A Proofs

A.1 Proof of Lemma 4.1

$$\begin{bmatrix}\dot\theta\\ \dot\phi\end{bmatrix} = \begin{bmatrix}\phi\\ -\theta\phi^T\phi + (D-\theta\,\mathrm{Sym}(\theta^T D))/\eta\end{bmatrix}, \tag{1}$$

where $\eta > 0$ will be later used in the discrete-time algorithm as the step-size and

$$D = D(\theta,\phi) := -\gamma\phi - \nabla\mathcal{L}(\theta). \tag{2}$$

Lemma 4.1 *For the dynamical system* (1) *defined on* $\mathbb{R}^{n\times p}\times\mathbb{R}^{n\times p}$*, if* $\theta(0)^T\theta(0) = I$ *and* $\mathrm{Sym}(\theta(0)^T\phi(0)) = 0$*, then* $\theta(t)^T\theta(t) = I$ *and* $\mathrm{Sym}(\theta(t)^T\phi(t)) = 0$ *hold for all* $t \geq 0$*.*

Proof. Consider the following dynamical system defined in a local neighborhood of $T\mathrm{St}(n,p)$ in $\mathbb{R}^{n\times p}\times\mathbb{R}^{n\times p}$ where $\theta^T\theta$ is invertible:

$$\begin{cases}\dot\theta = \phi\\ \dot\phi = -\theta(\theta^T\theta)^{-1}\phi^T\phi + (D-\theta(\theta^T\theta)^{-1}\,\mathrm{Sym}(\theta^T D))/\eta\end{cases} \tag{3}$$

Clearly, systems (1) and (3) become identical for $(\theta,\phi)\in T\mathrm{St}(n,p)$. Along each trajectory of the system (3) with $\theta(0)^T\theta(0) = I$ and $\mathrm{Sym}(\theta(0)^T\phi(0)) = 0$, we have

$$\begin{aligned}\frac{d}{dt}(\theta(t)^T\theta(t)) &= \theta(t)^T\dot\theta(t) + \dot\theta(t)^T\theta(t)\\ &= \theta(t)^T\phi(t) + \phi(t)^T\theta(t)\\ &= 2\,\mathrm{Sym}(\theta(t)^T\phi(t)),\end{aligned} \tag{4}$$

which is 0 when $t = 0$ by the given condition. It is easy to check that along each trajectory of the system (3),

$$\frac{d\,\mathrm{Sym}(\theta(t)^T\phi(t))}{dt} = \mathrm{Sym}(\dot\theta^T\phi) + \mathrm{Sym}(\theta^T\dot\phi) \equiv 0. \tag{5}$$

By simple integration of (4) and (5) in t, one can see that if $\theta(0)^T\theta(0) = I$ and $\mathrm{Sym}(\theta(0)^T\phi(0)) = 0$ then $\theta(t)^T\theta(t) = I$ and $\mathrm{Sym}(\theta(t)^T\phi(t)) = 0$ for all $t \geq 0$ for (3). This property also holds true for (1) since (3) coincides with (1) on $T\mathrm{St}(n,p)$. ■

A.2 Proof of Lemma 4.2

Lemma 4.2 *Assume that* $\theta^* = \arg\min_{\theta\in\mathrm{St}(n,p)}\mathcal{L}(\theta)$ *uniquely exists, and let* $c_0 \geq 0$ *such that* $(\theta^*,0)$ *is the only point in* $\{(\theta,0)\in\Omega : \mathrm{grad}_\theta\,\mathcal{L}(\theta) = 0\}$*, where* $\Omega = \{(\theta,\phi)\in T\mathrm{St}(n,p) : \frac{\eta}{2}\|\phi\|^2 + \mathcal{L}(\theta) \leq c_0\}$*. Assume that* $\mathcal{L}$ *is* C^2 *on* $\{\theta\in\mathrm{St}(n,p) : \mathcal{L}(\theta)\leq c_0\}$*. Then each trajectory of* (1) *starting in* Ω *stays in* Ω *for all forward time and asymptotically converges to* $(\theta^*,0)$ *as time tends to infinity.*

Proof. Let $(\theta(t),\phi(t))_{t\geq 0}$ be a trajectory of the dynamical system (1) starting in $\Omega\subset T\mathrm{St}(n,p)$. By Lemma 4.1, $\theta(t)^T\theta(t) = I$ and $\mathrm{Sym}(\theta(t)^T\phi(t)) = 0$ for all $t\geq 0$. Besides, Ω is compact because $\|\theta\|^2 = \mathrm{tr}(\theta^T\theta) = p$ and $\|\phi\|^2 \leq \frac{2}{\eta}(c_0 - \mathcal{L}(\theta)) \leq 2c_0/\eta$ for all $(\theta,\phi)\in\Omega$.

Let $U = U(\theta,\phi) = \frac{\eta}{2}\|\phi\|^2 + \mathcal{L}(\theta)$, we have $U\geq 0$ for all $(\theta,\phi)\in\Omega$, and

$$\begin{aligned}\frac{d}{dt}U(\theta(t),\phi(t)) &= \eta\langle\phi,\dot\phi\rangle + \langle\nabla\mathcal{L}(\theta),\dot\theta\rangle\\ &= -\gamma\|\phi\|^2 - \langle\theta^T\phi, \eta\phi^T\phi + \mathrm{Sym}(\theta^T D)\rangle\\ &= -\gamma\|\phi\|^2 - \langle\mathrm{Sym}(\theta^T\phi), \eta\phi^T\phi + \mathrm{Sym}(\theta^T D)\rangle\\ &= -\gamma\|\phi\|^2 \leq 0,\end{aligned} \tag{6}$$

where the last inequality holds as equality if and only if $\phi = 0$. Therefore, U is non-increasing along the trajectory, in particular, $(\theta(t),\phi(t))\in\Omega$ for all $t\geq 0$.

