Supplementary Material:
Riemannian Stochastic Recursive Gradient Algorithm

Hiroyuki Kasai [1] **Hiroyuki Sato** [2] **Bamdev Mishra** [3]

In this supplementary material, we present overviews of the manifolds of interest, a complete proof of the convergence analysis, and details of additional numerical experiments. Hereafter, we use $\mathbb{E}[\cdot]$ to indicate the expectation with respect to the joint distribution of all random variables. For example, w_t is determined by the realizations of the independent random variables $\{i_1, i_2, \dots, i_{t-1}\}$, and the total expectation of $f(w_t)$ for any $t \in \mathbb{N}$ can be taken as $\mathbb{E}[f(w_t)] = \mathbb{E}_{i_1}\mathbb{E}_{i_2}\dots\mathbb{E}_{i_{t-1}}[f(w_t)]$. We also use $\mathbb{E}[\cdot|\mathcal{F}_t]$ to denote an expected value taken with respect to the distribution of the random variable i_t.

A SPD manifold and Grassmann manifold

A.1 SPD manifold $\mathcal{S}^d_{++}$

We designate the space of $d \times d$ SPD matrices as the SPD manifold, $\mathcal{S}^d_{++}$. If we endow $\mathcal{S}^d_{++}$ with the affine-invariant Riemannian metric (AIRM) (Pennec et al., 2006) defined by $\langle \xi_{\mathbf{X}}, \eta_{\mathbf{X}} \rangle_{\mathbf{X}} = \mathrm{trace}(\xi_{\mathbf{X}}\mathbf{X}^{-1}\eta_{\mathbf{X}}\mathbf{X}^{-1})$ for $\xi_{\mathbf{X}}, \eta_{\mathbf{X}} \in T_{\mathbf{X}}\mathcal{S}^d_{++}$ at $\mathbf{X} \in \mathcal{S}^d_{++}$, the SPD manifold $\mathcal{S}^d_{++}$ forms a Riemannian manifold. An efficient retraction is proposed as follows (Jeuris et al., 2012): $R_{\mathbf{X}}(\xi_{\mathbf{X}}) = \mathbf{X} + \xi_{\mathbf{X}} + \frac{1}{2}\xi_{\mathbf{X}}\mathbf{X}^{-1}\xi_{\mathbf{X}}$. This maps $\xi_{\mathbf{X}}$ onto $\mathcal{S}^d_{++}$ for all $\xi_{\mathbf{X}} \in T_{\mathbf{X}}\mathcal{S}^d_{++}$. Previously, Huang et al. proposed an efficient isometric vector transport (Huang et al., 2015b; Yuan et al., 2016) defined as $\mathcal{T}_{S_\eta}\xi_{\mathbf{X}} = B_{\mathbf{Y}}B^{\flat}_{\mathbf{X}}\xi_{\mathbf{X}}$, where $\mathbf{Y} = R_{\mathbf{X}}(\xi_{\mathbf{X}})$ and $a^\flat$ denotes the flat of $a \in T_w\mathcal{M}$; i.e., $a^\flat : T_w\mathcal{M} \to \mathbb{R} : v \mapsto \langle a, v \rangle_w$. $B_{\mathbf{X}}$ and $B_{\mathbf{Y}}$ are the orthonormal bases of $T_{\mathbf{X}}\mathcal{S}^d_{++}$ and $T_{\mathbf{Y}}\mathcal{S}^d_{++}$, respectively, where the basis is calculated based on the Cholesky decomposition.

A.2 Grassmann manifold $\mathrm{Gr}(r, d)$

A point on the Grassmann manifold is an equivalence class represented by a $d \times r$ orthogonal matrix $\mathbf{U}$ with orthonormal columns: $\mathbf{U}^T\mathbf{U} = \mathbf{I}$. Two orthogonal matrices represent the same element on the Grassmann manifold if they can be transformed into each other by right multiplication of an $r \times r$ orthogonal matrix $\mathbf{O} \in \mathcal{O}(r)$. We state that these two matrices are equivalent. In other words, an element of $\mathrm{Gr}(r, d)$ is identified with a set of equivalent $d \times r$ orthogonal matrices $[\mathbf{U}] := \{\mathbf{UO} : \mathbf{O} \in \mathcal{O}(r)\}$. That is, $\mathrm{Gr}(r, d) := \mathrm{St}(r, d)/\mathcal{O}(r)$, where $\mathrm{St}(r, d)$ is the *Stiefel manifold*, which is the set of matrices of size $d \times r$ with orthonormal columns. The Grassmann manifold has the structure of a Riemannian quotient manifold (Absil et al., 2008). The retraction $R_{\mathbf{U}(0)}(\xi) = \mathrm{qf}(\mathbf{U}(0) + t\xi)(=: \mathbf{U}(t))$, which extracts the orthonormal factor based on QR decomposition, is widely used. Further, a commonly used vector transport employs an orthogonal projection of $t\xi$ to the horizontal space at $\mathbf{U}(t)$, i.e., $(\mathbf{I} - \mathbf{U}(t)\mathbf{U}(t)^T)t\xi$ (Absil et al., 2008).

[1] The University of Electro-Communications, Japan. [2] Kyoto University, Japan. [3] Microsoft, India. Correspondence to: Hiroyuki Kasai <kasai@is.uec.ac.jp>.

Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

B Proofs

B.1 Essential lemmas

B.1.1 Proof of Lemmas 3.5 and 3.6

Taylor's theorem is generalized to Riemannian manifolds (Absil et al., 2008) and addresses the exponential mapping instead of the retraction. Lemma 3.2 in (Huang et al., 2015b) applies Taylor's theorem on the retraction by newly introducing a function along a curve on the manifold.

Proof. For self-completeness, we show a proof similar to that of Lemma 3.2 in (Huang et al., 2015b). We define $\xi = \alpha\eta$ with $\eta = \xi/\|\xi\|_w$, i.e., $\alpha = \|\xi\|_w$. From Taylor's theorem, we have

$$\begin{aligned} f(z) - f(w) &= f(R_w(\alpha\eta)) - f(R_w(0)) \\ &= \frac{d}{d\tau} f(R_w(\tau\eta))\Big|_{\tau=0} \cdot \alpha + \frac{1}{2}\frac{d^2}{d\tau^2} f(R_w(\tau\eta))\Big|_{\tau=p} \cdot \alpha^2 \\ &= \langle \mathrm{grad} f(w), \alpha\eta \rangle_w + \frac{1}{2}\frac{d^2}{d\tau^2} f(R_w(\tau\eta))\Big|_{\tau=p} \cdot \alpha^2 \\ &\leq \langle \mathrm{grad} f(w), \xi \rangle_w + \frac{1}{2} L \|\xi\|_w^2, \end{aligned}$$

where $0 \leq p \leq \alpha$. This completes the proof of Lemma 3.5. Lemma 3.6 can be proved in a similar manner. □

B.1.2 Proof of Lemma 3.8

Proof. From Lemma 8 in (Huang et al., 2015a), we have

$$\left\| P(\gamma)_z^w \mathrm{grad} f(z) - \mathrm{grad} f(w) - \int_0^1 P(\gamma)_{\gamma(t)}^w \mathrm{Hess} f(\gamma(t)) P(\gamma)_w^{\gamma(t)} \eta dt \right\|_w \leq b_0 \|\eta\|_w^2,$$

where b_0 is identical to $C_h\theta$ by applying Assumption (1.4) and Lemma 3.7 for Lemma 8 in (Huang et al., 2015a). Hence, from the triangle inequality, we have

$$\begin{aligned} \|P(\gamma)_z^w \mathrm{grad} f(z) - \mathrm{grad} f(w)\|_w &\leq \left\| \int_0^1 P(\gamma)_{\gamma(t)}^w \mathrm{Hess} f(\gamma(t)) P(\gamma)_w^{\gamma(t)} \eta dt \right\|_w + C_h\theta \|\eta\|_w^2 \\ &\leq \int_0^1 \|P(\gamma)_{\gamma(t)}^w \mathrm{Hess} f(\gamma(t)) P(\gamma)_w^{\gamma(t)} \eta\|_w dt + C_h\theta\|\eta\|_w^2 \\ &\leq C_h(1 + C_\eta\theta)\|\eta\|_w. \end{aligned}$$

As $L_l := C_h(1 + C_\eta\theta)$, this completes the proof. □

Note that C_η is uniformly determined in each algorithm. As for Algorithm 1, the constant L_l is derived as $L_l = C_h(1 + 3\theta C_g)$ for any $t > 1$, because the triangle inequality yields the following, from (2):

$$\begin{aligned} \|v_t\|_{w_t} &= \|\mathrm{grad} f_{i_t}(w_t) - \mathcal{T}_{w_{t-1}}^{w_t} \mathrm{grad} f_{i_t}(w_{t-1}) + \mathcal{T}_{w_{t-1}}^{w_t} v_{t-1}\|_{w_t} \\ &\leq C_g + C_g + C_g \ = \ 3C_g. \end{aligned}$$

That is, C_η can be chosen as $3C_g$. Here, we used the property that $\mathcal{T}$ is an isometry in Assumption 1.

