \tag{114}$$

D.6 Retraction

In general, the points except the origin ($\boldsymbol{p}\left(0\right)=\boldsymbol{x}$) of the tangent space $T_{\boldsymbol{x}}\mathcal{M}$ at $\boldsymbol{x}$ on the manifold $\mathcal{M}$ are not elements on $\mathcal{M}$ ($\boldsymbol{p}\left(t\right)\in T_{\boldsymbol{x}}\mathcal{M}, t\neq 0$). Therefore, if the result of the operation on the tangent space is to be used at another point on the manifold $\mathcal{M}$, it is necessary to map $\boldsymbol{p}\left(t\right)$ to the manifold $\mathcal{M}$. The map from a tangent space to a manifold is referred to as an exponential map. However, because the exponential map is computationally expensive, retraction based on numerical linear algebra is often used as an alternative (Zhu & Sato (2021)). Retraction is a method for approximating an exponential map to first order while maintaining global convergence in optimization algorithms on Riemannian manifolds. The most commonly used retractions on $\mathrm{Gr}(k,D)$ are methods based on QR decomposition or singular-value decomposition (SVD) (Absil et al. (2008); Zhu & Sato (2021)). In addition, a retraction based on the Cayley transform is introduced in Zhu & Sato (2021). This retraction is closely related to the Cayley transform on $\mathrm{St}(k,D)$ (Wen & Yin (2013); Xiaojing (2017); Zhu & Duan (2019)) and the Projected polynomial retraction (Gawlik & Leok (2018a)).

D.6.1 Exponential Map and Retraction

Geodesics on $\mathrm{Gr}(k,D)$ can be expressed as the equivalence class $\left[\exp^{\mathrm{Gr}}_{\boldsymbol{Y}}\left(t\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}\right)\right]$, where

$$\exp^{\mathrm{Gr}}_{\boldsymbol{Y}}\left(t\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}\right)=\left[\begin{array}{cc}\boldsymbol{Y} & \boldsymbol{Y}_{\perp}\end{array}\right]\exp\left(t\mathfrak{B}\right)\boldsymbol{I}_{D\times k}. \tag{115}$$

Here, exp on the right-hand side is the matrix exponential, and $\mathfrak{B}=\left[\begin{array}{cc}\boldsymbol{0}_k & -\boldsymbol{B}^{\top}\\ \boldsymbol{B} & \boldsymbol{0}_{D-k}\end{array}\right]\in\mathrm{skew}\left(D\right)$, where $\boldsymbol{B}$ satisfies $\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}=\boldsymbol{Y}_{\perp}\boldsymbol{B}$. We can use a following exponential map that is mathematically equivalent to (115) (Edelman et al. (1998)):

$$\exp^{\mathrm{Gr}}_{\boldsymbol{Y}}\left(t\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}\right):=\left\{\boldsymbol{Y}\boldsymbol{V}\cos\left(\boldsymbol{\Sigma}t\right)+\boldsymbol{U}\sin\left(\boldsymbol{\Sigma}t\right)\right\}\boldsymbol{V}^{\top}, \tag{116}$$

where $\boldsymbol{U},\boldsymbol{\Sigma},\boldsymbol{V}^{\top}=\mathrm{SVD}\left(\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}\right)$.

Further, we can use the Padé approximation to approximate geodesics on Grassmann manifolds as follows:

$$\boldsymbol{Y}(t)=\left[\begin{array}{cc}\boldsymbol{Y} & \boldsymbol{Y}_{\perp}\end{array}\right]r_m(t\mathfrak{B})\boldsymbol{I}_{D\times k}\approx\exp^{\mathrm{Gr}}_{\boldsymbol{Y}}\left(t\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}\right), \tag{117}$$

where $r_m\left(\boldsymbol{X}\right)$ is the mth-order diagonal Padé approximation to the matrix exponential $\exp\left(\boldsymbol{X}\right)$. See the expression of $r_m\left(\boldsymbol{X}\right)$ in Moler & Loan (2003). The simplest member of this class is surely the first-order Padé approximation

$$\begin{aligned}\overline{R}_{\boldsymbol{Y}}\left(t\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}\right):=&\left[\begin{array}{cc}\boldsymbol{Y} & \boldsymbol{Y}_{\perp}\end{array}\right]r_1\left(t\mathfrak{B}\right)\boldsymbol{I}_{D\times k} &(118)\\ =&\left[\begin{array}{cc}\boldsymbol{Y} & \boldsymbol{Y}_{\perp}\end{array}\right]\left(I_n-\frac{t}{2}\mathfrak{B}\right)^{-1}\left(\boldsymbol{I}_D+\frac{t}{2}\mathfrak{B}\right)\boldsymbol{I}_{D\times k}, &(119)\end{aligned}$$