Let $E = \{(\theta,\phi)\in\Omega : \phi = 0\}$, and let $M\subset E$ be the largest invariant set for (1) in E. Clearly, $(\theta^*,0)\in M$. Now, consider a trajectory $(\theta(t),\phi(t))_{t\geq 0}$ in $M\subset E$, we have $\phi(t)\equiv 0$, thus we have $0 \equiv \dot{\phi}(t) = -\theta(t)\phi^T(t)\phi(t) + (D - \theta(t)\operatorname{Sym}(\theta(t)^T D))/\eta = -\operatorname{grad}_\theta \mathcal{L}(\theta(t))/\eta$, and thus $\operatorname{grad}_\theta \mathcal{L}(\theta(t)) \equiv 0$, which implies that $M = \{(\theta^*, 0)\}$. By LaSalle's Invariance Principle, the desired result follows. ■

Alternatively, we can have the following lemma.

Lemma 4.2A *Fix a $c \geq 0$ such that $\Omega = \{(\theta, \phi) \in T\mathrm{St}(n,p) : \frac{\eta}{2}\|\phi\|^2 + \mathcal{L}(\theta) \leq c\}$ is nonempty, and let $L = \{(\theta, 0) \in \Omega : \operatorname{grad}_\theta \mathcal{L}(\theta) = 0\}$ be the set of equilibrium points in Ω. Assume that $\mathcal{L}$ is C^2 on $\{\theta \in \mathrm{St}(n,p) : \mathcal{L}(\theta) \leq c_0\}$. Then each trajectory of (1) starting in Ω stays in Ω for all forward time and asymptotically converges to L as time tends to infinity.*

Proof. Let $(\theta(t), \phi(t))_{t\geq 0}$ be a trajectory of the dynamical system (1) starting in $\Omega \subset T\mathrm{St}(n,p)$. By Lemma 4.1, $\theta(t)^T\theta(t) = I$ and $\operatorname{Sym}(\theta(t)^T\phi(t)) = 0$ for all $t \geq 0$. Besides, Ω is compact because $\|\theta\|^2 = \operatorname{tr}(\theta^T\theta) = p$ and $\|\phi\|^2 \leq \frac{2}{\eta}(c_0 - \mathcal{L}(\theta)) \leq 2c_0/\eta$ for all $(\theta, \phi) \in \Omega$.

Let $U = U(\theta, \phi) = \frac{\eta}{2}\|\phi\|^2 + \mathcal{L}(\theta)$, we have $U \geq 0$ for all $(\theta, \phi) \in \Omega$, and

$$\begin{aligned}
\frac{d}{dt}U(\theta(t), \phi(t)) &= \eta\langle\phi, \dot{\phi}\rangle + \langle\nabla\mathcal{L}(\theta), \dot{\theta}\rangle \\
&= -\gamma\|\phi\|^2 - \langle\theta^T\phi, \eta\phi^T\phi + \operatorname{Sym}(\theta^T D)\rangle \\
&= -\gamma\|\phi\|^2 - \langle\operatorname{Sym}(\theta^T\phi), \eta\phi^T\phi + \operatorname{Sym}(\theta^T D)\rangle \\
&= -\gamma\|\phi\|^2 \leq 0,
\end{aligned} \tag{7}$$

where the last inequality holds as equality if and only if $\phi = 0$. Therefore, U is non-increasing along the trajectory, in particular, $(\theta(t), \phi(t)) \in \Omega$ for all $t \geq 0$.

Let $E = \{(\theta, \phi) \in \Omega : \phi = 0\}$, and let $M \subset E$ be the largest invariant set for (1) in E. Clearly, $L \subset M$. Now, consider a trajectory $(\theta(t), \phi(t))_{t\geq 0}$ in $M \subset E$, we have $\phi(t) \equiv 0$, thus we have $0 \equiv \dot{\phi}(t) = -\theta(t)\phi^T(t)\phi(t) + (D - \theta(t)\operatorname{Sym}(\theta(t)^T D))/\eta = -\operatorname{grad}_\theta \mathcal{L}(\theta(t))/\eta$, and thus $\operatorname{grad}_\theta \mathcal{L}(\theta(t)) \equiv 0$, which implies that $M \subset L$. Therefore, $M = L$, and by LaSalle's Invariance Principle, the desired result follows. ■