B.1.3 Lemma B.1

Lemma B.1. *Suppose that Assumption 1 holds and f is upper-Hessian bounded. Let w^* be an optimal solution to problem (1). Consider Algorithm 1 with a constant step size α. Then, we have*

$$\sum_{t=0}^{m} \mathbb{E}[\|\mathrm{grad} f(w_t)\|_{w_t}^2] \ \leq \ \frac{2}{\alpha}\mathbb{E}[f(w_0) - f(w^*)] + \sum_{t=0}^{m} \mathbb{E}[\|\mathrm{grad} f(w_t) - v_t\|_{w_t}^2] - (1 - L\alpha)\sum_{t=0}^{m} \mathbb{E}[\|v_t\|_{w_t}^2]. \quad \text{(A.1)}$$

Proof. This proof is the straightforward extension to the Riemannian setting from that of Lemma 1 in (Nguyen et al., 2017a). From Lemma 3.5, we have

$$\begin{aligned}
\mathbb{E}[f(w_{t+1})] &\leq \mathbb{E}[f(w_t)] - \mathbb{E}[\langle \mathrm{grad} f(w_t), \alpha v_t \rangle_{w_t}] + \frac{1}{2} L\alpha^2 \mathbb{E}[\|v_t\|^2_{w_t}] \\
&= \mathbb{E}[f(w_t)] - \frac{\alpha}{2}\mathbb{E}\left[\|\mathrm{grad} f(w_t)\|^2_{w_t} + \|v_t\|^2_{w_t} - \|\mathrm{grad} f(w_t) - v_t\|^2_{w_t}\right] + \frac{1}{2} L\alpha^2 \mathbb{E}[\|v_t\|^2_{w_t}] \\
&= \mathbb{E}[f(w_t)] - \frac{\alpha}{2}\mathbb{E}[\|\mathrm{grad} f(w_t)\|^2_{w_t}] + \frac{\alpha}{2}\mathbb{E}[\|\mathrm{grad} f(w_t) - v_t\|^2_{w_t}] - \left(\frac{\alpha}{2} - \frac{1}{2}L\alpha^2\right)\mathbb{E}[\|v_t\|^2_{w_t}].
\end{aligned}$$

Let w_{m+1} be a point obtained by performing Steps 8–10 in Algorithm 1 for $t = m$. Summing over $t = 0, 1, \ldots, m$ yields

$$\begin{aligned}
\mathbb{E}[f(w_{m+1})] &\leq \mathbb{E}[f(w_0)] - \frac{\alpha}{2}\sum_{t=0}^{m}\mathbb{E}[\|\mathrm{grad} f(w_t)\|^2_{w_t}] \\
&\qquad + \frac{\alpha}{2}\sum_{t=0}^{m}\mathbb{E}[\|\mathrm{grad} f(w_t) - v_t\|^2_{w_t}] - \left(\frac{\alpha}{2} - \frac{1}{2}L\alpha^2\right)\sum_{t=0}^{m}\mathbb{E}[\|v_t\|^2_{w_t}].
\end{aligned}$$

Then, we obtain

$$\begin{aligned}
\sum_{t=0}^{m}\mathbb{E}[\|\mathrm{grad} f(w_t)\|^2_{w_t}] &\leq \frac{2}{\alpha}\mathbb{E}[f(w_0) - f(w_{m+1})] + \sum_{t=0}^{m}\mathbb{E}[\|\mathrm{grad} f(w_t) - v_t\|^2_{w_t}] - (1 - L\alpha)\sum_{t=0}^{m}\mathbb{E}[\|v_t\|^2_{w_t}] \\
&\leq \frac{2}{\alpha}\mathbb{E}[f(w_0) - f(w^*)] + \sum_{t=0}^{m}\mathbb{E}[\|\mathrm{grad} f(w_t) - v_t\|^2_{w_t}] - (1 - L\alpha)\sum_{t=0}^{m}\mathbb{E}[\|v_t\|^2_{w_t}],
\end{aligned}$$

where the last inequality holds because w^* is a solution of f. This completes the proof. □

B.1.4 Lemma B.2

Lemma B.2. *Suppose that Assumption 1 holds. Consider v_t in Algorithm 1. Then, for any $t \geq 1$,*

$$\mathbb{E}[\|\mathrm{grad} f(w_t) - v_t\|^2_{w_t}] = \sum_{j=1}^{t}\mathbb{E}[\|v_j - \mathcal{T}^{w_j}_{w_{j-1}} v_{j-1}\|^2_{w_j}] - \sum_{j=1}^{t}\mathbb{E}[\|\mathrm{grad} f(w_j) - \mathcal{T}^{w_j}_{w_{j-1}}\mathrm{grad} f(w_{j-1})\|^2_{w_j}].$$

Proof. This proof is the straightforward extension to the Riemannian setting from that of Lemma 2 in (Nguyen et al., 2017a). First, we obtain the expectation of $v_j - \mathcal{T}^{w_j}_{w_{j-1}} v_{j-1}$ as

$$\begin{aligned}
\mathbb{E}[v_j - \mathcal{T}^{w_j}_{w_{j-1}} v_{j-1} | \mathcal{F}_j] &= \mathbb{E}[\mathrm{grad} f_{i_j}(w_j) - \mathcal{T}^{w_j}_{w_{j-1}}\mathrm{grad} f_{i_j}(w_{j-1}) | \mathcal{F}_j] \\
&= \mathrm{grad} f(w_j) - \mathcal{T}^{w_j}_{w_{j-1}}\mathrm{grad} f(w_{j-1}). \qquad \text{(A.2)}
\end{aligned}$$

Then, we have

$$\begin{aligned}
\mathbb{E}[\|\mathrm{grad} f(w_j) - v_j\|^2_{w_j} | \mathcal{F}_j] &= \mathbb{E}[\|[\mathcal{T}^{w_j}_{w_{j-1}}\mathrm{grad} f(w_{j-1}) - \mathcal{T}^{w_j}_{w_{j-1}} v_{j-1}] + \mathrm{grad} f(w_j) - \mathcal{T}^{w_j}_{w_{j-1}}\mathrm{grad} f(w_{j-1}) \\
&\qquad - [v_j - \mathcal{T}^{w_j}_{w_{j-1}} v_{j-1}]\|^2_{w_j} | \mathcal{F}_j] \\
&= \|\mathcal{T}^{w_j}_{w_{j-1}}\mathrm{grad} f(w_{j-1}) - \mathcal{T}^{w_j}_{w_{j-1}} v_{j-1}\|^2_{w_j} + \|\mathrm{grad} f(w_j) - \mathcal{T}^{w_j}_{w_{j-1}}\mathrm{grad} f(w_{j-1})\|^2_{w_j} \\
&\qquad + \mathbb{E}[\|v_j - \mathcal{T}^{w_j}_{w_{j-1}} v_{j-1}\|^2_{w_j} | \mathcal{F}_j] \\
&\qquad + 2\langle \mathcal{T}^{w_j}_{w_{j-1}}\mathrm{grad} f(w_{j-1}) - \mathcal{T}^{w_j}_{w_{j-1}} v_{j-1}, \mathrm{grad} f(w_j) - \mathcal{T}^{w_j}_{w_{j-1}}\mathrm{grad} f(w_{j-1})\rangle_{w_j} \\
&\qquad - 2\langle \mathcal{T}^{w_j}_{w_{j-1}}\mathrm{grad} f(w_{j-1}) - \mathcal{T}^{w_j}_{w_{j-1}} v_{j-1}, \mathbb{E}[v_j - \mathcal{T}^{w_j}_{w_{j-1}} v_{j-1} | \mathcal{F}_j]\rangle_{w_j} \\
&\qquad - 2\langle \mathrm{grad} f(w_j) - \mathcal{T}^{w_j}_{w_{j-1}}\mathrm{grad} f(w_{j-1}), \mathbb{E}[v_j - \mathcal{T}^{w_j}_{w_{j-1}} v_{j-1} | \mathcal{F}_j]\rangle_{w_j} \\
&\overset{(A.2)}{=} \|\mathcal{T}^{w_j}_{w_{j-1}}\mathrm{grad} f(w_{j-1}) - \mathcal{T}^{w_j}_{w_{j-1}} v_{j-1}\|^2_{w_j} - \|\mathrm{grad} f(w_j) - \mathcal{T}^{w_j}_{w_{j-1}}\mathrm{grad} f(w_{j-1})\|^2_{w_j} \\
&\qquad + \mathbb{E}[\|v_j - \mathcal{T}^{w_j}_{w_{j-1}} v_{j-1}\|^2_{w_j} | \mathcal{F}_j].
\end{aligned}$$