which is also known as the Cayley transform. From the error expression $\exp\left(\boldsymbol{Y}\right)=r_m\left(\boldsymbol{Y}\right)+O\left(\|\boldsymbol{Y}\|^{2m+1}\right)$ of the Padé approximation, we have

$$\overline{R}_{\boldsymbol{Y}}\left(t\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}\right)=\exp^{\mathrm{Gr}}_{\boldsymbol{Y}}\left(t\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}\right)+O\left(t^{2m+1}\left\|\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}\right\|^{2m+1}\right), \tag{120}$$

which is also given by Theorem 3 in Gawlik & Leok (2018b).

D.6.2 Horizontal Retraction

From Definition 3 in Zhu & Sato (2021), (119) is a horizontal retraction, and

$$R_{[\boldsymbol{Y}]}\left(t\boldsymbol{\xi}_{[\boldsymbol{Y}]}\right) := \left[\overline{R}_{\boldsymbol{Y}}\left(t\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}}\right)\right] \tag{121}$$

is a retraction on $\mathrm{Gr}(k, D)$ as a quotient manifold defined by (2). This is because $\overline{R}$ satisfies the invariance condition that $\overline{R}_{\boldsymbol{Y}}\left(t\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}}\right) \sim \overline{R}_{\boldsymbol{Y}'}\left(t\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}'}\right)$ for all $\boldsymbol{Y} \in \mathrm{St}(k, D)\,, \boldsymbol{Y}' \in \mathrm{St}(k, D)\,, \overline{\boldsymbol{\xi}}_{\boldsymbol{Y}} \in T^{\mathrm{h}}_{\boldsymbol{Y}}\,\mathrm{St}(k, D)$ and $\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}'} \in T^{\mathrm{h}}_{\boldsymbol{Y}'}\,\mathrm{St}(k, D)$ such that $\boldsymbol{Y} \sim \boldsymbol{Y}'$ and $\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}}$ and $\overline{\boldsymbol{\xi}}_{\boldsymbol{Y}'}$ are horizontal lifts of $\boldsymbol{\xi}_{[\boldsymbol{Y}]} \in T_{[\boldsymbol{Y}]}\,\mathrm{Gr}(k, D)$ at $\boldsymbol{Y}$ and $\boldsymbol{Y}'$, respectively.

In low-rank cases, we can obtain an economical version of (119) as follows (Zhu & Sato (2021)).

$$\overline{R}_{\boldsymbol{Y}}\left(t\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}}\right) = \boldsymbol{Y} + t\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}} - \left(\frac{t^2}{2}\boldsymbol{Y} + \frac{t^3}{4}\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}}\right)\left(\boldsymbol{I}_k + \frac{t^2}{4}\overline{\boldsymbol{\xi}}^{\mathrm{h}\top}_{\boldsymbol{Y}}\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}}\right)^{-1}\overline{\boldsymbol{\xi}}^{\mathrm{h}\top}_{\boldsymbol{Y}}\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}}. \tag{122}$$

The inverse retraction $\left(\overline{R^{-1}_{[\boldsymbol{Y}]}}\right)^{\mathrm{h}}_{\boldsymbol{Y}} : \mathrm{St}(k, D) \to T^{\mathrm{h}}_{\boldsymbol{Y}}\,\mathrm{St}(k, D)$ of $\overline{R}_{\boldsymbol{Y}}\left(\overline{\boldsymbol{\xi}}^{\mathrm{h}}_{\boldsymbol{Y}}\right)$ is the following:

$$\begin{aligned}
\left(\overline{R^{-1}_{[\boldsymbol{Y}]}\left([\boldsymbol{X}]\right)}\right)^{\mathrm{h}}_{\boldsymbol{Y}} &= \overline{R}^{-1}_{\boldsymbol{Y}}\left(\boldsymbol{X}\right) && (123)\\
&= 2\boldsymbol{Y}_{\perp}\boldsymbol{Y}_{\perp}^{\top}\boldsymbol{X}\left(\boldsymbol{I}_k + \boldsymbol{Y}^{\top}\boldsymbol{X}\right)^{-1} && (124)\\
&= 2\left(\boldsymbol{X} - \boldsymbol{Y}\boldsymbol{Y}^{\top}\boldsymbol{X}\right)\left(\boldsymbol{I}_k + \boldsymbol{Y}^{\top}\boldsymbol{X}\right)^{-1}. && (125)
\end{aligned}$$