A.3 Proof of Lemma 4.4

Let $V : \mathbb{R}^{n\times p} \times \mathbb{R}^{n\times p} \to \mathbb{R}_{\geq 0}$ be a function defined as

$$V(\theta, \phi) = \frac{k_1}{4}\left\|\theta^T\theta - I\right\|^2 + \frac{k_2}{2}\left\|\operatorname{Sym}(\theta^T\phi)\right\|^2, \tag{8}$$

where $k_1, k_2 > 0$. We have

$$V^{-1}(0) = \{(\theta, \phi) \in \mathbb{R}^{n\times p} \times \mathbb{R}^{n\times p} : \theta^T\theta = I, \operatorname{Sym}(\theta^T\phi) = 0\} = T\mathrm{St}(n,p), \tag{9}$$

and

$$\nabla V(\theta, \phi) = \begin{bmatrix} \nabla_\theta V(\theta, \phi) \\ \nabla_\phi V(\theta, \phi) \end{bmatrix} = \begin{bmatrix} k_1\theta(\theta^T\theta - I) + k_2\phi\operatorname{Sym}(\theta^T\phi) \\ k_2\theta\operatorname{Sym}(\theta^T\phi) \end{bmatrix}. \tag{10}$$

Lemma 4.4 *For each $0 < c < k_1/4$, the set of all critical points of V in $V^{-1}([0,c])$ is $V^{-1}(0)$.*

Proof. Since 0 is the minimum value of V, every point in $V^{-1}(0)$ is a critical point of V. Let (θ, ϕ) be a critical point of V in $V^{-1}([0,c])$, we then have

$$\begin{cases} k_1\theta(\theta^T\theta - I) + k_2\phi\operatorname{Sym}(\theta^T\phi) = 0 \\ k_2\theta\operatorname{Sym}(\theta^T\phi) = 0 \end{cases} \tag{11}$$

By left-multiplying the second equation in (11) by $(\theta^T\theta)^{-1}\theta^T$, we have $\operatorname{Sym}(\theta^T\phi) = 0$. Plug that in the first equation and we have $\theta(\theta^T\theta - I) = 0$, which yields $\theta^T\theta = I$ by left-multiplying $(\theta^T\theta)^{-1}\theta^T$, completing the proof. ■

A.4 Proof of Lemma 4.5

Let $W = S \times \mathbb{R}^{n\times p}$, where

$$S = \{\theta \in \mathbb{R}^{n\times p} : \left\|\theta^T\theta - I\right\| < 1\} \tag{12}$$

is an open neighborhood of $\mathrm{St}(n,p)$ such that $\theta^T\theta$ is invertible for all $\theta \in S$.

$$X(\theta,\phi) = \begin{bmatrix} X_\theta(\theta,\phi) \\ X_\phi(\theta,\phi) \end{bmatrix} = \begin{bmatrix} \phi - \theta(\theta^T\theta)^{-1}\operatorname{Sym}(\theta^T\phi) \\ \theta(\theta^T\theta)^{-1}((\theta^T\theta)^{-1}\theta^T\phi\operatorname{Sym}(\theta^T\phi) - \phi^T\phi) + (D - \theta(\theta^T\theta)^{-1}\operatorname{Sym}(\theta^T D))/\eta \end{bmatrix}$$

with D defined in (2), and $\alpha > 0$ is the feedback coefficient.

Lemma 4.5 $\langle \nabla V(\theta,\phi), X(\theta,\phi)\rangle = 0, \forall(\theta,\phi) \in W$.

Proof. We have

$$\langle \nabla_\theta V, X_\theta\rangle = \langle k_1\theta(\theta^T\theta - I) + k_2\phi\operatorname{Sym}(\theta^T\phi), \phi - \theta(\theta^T\theta)^{-1}\operatorname{Sym}(\theta^T\phi)\rangle, \tag{13}$$

where

$$\langle \theta(\theta^T\theta - I), \phi - \theta(\theta^T\theta)^{-1}\operatorname{Sym}(\theta^T\phi)\rangle = \langle \theta^T\theta - I, \theta^T\phi - \operatorname{Sym}(\theta^T\phi)\rangle = 0. \tag{14}$$

Hence

$$\begin{aligned} \langle \nabla_\theta V, X_\theta\rangle &= k_2\langle \phi\operatorname{Sym}(\theta^T\phi), \phi - \theta(\theta^T\theta)^{-1}\operatorname{Sym}(\theta^T\phi)\rangle \\ &= k_2\langle \operatorname{Sym}(\theta^T\phi), \phi^T\phi - \phi^T\theta(\theta^T\theta)^{-1}\operatorname{Sym}(\theta^T\phi)\rangle. \end{aligned} \tag{15}$$

Meanwhile, we have

$$\langle \nabla_\phi V, X_\phi\rangle = k_2\langle \operatorname{Sym}(\theta^T\phi), (\theta^T\theta)^{-1}\theta^T\phi\operatorname{Sym}(\theta^T\phi) - \phi^T\phi\rangle. \tag{16}$$

Now it suffices to show that

$$\langle \operatorname{Sym}(\theta^T\phi), \phi^T\theta(\theta^T\theta)^{-1}\operatorname{Sym}(\theta^T\phi)\rangle = \langle \operatorname{Sym}(\theta^T\phi), (\theta^T\theta)^{-1}\theta^T\phi\operatorname{Sym}(\theta^T\phi)\rangle. \tag{17}$$