Taking the total expectation for the above, we obtain

$$\begin{aligned}\mathbb{E}[\|\mathrm{grad}f(w_j)-v_j\|^2_{w_j}] &= \mathbb{E}[\|\mathcal{T}^{w_j}_{w_{j-1}}\mathrm{grad}f(w_{j-1})-\mathcal{T}^{w_j}_{w_{j-1}}v_{j-1}\|^2_{w_j}]-\mathbb{E}[\|\mathrm{grad}f(w_j)-\mathcal{T}^{w_j}_{w_{j-1}}\mathrm{grad}f(w_{j-1})\|^2_{w_j}]\\ &\quad +\mathbb{E}[\|v_j-\mathcal{T}^{w_j}_{w_{j-1}}v_{j-1}\|^2_{w_j}].\end{aligned}$$

As $\|\mathrm{grad}f(w_0)-v_0\|^2_{w_j}=0$ and $\|\mathcal{T}^{w_j}_{w_{j-1}}\mathrm{grad}f(w_{j-1})-\mathcal{T}^{w_j}_{w_{j-1}}v_{j-1}\|_{w_j}=\|\mathrm{grad}f(w_{j-1})-v_{j-1}\|_{w_j}$, summing over $j=1,2,\ldots,t$ yields

$$\mathbb{E}[\|\mathrm{grad}f(w_t)-v_t\|^2_{w_t}] = \sum_{j=1}^{t}\mathbb{E}[\|v_j-\mathcal{T}^{w_j}_{w_{j-1}}v_{j-1}\|^2_{w_j}]-\sum_{j=1}^{t}\mathbb{E}[\|\mathrm{grad}f(w_j)-\mathcal{T}^{w_j}_{w_{j-1}}\mathrm{grad}f(w_{j-1})\|^2_{w_j}]. \quad \text{(A.3)}$$

This completes the proof. □

B.2 Proofs of retraction-convex functions

B.2.1 Assumptions B.3 and B.4 and Lemma B.5

In this subsubsection, we first state an assumption.

Assumption B.3. *For all $w,z\in\mathcal{U}$, we have*

$$\frac{1}{L}\|P^z_w\mathrm{grad}f_i(w)-\mathrm{grad}f_i(z)\|^2_z \leq \langle P^z_w\mathrm{grad}f_i(w)-\mathrm{grad}f_i(z),\mathrm{Exp}^{-1}_z(w)\rangle_z,\quad i=1,2,\ldots,n, \quad \text{(A.4)}$$

where L is in Definition 3.1 and $P^z_w(\cdot)$ is a parallel translation operator along the retraction curve from w to z.

Note that (A.4) is equivalent to the condition that f is L-smooth and convex in the Euclidean setting. Note also that if the two L in the above equation and in Lemma 3.5 are different, we can newly define the maximum of the two values as L.

Assumption B.4. *There exists a constant $a_1>0$ such that, for any $w,z\in\mathcal{U}\subset\mathcal{M}$, $\xi=R^{-1}_w(z)$, $\eta=\mathrm{Exp}^{-1}_w(z)$, it holds*

$$\|P(\gamma_R)^z_w\chi-P(\gamma_g)^z_w\chi\|_z\leq a_1\|\xi\|_w\|\chi\|_w,\quad \chi\in T_w\mathcal{M},$$

where $\gamma_R(t):=R_w(t\xi)$ and $\gamma_g(t):=\mathrm{Exp}_w(t\eta)$. Furthermore, this a_1 is sufficiently small.

Lemma B.5. *Let R be a retraction on $\mathcal{M}$. Suppose that f is lower-Hessian bounded and that Assumptions 1 and B.4 hold. Then, there exists a constant $a_0>0$ such that, for all $w,z\in\mathcal{U}$, we have*

$$(a_0\mu-a_1C_g)\|R^{-1}_w(z)\|_w \leq \|\mathrm{grad}f(w)-P^w_z\mathrm{grad}f(z)\|_w, \quad \text{(A.5)}$$

where μ is in Definition 3.2 and $P^w_z(\cdot)$ is a parallel translation operator along the curve defined by R from z to w.

Proof. Considering the exponential mapping case in Lemma 3.6, we obtain the two inequalities as

$$\begin{aligned}f(z) &\geq f(w)+\langle\mathrm{grad}f(w),\mathrm{Exp}^{-1}_w(z)\rangle_w+\frac{\mu}{2}\|\mathrm{Exp}^{-1}_w(z)\|^2_w,\\ f(w) &\geq f(z)+\langle\mathrm{grad}f(z),\mathrm{Exp}^{-1}_z(w)\rangle_z+\frac{\mu}{2}\|\mathrm{Exp}^{-1}_z(w)\|^2_z.\end{aligned}$$

Hence, adding and rearranging the two inequalities yields

$$\begin{aligned}\mu\|\mathrm{Exp}^{-1}_w(z)\|^2_w &\leq -\langle\mathrm{grad}f(w),\mathrm{Exp}^{-1}_w(z)\rangle_w-\langle\mathrm{grad}f(z),\mathrm{Exp}^{-1}_z(w)\rangle_z\\ &= -\langle\mathrm{grad}f(w),\mathrm{Exp}^{-1}_w(z)\rangle_w-\langle P(\gamma)^w_z\mathrm{grad}f(z),P(\gamma)^w_z\mathrm{Exp}^{-1}_z(w)\rangle_w\\ &= -\langle\mathrm{grad}f(w),\mathrm{Exp}^{-1}_w(z)\rangle_w-\langle P(\gamma)^w_z\mathrm{grad}f(z),-\mathrm{Exp}^{-1}_w(z)\rangle_w\\ &= -\langle\mathrm{grad}f(w)-P(\gamma)^w_z\mathrm{grad}f(z),\mathrm{Exp}^{-1}_w(z)\rangle_w\\ &\leq \|\mathrm{grad}f(w)-P(\gamma)^w_z\mathrm{grad}f(z)\|_w\|\mathrm{Exp}^{-1}_w(z)\|_w,\end{aligned}$$

where $P(\gamma)$ is the parallel translation along the geodesic γ and $P(\gamma)^w_z\mathrm{Exp}^{-1}_z(w)=-\mathrm{Exp}^{-1}_w(z)$ is incorporated in the second equality. The last inequality incorporates the Cauchy-Schwarz inequality. Here, according to Lemma 3 in (Huang et al., 2015a) for a constant $a_0>0$, we have

$$a_0\|R^{-1}_w(z)\|_w \leq \|\mathrm{Exp}^{-1}_w(z)\|_w.$$

Furthermore, we can evaluate $\|\mathrm{grad}f(w) - P(\gamma)_z^w \mathrm{grad}f(z)\|_w$ by using Assumption B.4 as

$$\begin{array}{rcl} \|\mathrm{grad}f(w) - P(\gamma)_z^w \mathrm{grad}f(z)\|_w & = & \|\mathrm{grad}f(w) - P_z^w \mathrm{grad}f(z) + P_z^w \mathrm{grad}f(z) - P(\gamma)_z^w \mathrm{grad}f(z)\|_w \\ & \leq & \|\mathrm{grad}f(w) - P_z^w \mathrm{grad}f(z)\|_w + \|(P_z^w - P(\gamma)_z^w)\mathrm{grad}f(z)\|_w \\ & \leq & \|\mathrm{grad}f(w) - P_z^w \mathrm{grad}f(z)\|_w + a_1 C_g \|R_w^{-1}(z)\|_w. \end{array}$$

Considering these three inequalities, we obtain the desired result. This completes the proof. □

B.2.2 Proof of Lemma 3.9

This subsection provides the proof of Lemma 3.9.

Proof. We have

$$\begin{array}{rcl} \langle \mathrm{Exp}_w^{-1}(z), \xi\rangle_w - \langle R_w^{-1}(z), \xi\rangle_w & = & \langle \mathrm{Exp}_w^{-1}(z) - R_w^{-1}(z), \xi\rangle_w \\ & \leq & \|\mathrm{Exp}_w^{-1}(z) - R_w^{-1}(z)\|_w \|\xi\|_w \\ & \leq & 2c_R C_g \|R_w^{-1}(z)\|_w^2, \end{array} \tag{A.6}$$

where the last inequality incorporates Assumption (1.6). Defining $2c_R C_g$ as ν gives the desired result. This ends the proof. □

Note that, when the retraction is close to the exponential mapping $\mathrm{Exp}_w(z)$, ν becomes close to zero.

B.2.3 Lemma B.6

Lemma B.6. *Suppose that Assumptions 1 and B.3 hold and f is upper-Hessian bounded. Consider v_t in Algorithm 1 with a constant step size $\alpha > 0$. Then, for any $t \geq 1$,*

$$\sum_{t=0}^{m} \mathbb{E}[\|\mathrm{grad}f(w_t) - v_t\|_{w_t}^2] \quad \leq \quad \frac{m\alpha L}{2-\alpha L}\mathbb{E}[\|v_0\|_{w_0}^2] + \psi(\alpha)\sum_{t=0}^{m}\sum_{j=1}^{t}\mathbb{E}[\|v_{j-1}\|_{w_{j-1}}^2],$$

where $\psi(\alpha) = \dfrac{2(2L_l + 2\theta C_g + L)\theta C_g \alpha^2}{2-\alpha L}$.

Proof. The expectation of the bound of the norm of v_t is first derived.

$$\begin{array}{rcl} \mathbb{E}[\|v_t\|_{w_t}^2|\mathcal{F}_t] & = & \mathbb{E}[\|\mathcal{T}_{w_{t-1}}^{w_t} v_{t-1} - (\mathcal{T}_{w_{t-1}}^{w_t}\mathrm{grad}f_{i_t}(w_{t-1}) - \mathrm{grad}f_{i_t}(w_t))\|_{w_t}^2|\mathcal{F}_t] \\ & = & \|\mathcal{T}_{w_{t-1}}^{w_t} v_{t-1}\|_{w_t}^2 + \mathbb{E}[\|\mathcal{T}_{w_{t-1}}^{w_t}\mathrm{grad}f_{i_t}(w_{t-1}) - \mathrm{grad}f_{i_t}(w_t))\|_{w_t}^2|\mathcal{F}_t] \\ & & -2\mathbb{E}[\langle \mathcal{T}_{w_{t-1}}^{w_t}\mathrm{grad}f_{i_t}(w_{t-1}) - \mathrm{grad}f_{i_t}(w_t), \mathcal{T}_{w_{t-1}}^{w_t} v_{t-1}\rangle_{w_t}|\mathcal{F}_t]. \end{array} \tag{A.7}$$

The third term in (A.7) is evaluated as

$$
\begin{aligned}
& -2\mathbb{E}[\langle \mathcal{T}^{w_t}_{w_{t-1}} \mathrm{grad} f_{i_t}(w_{t-1}) - \mathrm{grad} f_{i_t}(w_t), \mathcal{T}^{w_t}_{w_{t-1}} v_{t-1} \rangle_{w_t} | \mathcal{F}_t] \\
= \quad & -2\mathbb{E}[\langle \mathrm{grad} f_{i_t}(w_{t-1}) - (\mathcal{T}^{w_t}_{w_{t-1}})^{-1} \mathrm{grad} f_{i_t}(w_t), v_{t-1} \rangle_{w_{t-1}} | \mathcal{F}_t] \\
= \quad & -2\mathbb{E}[\langle \mathrm{grad} f_{i_t}(w_{t-1}) - P^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t) + P^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t) - (\mathcal{T}^{w_t}_{w_{t-1}})^{-1} \mathrm{grad} f_{i_t}(w_t), v_{t-1} \rangle_{w_{t-1}} | \mathcal{F}_t] \\
= \quad & -\frac{2}{\alpha}\mathbb{E}[\langle \mathrm{grad} f_{i_t}(w_{t-1}) - P^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t), \alpha v_{t-1} \rangle_{w_{t-1}} | \mathcal{F}_t] \\
& -2\mathbb{E}[\langle P^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t) - (\mathcal{T}^{w_t}_{w_{t-1}})^{-1} \mathrm{grad} f_{i_t}(w_t), v_{t-1} \rangle_{w_{t-1}} | \mathcal{F}_t] \\
= \quad & -\frac{2}{\alpha}\mathbb{E}[\langle -\mathrm{grad} f_{i_t}(w_{t-1}) + P^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t), -\alpha v_{t-1} \rangle_{w_{t-1}} | \mathcal{F}_t] \\
& -2\mathbb{E}[\langle P^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t) - (\mathcal{T}^{w_t}_{w_{t-1}})^{-1} \mathrm{grad} f_{i_t}(w_t), v_{t-1} \rangle_{w_{t-1}} | \mathcal{F}_t] \\
\overset{(A.4)}{\leq} \quad & -\frac{2}{\alpha}\mathbb{E}\left[\frac{1}{L}\| \mathrm{grad} f_{i_t}(w_{t-1}) - P^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t)\|^2_{w_{t-1}} - \nu \|\alpha v_{t-1}\|^2_{w_{t-1}} | \mathcal{F}_t\right] \\
& -2\mathbb{E}[\langle P^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t) - (\mathcal{T}^{w_t}_{w_{t-1}})^{-1} \mathrm{grad} f_{i_t}(w_t), v_{t-1} \rangle_{w_{t-1}} | \mathcal{F}_t] \\
= \quad & -\frac{2}{\alpha L}\mathbb{E}[\| \mathrm{grad} f_{i_t}(w_{t-1}) - P^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t)\|^2_{w_{t-1}} | \mathcal{F}_t] + 2\nu\alpha \|v_{t-1}\|^2_{w_{t-1}} \\
& -2\mathbb{E}[\langle P^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t) - (\mathcal{T}^{w_t}_{w_{t-1}})^{-1} \mathrm{grad} f_{i_t}(w_t), v_{t-1} \rangle_{w_{t-1}} | \mathcal{F}_t],
\end{aligned}
\tag{A.8}
$$

where the inequality incorporates Lemma 3.9. We now proceed to evaluate the first and third terms in (A.8) separately. The first term in (A.8) is further calculated as

$$
\begin{aligned}
& -\frac{2}{\alpha L}\mathbb{E}[\| \mathrm{grad} f_{i_t}(w_{t-1}) - P^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t)\|^2_{w_{t-1}} | \mathcal{F}_t] \\
= \quad & -\frac{2}{\alpha L}\mathbb{E}[\| \mathrm{grad} f_{i_t}(w_{t-1}) - \mathcal{T}^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t) - P^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t) + \mathcal{T}^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t)\|^2_{w_{t-1}} | \mathcal{F}_t] \\
= \quad & -\frac{2}{\alpha L}\mathbb{E}[\| \mathrm{grad} f_{i_t}(w_{t-1}) - \mathcal{T}^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t)\|^2_{w_{t-1}} | \mathcal{F}_t] \\
& -\frac{2}{\alpha L}\mathbb{E}[\| P^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t) - \mathcal{T}^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t)\|^2_{w_{t-1}} | \mathcal{F}_t] \\
& +\frac{4}{\alpha L}\mathbb{E}[\langle \mathrm{grad} f_{i_t}(w_{t-1}) - \mathcal{T}^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t), P^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t) - \mathcal{T}^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t) \rangle_{w_{t-1}} | \mathcal{F}_t] \\
\leq \quad & -\frac{2}{\alpha L}\mathbb{E}[\| \mathrm{grad} f_{i_t}(w_{t-1}) - \mathcal{T}^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t)\|^2_{w_{t-1}} | \mathcal{F}_t] \\
& +\frac{4}{\alpha L}\mathbb{E}[\| \mathrm{grad} f_{i_t}(w_{t-1}) - \mathcal{T}^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t)\|_{w_{t-1}} \cdot \| P^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t) - \mathcal{T}^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t)\|_{w_{t-1}} | \mathcal{F}_t].
\end{aligned}
\tag{A.9}
$$