Indeed

$$\begin{aligned} \langle \operatorname{Sym}(\theta^T\phi), \phi^T\theta(\theta^T\theta)^{-1}\operatorname{Sym}(\theta^T\phi)\rangle &= \langle (\operatorname{Sym}(\theta^T\phi))^2, \phi^T\theta(\theta^T\theta)^{-1}\rangle \\ &= \langle (\operatorname{Sym}(\theta^T\phi))^2, (\theta^T\theta)^{-1}\theta^T\phi\rangle \\ &= \langle \operatorname{Sym}(\theta^T\phi), (\theta^T\theta)^{-1}\theta^T\phi\operatorname{Sym}(\theta^T\phi)\rangle, \end{aligned} \tag{18}$$

completing the proof. ■

A.5 Proof of Lemma 4.6

For $(\theta,\phi) \in W$, we define $L \in \mathbb{R}^{2n\times 2n}$ as

$$L = L(\theta,\phi) = \begin{bmatrix} \frac{1}{4k_1}\theta(\theta^T\theta)^{-2}\theta^T & -\frac{1}{4k_1}\theta(\theta^T\theta)^{-2}\theta^T\phi(\theta^T\theta)^{-1}\theta^T \\ -\frac{1}{4k_1}\theta(\theta^T\theta)^{-1}\phi^T\theta(\theta^T\theta)^{-2}\theta^T & L_2(\theta,\phi) \end{bmatrix},$$

where $L_2(\theta,\phi) = \frac{1}{4k_1}\theta(\theta^T\theta)^{-1}\phi^T\theta(\theta^T\theta)^{-2}\theta^T\phi(\theta^T\theta)^{-1}\theta^T + \frac{1}{2k_2}\theta(\theta^T\theta)^{-2}\theta^T$.

Lemma 4.6 $\langle \nabla V(\theta,\phi), L(\theta,\phi)\nabla V(\theta,\phi)\rangle = V(\theta,\phi), \forall(\theta,\phi) \in W$.

Proof. By direct computation, it is easy to check that

$$L = L(\theta,\phi) = M^T(\theta,\phi)N(\theta,\phi)M(\theta,\phi), \tag{19}$$

where

$$\begin{aligned} M(\theta,\phi) &= \begin{bmatrix} I_n & -\phi(\theta^T\theta)^{-1}\theta^T \\ 0_n & I_n \end{bmatrix}, \\ N(\theta,\phi) &= \begin{bmatrix} \frac{1}{4k_1}\theta(\theta^T\theta)^{-2}\theta^T & 0_n \\ 0_n & \frac{1}{2k_2}\theta(\theta^T\theta)^{-2}\theta^T \end{bmatrix}. \end{aligned}$$

First we compute

$$M\nabla V = \begin{bmatrix} I_n & -\phi(\theta^T\theta)^{-1}\theta^T \\ 0_n & I_n \end{bmatrix}\begin{bmatrix} k_1\theta(\theta^T\theta - I) + k_2\phi\operatorname{Sym}(\theta^T\phi) \\ k_2\theta\operatorname{Sym}(\theta^T\phi) \end{bmatrix} = \begin{bmatrix} k_1\theta(\theta^T\theta - I) \\ k_2\theta\operatorname{Sym}(\theta^T\phi) \end{bmatrix} \tag{20}$$

and then we have

$$NM\nabla V = \begin{bmatrix} \frac{1}{4k_1}\theta(\theta^T\theta)^{-2}\theta^T & 0_n \\ 0_n & \frac{1}{2k_2}\theta(\theta^T\theta)^{-2}\theta^T \end{bmatrix} \begin{bmatrix} k_1\theta(\theta^T\theta - I) \\ k_2\theta\,\mathrm{Sym}(\theta^T\phi) \end{bmatrix}$$
$$= \begin{bmatrix} \theta(\theta^T\theta)^{-1}(\theta^T\theta - I)/4 \\ \theta(\theta^T\theta)^{-1}\,\mathrm{Sym}(\theta^T\phi)/2 \end{bmatrix}. \quad (21)$$

Thus we have

$$\begin{aligned} \langle \nabla V, L\nabla V\rangle =& \langle \nabla V, M^T NM\nabla V\rangle \\ =& \langle M\nabla V, NM\nabla V\rangle \\ =& \langle k_1\theta(\theta^T\theta - I), \theta(\theta^T\theta)^{-1}(\theta^T\theta - I)/4\rangle + \langle k_2\theta\,\mathrm{Sym}(\theta^T\phi), \theta(\theta^T\theta)^{-1}\,\mathrm{Sym}(\theta^T\phi)/2\rangle \\ =& \frac{k_1}{4}\left\|\theta^T\theta - I\right\|^2 + \frac{k_2}{2}\left\|\mathrm{Sym}(\theta^T\phi)\right\|^2 = V, \end{aligned} \quad (22)$$

completing the proof. ■

A.6 Proof of Theorem 4.8

Theorem 4.8 *With the approximation $(\theta^T\theta)^{-1} \approx 2I - \theta^T\theta$, the additional time complexity of Algorithm 1 is $O(p^2n)$.*

Proof. The additional time complexity comes from matrix multiplications and additions, where the time complexity from additions is relatively negligible. Specifically, the algorithm contains $(p \times n)$-by-$(n \times p)$, $(p \times p)$-by-$(p \times n)$, and $(p \times p)$-by-$(p \times p)$ matrix multiplications, which have time complexity $O(p^2n)$, $O(p^2n)$, and $O(p^3)$, respectively. Since $n \geq p$, the result follows. Refer to section C for the details of the implementation. ■

A.7 Proof of Theorem 4.9

Theorem 4.9 *Assume there exists a constant $c > 0$ such that $\|\phi\| \leq c$ for all timesteps when using Algorithm 1 with the approximation $(\theta^T\theta)^{-1} \approx 2I - \theta^T\theta$. For any given $0 < \epsilon < 1$, there exist $\eta = \eta(\epsilon) > 0$ and $\alpha = \alpha(\epsilon) > 0$ so that $\left\|\theta^T\theta - I\right\| \leq \epsilon$ holds for all timesteps.*