Here, we note that

$$
\begin{aligned}
& \| \mathrm{grad} f_{i_t}(w_{t-1}) - \mathcal{T}^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t)\|_{w_{t-1}} \\
= \quad & \| \mathrm{grad} f_{i_t}(w_{t-1}) - P^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t) + P^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t) - \mathcal{T}^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t)\|_{w_{t-1}} \\
\leq \quad & \| \mathrm{grad} f_{i_t}(w_{t-1}) - P^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t)\|_{w_{t-1}} + \| P^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t) - \mathcal{T}^{w_{t-1}}_{w_t} \mathrm{grad} f_{i_t}(w_t)\|_{w_{t-1}} \\
\leq \quad & L_l \| R^{-1}_{w_{t-1}}(w_t)\|_{w_{t-1}} + \theta\alpha \|v_{t-1}\|_{w_{t-1}} \|\mathrm{grad} f_{i_t}(w_t)\|_{w_t} \\
\leq \quad & (L_l + \theta C_g)\alpha \|v_{t-1}\|_{w_{t-1}},
\end{aligned}
$$

where the first inequality incorporates the triangle inequality and the second inequality incorporates Lemmas 3.7 and 3.8.

Substituting this result into (A.9) yields the first term in (A.8) as

$$\begin{aligned} & -\frac{2}{\alpha L}\mathbb{E}[\|\mathrm{grad} f_{i_t}(w_{t-1}) - P_{w_t}^{w_{t-1}}\mathrm{grad} f_{i_t}(w_t)\|^2_{w_{t-1}}|\mathcal{F}_t] \\ \leq \ & -\frac{2}{\alpha L}\mathbb{E}[\|\mathrm{grad} f_{i_t}(w_{t-1}) - \mathcal{T}_{w_t}^{w_{t-1}}\mathrm{grad} f_{i_t}(w_t)\|^2_{w_{t-1}}|\mathcal{F}_t] + \frac{4}{\alpha L}(L_l+\theta C_g)\alpha\|v_{t-1}\|_{w_{t-1}}\cdot\theta C_g\alpha\|v_{t-1}\|_{w_{t-1}} \\ = \ & -\frac{2}{\alpha L}\mathbb{E}[\|\mathrm{grad} f_{i_t}(w_{t-1}) - \mathcal{T}_{w_t}^{w_{t-1}}\mathrm{grad} f_{i_t}(w_t)\|^2_{w_{t-1}}|\mathcal{F}_t] + \frac{4(L_l+\theta C_g)\theta C_g\alpha}{L}\|v_{t-1}\|^2_{w_{t-1}}. \end{aligned} \tag{A.10}$$

On the other hand, the third term in (A.8) is calculated as

$$\begin{aligned} & -2\mathbb{E}[\langle P_{w_t}^{w_{t-1}}\mathrm{grad} f_{i_t}(w_t) - (\mathcal{T}_{w_{t-1}}^{w_t})^{-1}\mathrm{grad} f_{i_t}(w_t), v_{t-1}\rangle_{w_{t-1}}|\mathcal{F}_t] \\ \leq \ & 2\mathbb{E}[\|P_{w_t}^{w_{t-1}}\mathrm{grad} f_{i_t}(w_t) - (\mathcal{T}_{w_{t-1}}^{w_t})^{-1}\mathrm{grad} f_{i_t}(w_t)\|_{w_{t-1}}|\mathcal{F}_t]\cdot\|v_{t-1}\|_{w_{t-1}} \\ \leq \ & 2\theta C_g\alpha\|v_{t-1}\|^2_{w_{t-1}}. \end{aligned} \tag{A.11}$$

Substituting (A.10) and (A.11) into (A.7) yields

$$\begin{aligned} \mathbb{E}[\|v_t\|^2_{w_t}|\mathcal{F}_t] \ = \ & \|v_{t-1}\|^2_{w_{t-1}} + \mathbb{E}[\|\mathcal{T}_{w_{t-1}}^{w_t}\mathrm{grad} f_{i_t}(w_{t-1}) - \mathrm{grad} f_{i_t}(w_t)\|^2_{w_t} \\ & -2\langle\mathcal{T}_{w_{t-1}}^{w_t}\mathrm{grad} f_{i_t}(w_{t-1}) - \mathrm{grad} f_{i_t}(w_t), \mathcal{T}_{w_{t-1}}^{w_t}v_{t-1}\rangle_{w_t}|\mathcal{F}_t] \\ \leq \ & \|v_{t-1}\|^2_{w_{t-1}} + \mathbb{E}[\|\mathcal{T}_{w_{t-1}}^{w_t}\mathrm{grad} f_{i_t}(w_{t-1}) - \mathrm{grad} f_{i_t}(w_t)\|^2_{w_t}|\mathcal{F}_t] \\ & -\frac{2}{\alpha L}\mathbb{E}[\|\mathrm{grad} f_{i_t}(w_{t-1}) - \mathcal{T}_{w_t}^{w_{t-1}}\mathrm{grad} f_{i_t}(w_t)\|^2_{w_{t-1}}|\mathcal{F}_t] \\ & +\frac{2((2L_l+2\theta C_g+L)\theta C_g+\nu L)\alpha}{L}\|v_{t-1}\|^2_{w_{t-1}} \\ = \ & \|v_{t-1}\|^2_{w_{t-1}} + \left(1-\frac{2}{\alpha L}\right)\mathbb{E}[\|\mathcal{T}_{w_{t-1}}^{w_t}\mathrm{grad} f_{i_t}(w_{t-1}) - \mathrm{grad} f_{i_t}(w_t)\|^2_{w_t}|\mathcal{F}_t] \\ & +\frac{2((2L_l+2\theta C_g+L)\theta C_g+\nu L)\alpha}{L}\|v_{t-1}\|^2_{w_{t-1}} \\ \overset{(2)}{=} \ & \|v_{t-1}\|^2_{w_{t-1}} + \left(1-\frac{2}{\alpha L}\right)\mathbb{E}[\|v_t - \mathcal{T}_{w_{t-1}}^{w_t}v_{t-1}\|^2_{w_t}|\mathcal{F}_t] \\ & +\frac{2((2L_l+2\theta C_g+L)\theta C_g+\nu L)\alpha}{L}\|v_{t-1}\|^2_{w_{t-1}}. \end{aligned} \tag{A.12}$$

After taking the expectation, the equations are rearranged to yield

$$\begin{aligned} & \mathbb{E}[\|v_t - \mathcal{T}_{w_{t-1}}^{w_t}v_{t-1}\|^2_{w_t}] \\ \leq \ & \frac{\alpha L}{2-\alpha L}\left[\mathbb{E}[\|v_{t-1}\|^2_{w_{t-1}}] - \mathbb{E}[\|v_t\|^2_{w_t}] + \frac{2((2L_l+2\theta C_g+L)\theta C_g+\nu L)\alpha}{L}\mathbb{E}[\|v_{t-1}\|^2_{w_{t-1}}]\right]. \end{aligned}$$

Consider the above inequality in which t is replaced with $j \in \{1,2,\ldots,t\}$. Summing this inequality over $j = 1,2,\ldots,t$ $(t \geq 1)$ yields

$$\begin{aligned} & \sum_{j=1}^{t}\mathbb{E}[\|v_j - \mathcal{T}_{w_{j-1}}^{w_j}v_{j-1}\|^2_{w_j}] \\ \leq \ & \frac{\alpha L}{2-\alpha L}[\mathbb{E}[\|v_0\|^2_{w_0}] - \mathbb{E}[\|v_t\|^2_{w_t}]] + \sum_{j=1}^{t}\frac{2((2L_l+2\theta C_g+L)\theta C_g+\nu L)\alpha^2}{2-\alpha L}\mathbb{E}[\|v_{t-1}\|^2_{w_{t-1}}]. \end{aligned}$$

Defining $\psi(\alpha) = \dfrac{2((2L_l + 2\theta C_g + L)\theta C_g + \nu L)\alpha^2}{2 - \alpha L}$, we obtain the following from Lemma B.2:

$$\begin{aligned}
\mathbb{E}[\|\mathrm{grad}f(w_t) - v_t\|^2_{w_t}] &\leq \sum_{j=1}^{t} \mathbb{E}[\|v_j - \mathcal{T}^{w_j}_{w_{j-1}} v_{j-1}\|^2_{w_j}] \\
&\leq \frac{\alpha L}{2 - \alpha L}[\mathbb{E}[\|v_0\|^2_{w_0}] - \mathbb{E}[\|v_t\|^2_{w_t}]] + \sum_{j=1}^{t} \psi(\alpha)\mathbb{E}[\|v_{j-1}\|^2_{w_{j-1}}].
\end{aligned}$$

Finally, summing over $t = 1, 2, \ldots, m$ yields

$$\begin{aligned}
\sum_{t=1}^{m} \mathbb{E}[\|\mathrm{grad}f(w_t) - v_t\|^2_{w_t}] &\leq \frac{\alpha L}{2 - \alpha L} \sum_{t=1}^{m} \mathbb{E}[\|v_0\|^2_{w_0}] + \psi(\alpha) \sum_{t=1}^{m} \sum_{j=1}^{t} \mathbb{E}[\|v_{j-1}\|^2_{w_{j-1}}] \\
&= \frac{m\alpha L}{2 - \alpha L} \mathbb{E}[\|v_0\|^2_{w_0}] + \psi(\alpha) \sum_{t=1}^{m} \sum_{j=1}^{t} \mathbb{E}[\|v_{j-1}\|^2_{w_{j-1}}].
\end{aligned}$$

This completes the proof. □

B.2.4 Proof of Theorem 4.1

Proof. Let us define $\psi(\alpha) = \dfrac{2((2L_l + 2\theta C_g + L)\theta C_g + \nu L)\alpha^2}{2 - \alpha L}$ as in the previous proof. From Lemmas B.1 and B.6, we have

$$\begin{aligned}
&\sum_{t=0}^{m} \mathbb{E}[\|\mathrm{grad}f(w_t)\|^2_{w_t}] \\
\leq\ & \frac{2}{\alpha}\mathbb{E}[f(w_0) - f(w^*)] + \sum_{t=0}^{m} \mathbb{E}[\|\mathrm{grad}f(w_t) - v_t\|^2_{w_t}] - (1 - L\alpha) \sum_{t=0}^{m} \mathbb{E}[\|v_t\|^2_{w_t}] \\
\leq\ & \frac{2}{\alpha}\mathbb{E}[f(w_0) - f(w^*)] + \frac{m\alpha L}{2 - \alpha L}\mathbb{E}[\|v_0\|^2_{w_0}] + \psi(\alpha) \sum_{t=1}^{m} \sum_{j=1}^{t} \mathbb{E}[\|v_{j-1}\|^2_{w_{j-1}}] - (1 - L\alpha) \sum_{t=0}^{m} \mathbb{E}[\|v_t\|^2_{w_t}] \\
\leq\ & \frac{2}{\alpha}\mathbb{E}[f(w_0) - f(w^*)] + \frac{m\alpha L}{2 - \alpha L}\mathbb{E}[\|v_0\|^2_{w_0}] \\
& + \psi(\alpha)\left[m\mathbb{E}\|v_0\|^2_{w_0} + (m-1)\mathbb{E}\|v_1\|^2_{w_1} + \cdots + \mathbb{E}\|v_{m-1}\|^2_{w_{m-1}}\right] - (1 - L\alpha) \sum_{t=0}^{m} \mathbb{E}[\|v_t\|^2_{w_t}] \\
\leq\ & \frac{2}{\alpha}\mathbb{E}[f(w_0) - f(w^*)] + \frac{m\alpha L}{2 - \alpha L}\mathbb{E}[\|v_0\|^2_{w_0}] + [m\psi(\alpha) - (1 - L\alpha)] \sum_{t=0}^{m} \mathbb{E}[\|v_t\|^2_{w_t}].
\end{aligned}$$

From the assumption, i.e., $(2((2L_l + 2\theta C_g + L)\theta C_g + \nu L)m - L^2)\alpha^2 + 3L\alpha - 2 \leq 0$, the third term is not greater than zero. Consequently, we obtain

$$\sum_{t=0}^{m} \mathbb{E}[\|\mathrm{grad}f(w_t)\|^2_{w_t}] \leq \frac{2}{\alpha}\mathbb{E}[f(w_0) - f(w^*)] + \frac{m\alpha L}{2 - \alpha L}\mathbb{E}[\|v_0\|^2_{w_0}].$$

The discussion above is for a single outer iteration. Then, for $s \geq 1$, we have $v_0 = \mathrm{grad}f(w_0) = \mathrm{grad}f(\tilde{w}^{s-1})$, because $w_0 = \tilde{w}^{s-1}$. We set $\tilde{w}^s = w_t$ where t is randomly selected from $\{0, 1, \ldots, m\}$. Consequently, we obtain

$$\begin{aligned}
\mathbb{E}[\|\mathrm{grad}f(\tilde{w}^s)\|^2_{\tilde{w}^s}] &\leq \frac{1}{m+1} \sum_{t=0}^{m} \mathbb{E}[\|\mathrm{grad}f(w_t)\|^2_{w_t}] \\
&\leq \frac{2}{\alpha(m+1)}\mathbb{E}[f(\tilde{w}^{s-1}) - f(w^*)] + \frac{\alpha L}{2 - \alpha L}\mathbb{E}[\|\mathrm{grad}f(\tilde{w}^{s-1})\|^2_{\tilde{w}^{s-1}}].
\end{aligned}$$

This completes the proof. □

B.2.5 Proof of Theorem 4.2

Proof. As the proof of Theorem 4.2 is similar to that for Theorem 3 of (Nguyen et al., 2017a), it is omitted. □

Suppose that we select the step size as constant $\alpha = \alpha^* := \frac{2}{3}\frac{3L - \sqrt{L^2+8\beta}}{2(L^2-\beta)}$. Then, φ in Theorem 4.2 is obtained as

$$\varphi \ = \ \frac{\alpha^* L}{2-\alpha^* L} \ = \ \frac{(3L-\sqrt{L^2+8\beta})L}{6(L^2-\beta)-(3L-\sqrt{L^2+8\beta})L} \leq \frac{(3L-\sqrt{L^2})L}{6(L^2-\beta)-3L^2+\sqrt{\beta}\sqrt{\beta+8\beta}} \ = \ \frac{2L^2}{3(L^2{-}\beta)},$$

which requires $(0 \leq)\ \beta \leq \dfrac{L^2}{3}$ to satisfy $\varphi < 1$.