Remark A.1 *Recall the following dynamical system defined on W:*

$$\begin{bmatrix} \dot\theta \\ \dot\phi \end{bmatrix} = X(\theta, \phi) - \frac{\alpha}{4}\begin{bmatrix} \theta(I - (\theta^T\theta)^{-1}) \\ \theta(\theta^T\theta)^{-1}(\phi^T\theta(\theta^T\theta)^{-1} + \theta^T\phi) \end{bmatrix}, \quad (23)$$

where

$$\begin{aligned} X(\theta,\phi) =& \begin{bmatrix} X_\theta(\theta,\phi) \\ X_\phi(\theta,\phi) \end{bmatrix} \\ =& \begin{bmatrix} \phi - \theta(\theta^T\theta)^{-1}\,\mathrm{Sym}(\theta^T\phi) \\ \theta(\theta^T\theta)^{-1}((\theta^T\theta)^{-1}\theta^T\phi\,\mathrm{Sym}(\theta^T\phi) - \phi^T\phi) + (D - \theta(\theta^T\theta)^{-1}\,\mathrm{Sym}(\theta^T D))/\eta \end{bmatrix} \end{aligned}$$

with D defined in (2)*, and $\alpha > 0$ is the feedback coefficient.*

In Theorem 4.3, we have shown the exponential stability of the tangent bundle of the Stiefel manifold for the continuous-time dynamical system (23)*. Intuitively, this stability will be carried over to its discretized system. We now analyze in detail the stability of the discretized algorithm, Algorithm 1.*

Recall the update rule of Algorithm 1:

$$\theta \leftarrow \theta + \eta[X_\theta(\theta,\phi) - \frac{\alpha}{4}\theta(I - (\theta^T\theta)^{-1})] \quad (24)$$
$$\phi \leftarrow \phi + \eta[X_\phi(\theta,\phi) - \frac{\alpha}{4}\theta(\theta^T\theta)^{-1}(\phi^T\theta(\theta^T\theta)^{-1} + \theta^T\phi)], \quad (25)$$

where

$$X_\theta(\theta,\phi) = \phi - \theta(\theta^T\theta)^{-1}\,\mathrm{Sym}(\theta^T\phi), \quad (26)$$
$$X_\phi(\theta,\phi) = \theta(\theta^T\theta)^{-1}((\theta^T\theta)^{-1}\theta^T\phi\,\mathrm{Sym}(\theta^T\phi) - \phi^T\phi) + (D - \theta(\theta^T\theta)^{-1}\,\mathrm{Sym}(\theta^T D))/\eta. \quad (27)$$

Proof. Let $\beta > 0$ be a constant to be determined later and we let $\alpha = 4\beta/\eta$ be a parameter depending on the value of η. Define the skew-symmetrization operator as

$$\mathrm{Skew}(B) = B - \mathrm{Sym}(B) = \frac{1}{2}(B - B^T), \tag{28}$$

for any square matrix B.

When we use the approximation $(\theta^T\theta)^{-1} \approx 2I - \theta^T\theta$, at each update step of θ we have

$$(\theta^T\theta - I)_{new} \approx (1-2\beta)(\theta^T\theta - I) + (\beta^2 - 2\beta)(\theta^T\theta - I)^2 + \beta^2(\theta^T\theta - I)^3 + F_\eta, \tag{29}$$

where

$$F_\eta = 2\eta\,\mathrm{Sym}(\tilde{X}_\theta^T\theta) - 2\eta\beta\,\mathrm{Sym}(\tilde{X}_\theta^T\theta(\theta^T\theta - I)) + \eta^2\tilde{X}_\theta^T\tilde{X}_\theta \tag{30}$$

with the approximation of X_θ

$$\tilde{X}_\theta = \phi - \theta(2I - \theta^T\theta)\,\mathrm{Sym}(\theta^T\phi). \tag{31}$$

Since $\|\phi\| \le c$, if we have $\left\|\theta^T\theta - I\right\| \le \epsilon$ at the current timestep, then $\|\theta\| \le c_1$ and $\left\|\tilde{X}_\theta\right\| \le c_2$, for some constants $c_1 = c_1(\epsilon) > 0$ and $c_2 = c_2(\epsilon) > 0$. Therefore, we have

$$\|F_\eta\| \le 2\eta c_1 c_2 + 2\eta\beta\epsilon c_1 c_2 + \eta^2 c_2^2 \tag{32}$$

and

$$\begin{aligned}\left\|(\theta^T\theta - I)_{new}\right\| &\le (1-2\beta)\epsilon + (\beta^2 - 2\beta)\epsilon^2 + \beta^2\epsilon^3 + \|F_\eta\| \\ &= (1 - 2\beta + 2\eta\beta c_1 c_2)\epsilon + (\beta^2 - 2\beta)\epsilon^2 + \beta^2\epsilon^3 + 2\eta c_1 c_2 + \eta^2 c_2^2.\end{aligned} \tag{33}$$

To make $\left\|(\theta^T\theta - I)_{new}\right\| \le \epsilon$, we first choose $\beta = \beta(\eta) > 0$ such that

$$1 - 2\beta + (\beta^2 - 2\beta)\epsilon + \beta^2\epsilon^2 < 1. \tag{34}$$

This is possible since the function $f(\beta) = 1 - 2\beta + (\beta^2 - 2\beta)\epsilon + \beta^2\epsilon^2$ satisfies $f(0) = 1$ and $f'(0) = -2 - 2\epsilon < 0$. In particular, $1 - 2\beta + (\beta^2 - 2\beta)\epsilon + \beta^2\epsilon^2 < 1$ for all $0 < \beta < \frac{2}{\epsilon}$. After we fix β, let $c_3 = c_3(\epsilon) \coloneqq 1 - 2\beta + (\beta^2 - 2\beta)\epsilon + \beta^2\epsilon^2 < 1$, and we choose $\eta = \eta(\epsilon) > 0$ such that

$$2\eta\beta c_1 c_2\epsilon + 2\eta c_1 c_2 + \eta^2 c_2^2 \le (1 - c_3)\epsilon. \tag{35}$$

This is possible since the continuous function $g(\eta) = 2\eta\beta c_1 c_2\epsilon + 2\eta c_1 c_2 + \eta^2 c_2^2$ satisfies $g(0) = 0$ and $(1 - c_3)\epsilon > 0$. In particular, the above inequality holds when

$$0 < \eta \le \frac{\sqrt{c_1^2(\beta\epsilon + 1)^2 + \epsilon(1 - c_3)} - c_1(\beta\epsilon + 1)}{c_2}, \tag{36}$$

completing the proof. ■

Remark A.2 *Without the approximation, at each update step of θ, we have*

$$\theta_{new} = \theta + \eta X_\theta - \beta\theta(I - (\theta^T\theta)^{-1}), \tag{37}$$

which gives

$$\begin{aligned}(\theta^T\theta - I)_{new} =\ & (1-2\beta)(\theta^T\theta - I) + \beta^2(\theta^T\theta + (\theta^T\theta)^{-1} - 2I) \\ & + 2\eta\,\mathrm{Sym}(\theta^T X_\theta) + \eta^2 X_\theta^T X_\theta - 2\eta\beta\,\mathrm{Sym}(X_\theta^T\theta(I - (\theta^T\theta)^{-1})) \\ =\ & (1-2\beta)(\theta^T\theta - I) + \beta^2(\theta^T\theta + (\theta^T\theta)^{-1} - 2I) \\ & + \eta^2 X_\theta^T X_\theta + 2\eta\beta\,\mathrm{Sym}((\theta^T\theta)^{-1}\,\mathrm{Skew}(\theta^T\phi))\end{aligned} \tag{38}$$

At each update step of ϕ, we have

$$\phi_{new} = \phi + \eta X_\phi(\theta, \phi) - \beta\theta(\theta^T\theta)^{-1}(\phi^T\theta(\theta^T\theta)^{-1} + \theta^T\phi), \tag{39}$$

which gives

$$\begin{aligned}[\mathrm{Sym}(\theta^T\phi)]_{new} =\ & (1-2\beta)\,\mathrm{Sym}(\theta^T\phi) + \beta^2(\mathrm{Sym}(\theta^T\phi) - (\theta^T\theta)^{-1}\,\mathrm{Sym}(\theta^T\phi)(\theta^T\theta)^{-1}) \\ & + \eta^2\,\mathrm{Sym}(X_\theta^T X_\phi) - \eta\beta\,\mathrm{Sym}(A),\end{aligned} \tag{40}$$

where

$$\begin{aligned} X_\theta^T X_\phi = & \operatorname{Skew}(\phi^T\theta)(\theta^T\theta)^{-1}((\theta^T\theta)^{-1}\theta^T\phi \operatorname{Sym}(\theta^T\phi) - \phi^T\phi) \\ & + \frac{1}{\eta}(\phi^T D - \phi^T\theta(\theta^T\theta)^{-1}\operatorname{Sym}(\theta^T D) - \operatorname{Sym}(\theta^T\phi)(\theta^T\theta)^{-1}\operatorname{Skew}(\theta^T D)) \end{aligned} \tag{41}$$

and

$$\begin{aligned} A = & \operatorname{Skew}(\phi^T\theta)(\theta^T\theta)^{-1}\phi^T\theta(\theta^T\theta)^{-1} + \operatorname{Skew}(\phi^T\theta)(\theta^T\theta)^{-1}\theta^T\phi \\ & + (I - (\theta^T\theta)^{-1})((\theta^T\theta)^{-1}\theta^T\phi \operatorname{Sym}(\theta^T\phi) - \phi^T\phi). \end{aligned} \tag{42}$$

We also have

$$\begin{aligned} (\phi^T\phi)_{new} = & \phi^T\phi - 2\beta \operatorname{Sym}(\phi^T\theta(\theta^T\theta)^{-1}\phi^T\theta(\theta^T\theta)^{-1}) \\ & + \beta^2(((\theta^T\theta)^{-1}\theta^T\phi + \phi^T\theta)(\theta^T\theta)^{-1}(\phi^T\theta(\theta^T\theta)^{-1} + \theta^T\phi)) \\ & + 2\eta \operatorname{Sym}(\phi^T X_\phi) + \eta^2 X_\phi^T X_\phi \\ & - 2\eta\beta \operatorname{Sym}(X_\phi^T\theta(\theta^T\theta)^{-1}(\phi^T\theta(\theta^T\theta)^{-1} + \theta^T\phi)), \end{aligned} \tag{43}$$