B.2.6 Proof of Theorem 4.3

Proof. The proof of Theorem 4.3 is similar to that for Theorem 4 of (Nguyen et al., 2017a), as

$$\begin{aligned}
\mathbb{E}[\|\mathrm{grad} f(\tilde{w}^s)\|^2_{\tilde{w}^s}] \ &\leq \ \frac{2}{\alpha(m+1)}\mathbb{E}[f(\tilde{w}^{s-1}) - f(w^*)] + \frac{\alpha L}{2-\alpha L}\mathbb{E}[\|\mathrm{grad} f(\tilde{w}^{s-1})\|^2_{\tilde{w}^{s-1}}] \\
&\leq \ \left(\frac{1}{\mu\alpha(m+1)} + \frac{\alpha L}{2-\alpha L}\right)\mathbb{E}[\|\mathrm{grad} f(\tilde{w}^{s-1})\|^2_{\tilde{w}^{s-1}}],
\end{aligned}$$

where the last inequality incorporates the relation of retraction μ-strongly convex function. □

To obtain the total complexity, we suppose that $\beta \leq \dfrac{L^2}{5}$, and select $\alpha = \alpha^* := \dfrac{1}{2}\alpha_l = \dfrac{1}{2}\dfrac{3L - \sqrt{L^2+8\beta}}{2(L^2-\beta)}$ and $m = 6.5\kappa$. Then, σ_m is obtained as

$$\begin{aligned}
\sigma_m \ &= \ \frac{1}{\mu\alpha^*(m+1)} + \frac{\alpha^* L}{2-\alpha^* L} \\
&= \ \frac{4(L^2-\beta)}{\mu(3L-\sqrt{L^2+8\beta})(m+1)} + \frac{(3L-\sqrt{L^2+8\beta})L}{8(L^2-\beta)-(3L-\sqrt{L^2+8\beta})L} \\
&\leq \ \frac{4L\kappa(1-\beta/L^2)}{(3L-\sqrt{L^2+8\beta})(m+1)} + \frac{(3L-\sqrt{L^2})L}{8(L^2-\beta)-3L^2+\sqrt{\beta}\sqrt{\beta+8\beta}} \\
&\leq \ \frac{8L(1-\beta/L^2)}{13(3L-\sqrt{L^2+8\beta})} + \frac{2L^2}{5(L^2-\beta)} \ \leq \ \frac{L^2(1-\beta/L^2)(3+\sqrt{1+8\beta/L^2})}{13(L^2-\beta)} + \frac{2L^2}{5(L^2-\beta)} \\
&\leq \ \frac{L^2(3+\sqrt{1+8\beta/L^2})}{13(L^2-\beta)} + \frac{2L^2}{5(L^2-\beta)} \ \leq \ \frac{L^2(3+\sqrt{13/5})}{13(L^2-\beta)} + \frac{2L^2}{5(L^2-\beta)} \\
&< \ \frac{2L^2}{5(L^2-\beta)} + \frac{2L^2}{5(L^2-\beta)} \ = \ \frac{4L^2}{5(L^2-\beta)} \ \leq \ 1.
\end{aligned}$$

B.2.7 Proof of Proposition 4.4

Proof. The norm of the difference between the two gradients of the successive iterates is bounded as

$$\begin{aligned}
\|\mathrm{grad} f(w_t) - \mathcal{T}_{w_{t-1}}^{w_t}\mathrm{grad} f(w_{t-1})\|^2_{w_t} \ &= \ \left\|\frac{1}{n}\sum_{i=1}^{n}[\mathrm{grad} f_i(w_t) - \mathcal{T}_{w_{t-1}}^{w_t}\mathrm{grad} f_i(w_{t-1})]\right\|^2_{w_t} \\
&\leq \ \frac{1}{n}\sum_{i=1}^{n}\left\|\mathrm{grad} f_i(w_t) - \mathcal{T}_{w_{t-1}}^{w_t}\mathrm{grad} f_i(w_{t-1})\right\|^2_{w_t} \\
&= \ \mathbb{E}[\|\mathrm{grad} f_{i_t}(w_t) - \mathcal{T}_{w_{t-1}}^{w_t}\mathrm{grad} f_{i_t}(w_{t-1})\|^2_{w_t}|\mathcal{F}_t].
\end{aligned}$$

Hence, noting that $1 - \dfrac{2}{\alpha L} \leq 0$, we obtain the following from (A.12):

$$\begin{aligned} \mathbb{E}[\|v_t\|_{w_t}^2 | \mathcal{F}_t] &\leq \|v_{t-1}\|_{w_{t-1}}^2 + \left(1 - \frac{2}{\alpha L}\right) \mathbb{E}[\|\mathcal{T}_{w_{t-1}}^{w_t} \mathrm{grad} f_{i_t}(w_{t-1}) - \mathrm{grad} f_{i_t}(w_t)\|_{w_t}^2 | \mathcal{F}_t] \\ &\quad + \frac{2((2L_l + 2\theta C_g + L)\theta C_g + \nu L)\alpha}{L} \|v_{t-1}\|_{w_{t-1}} \\ &\leq \|v_{t-1}\|_{w_{t-1}}^2 + \left(1 - \frac{2}{\alpha L}\right) \|\mathrm{grad} f(w_t) - \mathcal{T}_{w_{t-1}}^{w_t} \mathrm{grad} f(w_{t-1})\|_{w_t}^2 \\ &\quad + \frac{2((2L_l + 2\theta C_g + L)\theta C_g + \nu L)\alpha}{L} \|v_{t-1}\|_{w_{t-1}}. \end{aligned}$$

Noting that the coefficient of the second term is not greater than zero, the term is bounded as

$$\begin{aligned} & -\|\mathrm{grad} f(w_t) - \mathcal{T}_{w_{t-1}}^{w_t} \mathrm{grad} f(w_{t-1})\|_{w_t}^2 \\ = \;& -\|\mathrm{grad} f(w_t) - P_{w_{t-1}}^{w_t} \mathrm{grad} f(w_{t-1}) + P_{w_{t-1}}^{w_t} \mathrm{grad} f(w_{t-1}) - \mathcal{T}_{w_{t-1}}^{w_t} \mathrm{grad} f(w_{t-1})\|_{w_t}^2 \\ = \;& -\|\mathrm{grad} f(w_t) - P_{w_{t-1}}^{w_t} \mathrm{grad} f(w_{t-1})\|_{w_t}^2 - \|P_{w_{t-1}}^{w_t} \mathrm{grad} f(w_{t-1}) - \mathcal{T}_{w_{t-1}}^{w_t} \mathrm{grad} f(w_{t-1})\|_{w_t}^2 \\ & -2\langle \mathrm{grad} f(w_t) - P_{w_{t-1}}^{w_t} \mathrm{grad} f(w_{t-1}), P_{w_{t-1}}^{w_t} \mathrm{grad} f(w_{t-1}) - \mathcal{T}_{w_{t-1}}^{w_t} \mathrm{grad} f(w_{t-1}) \rangle_{w_t} \\ \overset{(A.5)}{\leq} \;& -(a_0\mu - a_1 C_g)^2 \|R_{w_{t-1}}^{-1}(w_t)\|_{w_{t-1}}^2 \\ & +2\|\mathrm{grad} f(w_t) - P_{w_{t-1}}^{w_t} \mathrm{grad} f(w_{t-1})\|_{w_t} \cdot \|P_{w_{t-1}}^{w_t} \mathrm{grad} f(w_{t-1}) - \mathcal{T}_{w_{t-1}}^{w_t} \mathrm{grad} f(w_{t-1})\|_{w_t} \\ \leq \;& -(a_0\mu - a_1 C_g)^2 \alpha^2 \|v_{t-1}\|_{w_{t-1}}^2 + 2L_l \|R_{w_{t-1}}^{-1}(w_t)\|_{w_{t-1}} \cdot \theta C_g \alpha \|v_{t-1}\|_{w_{t-1}} \\ = \;& -((a_0\mu - a_1 C_g)^2 - 2L_l \theta C_g) \alpha^2 \|v_{t-1}\|_{w_{t-1}}^2, \end{aligned}$$

where the first inequality incorporates the Cauchy-Schwarz inequality, and Lemma B.5. The second and third inequalities employ Lemmas 3.7 and 3.8. Consequently, we obtain