where

$$\begin{aligned} \phi^T X_\phi = & \phi^T\theta(\theta^T\theta)^{-1}((\theta^T\theta)^{-1}\theta^T\phi \operatorname{Sym}(\theta^T\phi) - \phi^T\phi) \\ & + \phi^T(D - \theta(\theta^T\theta)^{-1}\operatorname{Sym}(\theta^T D))/\eta. \end{aligned} \tag{44}$$

and

$$\begin{aligned} X_\phi^T X_\phi = & (\operatorname{Sym}(\theta^T\phi)\phi^T\theta(\theta^T\theta)^{-1} - \phi^T\phi)(\theta^T\theta)^{-1}((\theta^T\theta)^{-1}\theta^T\phi \operatorname{Sym}(\theta^T\phi) - \phi^T\phi) \\ & + \frac{2}{\eta}\operatorname{Sym}((\operatorname{Sym}(\theta^T\phi)\phi^T\theta(\theta^T\theta)^{-1})(\theta^T\theta)^{-1}\operatorname{Skew}(\theta^T D)) \\ & + \frac{1}{\eta^2}(D^T D - 2\operatorname{Sym}(D^T\theta(\theta^T\theta)^{-1}\operatorname{Sym}(\theta^T D)) + \operatorname{Sym}(\theta^T D)(\theta^T\theta)^{-1}\operatorname{Sym}(\theta^T D)). \end{aligned} \tag{45}$$

Inheriting the exponential stability from the continuous-time dynamical system, we can see in both (29) *and* (38) *that the discretized algorithm also has the ability to control the distance from θ to the Stiefel manifold by the term $(1-2\beta)(\theta^T\theta - I)$ with rate $1-2\beta$, even when the approximation is used. So intuitively we may choose β such that $0 < \beta \leq 0.5$, i.e., $0 \leq 1-2\beta < 1$. Specifically, we choose $\beta = 0.4$ in most of our experiments.*

Similarly, we can see in (40) *that the term $(1-2\beta)\operatorname{Sym}(\theta^T\phi)$ pulls $\operatorname{Sym}(\theta^T\phi)$ to 0 with rate $1-2\beta$ and the term $\beta^2(\operatorname{Sym}(\theta^T\phi) - (\theta^T\theta)^{-1}\operatorname{Sym}(\theta^T\phi)(\theta^T\theta)^{-1})$ vanishes if $\theta^T\theta = I$ or $\operatorname{Sym}(\theta^T\phi) = 0$.*

Note that if $\theta^T\theta = I$ and $\operatorname{Sym}(\theta^T\phi) = 0$, then the error terms $X_\theta^T X_\theta = \phi^T\phi$ and $\operatorname{Sym}(\theta^T\theta\operatorname{Skew}(\theta^T\phi)) = 0$ in (38)*. Therefore, the numerical stability practically depends on the value of $\left\|\phi^T\phi\right\|$. Although in our experiments we observe high numerical stability without any further care, we would like to provide a possible way to enhance the numerical stability in the cases when divergence occasionally happens. As $\left\|\phi^T\phi\right\| \leq \|\phi\|^2$, we can clip ϕ if $\|\phi\|$ is large so that $\left\|\phi^T\phi\right\|$ will not be too large. Similar ideas can also be found in a previous work [1] that proposes a gradient norm clipping strategy to deal with exploding gradients. By doing this, the errors $\eta^2 X_\theta^T X_\theta - 2\eta\beta\operatorname{Sym}(\theta^T\theta\operatorname{Skew}(\theta^T\phi)) = O(\eta)$ and the numerical stability will be theoretically guaranteed, as shown in the previous proof.*

B Detailed Experimental Settings

B.1 WideResNet on CIFAR-10/100

The model is trained for 200 epochs in total. When using FGD, the learning rate η is initialized as 0.2 and 0.05 for the parameters with and without orthogonality, respectively. When using other methods, the learning rate η is initialized as 0.1 for all parameters following the original papers. Note that we have also tried other settings of learning rate for other methods, where no improvement of performance is observed. For all methods, the learning rate of each parameter is multiplied by 0.2 at epochs 60, 120, and 160. The momentum coefficient γ is 0.1 for all the methods using momentum in all the experiments. The feedback coefficient α is $1.6/\eta$ depending on the current learning rate. All the other settings are in agreement with the original papers.

B.2 ResNet and VGG on CIFAR-10/100

For FGD, the learning rate η and the feedback coefficient α are the same as in the experiments using WideResNet, cf. Section B.1. All the other settings are in agreement with the original papers. We use ResNet110 in the $9n + 2$ version, where $9n + 2$ means the total depth is the total number of convolutional blocks times 9 plus 2. For each of the models, orthogonality is imposed only on the convolutional layers in the last two residual modules. The parameters of these layers constitute the majority of the whole network. This restriction on the range of parameters to impose orthogonality shows the best performance. We also apply the same restriction when using OCNN for fairness. This restriction also improves the performance of OCNN. Specifically, for all the models except ResNet110, FGD is only applied to the convolutional layers whose input channel number is 256 or 512, i.e., the layers in the last two residual modules, while for ResNet110 we only apply FGD to those with input channel number 32 or 64 that are also the layers in the last two residual modules.

B.3 ResNet and PreActResNet on ImageNet

Each model is trained for 120 epochs in total. For all the methods and all the parameters, the learning rate η is initialized as 0.1, which is multiplied by 0.1 at epochs 30, 60, and 90. The momentum coefficient $\gamma = 0.1$ for all the methods using momentum. The feedback coefficient α is 6 during the whole training process. All the other settings are in agreement with the original papers. For each model, orthogonality is imposed only on the convolutional layers in the last residual module whose output channel number is 512, which shows the best performance.

B.4 Experiments on SVHN

We use ResNet models on the SVHN [2] dataset. The settings are the same as those of the corresponding models in the experiments on CIFAR-10/100, cf. Section B.2. In Table 1, we report the test accuracy rates. Our method consistently outperforms the baseline method SGD with momentum in terms of accuracy.

Method	Res18	Rest34	Res50	Res101
SGD*	96.53	96.54	96.68	96.86
FGD*(ours)	**96.88**	**96.91**	**96.97**	**97.12**

Table 1: Test accuracy rates using ResNet on SVHN.

C Code

```
if p.grad is None:
    continue
d_p = p.grad.data
if not stiefel:  # original procedure
    if weight_decay != 0:
        d_p = d_p.add(p, alpha=weight_decay)
    if momentum != 0:
        param_state = self.state[p]
        if 'momentum_buffer' not in param_state:
            buf = param_state['momentum_buffer'] = torch.clone(d_p).detach()
        else:
            buf = param_state['momentum_buffer']
            buf.mul_(momentum).add_(d_p, alpha=1 - dampening)
        if nesterov:
            d_p = d_p.add(buf, alpha=momentum)
```

```
            else:
                d_p = buf
        p.add_(d_p, alpha=-baselr)
    else:
        # no weight decay
        p_2d = p.data.view(p.shape[0], -1)
        # d_p_2d = d_p.view_as(p_2d)
        eye_p_2d = torch.eye(p_2d.shape[0], device=p.device)
        inverse_approx_2d = eye_p_2d.mul(2).sub(p_2d.mm(p_2d.t()))
        lr = baselr * stiefel
        if momentum != 0:
            param_state = self.state[p]
            if 'momentum_buffer' not in param_state:  # v0
                buf = param_state['momentum_buffer'] = torch.zeros_like(p.data)
                buf.add_(d_p)
                buf_2d = buf.view_as(p_2d)
                q = buf_2d.mm(p_2d.t())
                q = q.add(q.t()).mul(0.5)
                buf_2d.sub_(q.mm(inverse_approx_2d).mm(p_2d))
                buf.mul_(-lr)
            else:
                buf = param_state['momentum_buffer']
                buf_2d = buf.view_as(p_2d)

                # C
                con = torch.zeros_like(p.data)
                con_2d = con.view_as(p_2d)
                con_2d.sub_(buf_2d.mm(buf_2d.t()).mm(inverse_approx_2d).mm(p_2d))

                # D
                rd = torch.zeros_like(p.data)
                rd_2d = rd.view_as(p_2d)
                rd.add_(buf, alpha=momentum - 1).add_(d_p, alpha=-lr)
                q = rd_2d.mm(p_2d.t())
                rd_2d.sub_(q.add(q.t()).mul(0.5).mm(inverse_approx_2d).mm(p_2d))

                # E
                ext = torch.zeros_like(p.data)
                ext_2d = ext.view_as(p_2d)
                q = buf_2d.mm(p_2d.t())
                ext_2d.add_(q.add(q.t()).mul(0.5).mm(q).mm(inverse_approx_2d)
                            .mm(inverse_approx_2d).mm(p_2d))

                # F_phi
                fb = torch.zeros_like(p.data)
                fb_2d = fb.view_as(p_2d)
                if feedback:
                    fb_2d.sub_((inverse_approx_2d.mm(p_2d).mm(buf_2d.t()) +
                                buf_2d.mm(p_2d.t())).mm(inverse_approx_2d).mm(p_2d),
                               alpha=feedback)

                buf.add_(con).add_(rd).add_(ext).add_(fb)

            # no Nesterov
            d_p = buf

            # f_tilde
            f_phi = torch.zeros_like(p.data)
            f_phi_2d = f_phi.view_as(p_2d)
```

```
            f_phi.add_(d_p)
            q = f_phi_2d.mm(p_2d.t())
            f_phi_2d.sub_(q.add(q.t()).mul(0.5)
                          .mm(inverse_approx_2d).mm(p_2d))

            # F_theta
            fb = torch.zeros_like(p.data)
            fb_2d = fb.view_as(p_2d)
            if feedback:
                fb_2d.sub_(p_2d.sub(inverse_approx_2d.mm(p_2d)),
                           alpha=feedback)
            p.data.add_(d_p).add_(fb)
        else:  # no momentum
            d_p.mul_(-lr)
            d_p_2d = d_p.view_as(p_2d)
            q = d_p_2d.mm(p_2d)
            q = q.add(q.t()).mul(0.5)
            d_p_2d.sub_(q.mm(inverse_approx_2d).mm(p_2d))
            if feedback:
                d_p_2d.sub_(p_2d.mm(p_2d.t()).mm(p_2d).sub(p_2d),
                            alpha=feedback)
            p.data.add_(d_p)
```

References

[1] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks,” in *International conference on machine learning*, pp. 1310–1318, PMLR, 2013.

[2] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, “Reading digits in natural images with unsupervised feature learning,” 2011.