$$\begin{aligned} & \mathbb{E}[\|v_t\|_{w_t}^2 | \mathcal{F}_t] \\ \leq \;& \|\mathcal{T}_{w_{t-1}}^{w_t} v_{t-1}\|_{w_t}^2 + \left(1 - \frac{2}{\alpha L}\right) \mathbb{E}[\|\mathcal{T}_{w_{t-1}}^{w_t} \mathrm{grad} f_{i_t}(w_{t-1}) - \mathrm{grad} f_{i_t}(w_t)\|_{w_t}^2 | \mathcal{F}_t] \\ & + \frac{2((2L_l + 2\theta C_g + L)\theta C_g + \nu L)\alpha}{L} \|v_{t-1}\|_{w_{t-1}}^2 \\ \leq \;& \|v_{t-1}\|_{w_{t-1}}^2 + \left(1 - \frac{2}{\alpha L}\right) ((a_0\mu - a_1 C_g)^2 - 2L_l \theta C_g) \alpha^2 \|v_{t-1}\|_{w_{t-1}}^2 \\ & + \frac{2((2L_l + 2\theta C_g + L)\theta C_g + \nu L)\alpha}{L} \|v_{t-1}\|_{w_{t-1}}^2 \\ \leq \;& \left(1 + \left(1 - \frac{2}{\alpha L}\right) ((a_0\mu - a_1 C_g)^2 - 2L_l \theta C_g) \alpha^2 + \frac{2((2L_l + 2\theta C_g + L)\theta C_g + \nu L)\alpha}{L}\right) \|v_{t-1}\|_{w_{t-1}}^2 \\ = \;& \left(1 - \left(\frac{2}{\alpha L} - 1\right) (a_0\mu - a_1 C_g)^2 \alpha^2 + \frac{(2(2L_l + 2\theta C_g + L - (\alpha L - 2)L_l)\theta C_g + 2\nu L)\alpha}{L}\right) \|v_{t-1}\|_{w_{t-1}}^2 \\ = \;& \left(1 - \left(\frac{2}{\alpha L} - 1\right) (a_0\mu - a_1 C_g)^2 \alpha^2 + \frac{2((4L_l + 2\theta C_g + L - \alpha L L_l)\theta C_g + \nu L)\alpha}{L}\right) \|v_{t-1}\|_{w_{t-1}}^2. \end{aligned}$$

Defining $\phi(\alpha) = \dfrac{2((4L_l + 2\theta C_g + L - \alpha L L_l)\theta C_g + \nu L)\alpha}{L}$, we obtain

$$\mathbb{E}[\|v_t\|_{w_t}^2 | \mathcal{F}_t] \leq \left(1 - \left(\frac{2}{\alpha L} - 1\right) (a_0\mu - a_1 C_g)^2 \alpha^2 + \phi(\alpha)\right) \|v_{t-1}\|_{w_{t-1}}^2.$$

Finally, denoting the coefficient of $\|v_{t-1}\|_{w_{t-1}}^2$ as λ, recursive calculation while taking the total expectation yields

$$\mathbb{E}[\|v_t\|_{w_t}^2] \;\leq\; \lambda \mathbb{E}[\|v_{t-1}\|_{w_{t-1}}^2] \;\leq\; \lambda^t \mathbb{E}[\|v_0\|_{w_0}^2] \;=\; \lambda^t \mathbb{E}[\|\mathrm{grad} f(w_0)\|_{w_0}^2].$$

When θ and ν are close to zero, $\phi(\alpha)$ becomes closely zero. Then, we obtain $\lambda < 1$. This completes the proof. □

B.3 Proofs of non-convex functions

This subsection presents the convergence analysis for non-convex functions. The proof strategy follows and extends that of Theorem 4.1 and (Nguyen et al., 2017b).

B.3.1 Lemma B.7

Lemma B.7. *Suppose that the conditions of Lemma 3.8 hold. Consider v_t in Algorithm 1 with a constant step size α. Then, for any $t \geq 1$,*

$$\|v_j - \mathcal{T}_{w_{j-1}}^{w_j} v_{j-1}\|_{w_j}^2 \quad \leq \quad 2(L_l^2 + \theta^2 C_g^2)\alpha^2 \|v_{j-1}\|_{w_{j-1}}^2 .$$

Proof. We have

$$\begin{aligned}
& \|v_j - \mathcal{T}_{w_{j-1}}^{w_j} v_{j-1}\|_{w_j}^2 \\
= \; & \|\mathrm{grad} f_{i_j}(w_j) - \mathcal{T}_{w_{j-1}}^{w_j} \mathrm{grad} f_{i_j}(w_{j-1})\|_{w_j}^2 \\
= \; & \|\mathrm{grad} f_{i_j}(w_j) - P_{w_{j-1}}^{w_j} \mathrm{grad} f_{i_j}(w_{j-1}) + P_{w_{j-1}}^{w_j} \mathrm{grad} f_{i_j}(w_{j-1}) - \mathcal{T}_{w_{j-1}}^{w_j} \mathrm{grad} f_{i_j}(w_{j-1})\|_{w_j}^2 \\
\leq \; & 2\|\mathrm{grad} f_{i_j}(w_j) - P_{w_{j-1}}^{w_j} \mathrm{grad} f_{i_j}(w_{j-1})\|_{w_j}^2 + 2\|P_{w_{j-1}}^{w_j} \mathrm{grad} f_{i_j}(w_{j-1}) - \mathcal{T}_{w_{j-1}}^{w_j} \mathrm{grad} f_{i_j}(w_{j-1})\|_{w_j}^2 \\
\leq \; & 2L_l^2 \|R_{w_{j-1}}^{-1}(w_j)\|_{w_{j-1}}^2 + 2\theta^2 \|\mathrm{grad} f_{i_j}(w_{j-1})\|_{w_{j-1}}^2 \|\alpha v_{j-1}\|_{w_{j-1}}^2 \\
= \; & 2(L_l^2 + \theta^2 C_g^2)\alpha^2 \|v_{j-1}\|_{w_{j-1}}^2 ,
\end{aligned} \tag{A.13}$$

where the first inequality incorporates $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ for any vectors a and b in a norm space. The second inequality employs Lemmas 3.7 and 3.8. This completes the proof. □

B.3.2 Proof of Theorem 4.5

Proof. From Lemmas B.2 and B.7, we obtain

$$\begin{aligned}
\mathbb{E}[\|\mathrm{grad} f(w_t) - v_t\|_{w_t}^2] \quad & \overset{(A.3)}{\leq} \quad \sum_{j=1}^{t} \mathbb{E}[\|v_j - \mathcal{T}_{w_{j-1}}^{w_j} v_{j-1}\|_{w_j}^2] \\
& \overset{(A.13)}{\leq} \quad \sum_{j=1}^{t} 2(L_l^2 + \theta^2 C_g^2)\alpha^2 \mathbb{E}[\|v_{j-1}\|_{w_{j-1}}^2].
\end{aligned}$$

Because $\|\mathrm{grad} f(w_0) - v_0\|_{w_j}^2 = 0$, summing over $t = 1, 2, \ldots, m$ yields

$$\begin{aligned}
\sum_{t=0}^{m} \mathbb{E}[\|v_t - \mathrm{grad} f(w_t)\|_{w_t}^2] \quad & \leq \quad \sum_{j=1}^{1} 2(L_l^2 + \theta^2 C_g^2)\alpha^2 \mathbb{E}[\|v_{j-1}\|_{w_{j-1}}^2] + \ldots + \sum_{j=1}^{m} 2(L_l^2 + \theta^2 C_g^2)\alpha^2 \mathbb{E}[\|v_{j-1}\|_{w_{j-1}}^2] \\
& = \quad 2(L_l^2 + \theta^2 C_g^2)\alpha^2 [m\mathbb{E}[\|v_0\|_{w_0}^2] + (m-1)\mathbb{E}[\|v_1\|_{w_1}^2] + \cdots + \mathbb{E}[\|v_{m-1}\|_{w_{m-1}}^2]].
\end{aligned}$$

Hence, it follows that

$$\begin{aligned}
& \sum_{t=0}^{m} \mathbb{E}[\|v_t - \mathrm{grad} f(w_t)\|_{w_t}^2] - (1 - L\alpha) \sum_{t=0}^{m} \mathbb{E}[\|v_t\|_{w_t}^2] \\
\leq \; & 2(L_l^2 + \theta^2 C_g^2)\alpha^2 [m\mathbb{E}[\|v_0\|_{w_0}^2] + (m-1)\mathbb{E}[\|v_1\|_{w_1}^2] + \cdots + \mathbb{E}[\|v_{m-1}\|_{w_{m-1}}^2]] - (1 - L\alpha) \sum_{t=0}^{m} \mathbb{E}[\|v_t\|_{w_t}^2] \\
\leq \; & 2(L_l^2 + \theta^2 C_g^2)\alpha^2 [m\mathbb{E}[\|v_0\|_{w_0}^2] + m\mathbb{E}[\|v_1\|_{w_1}^2] + \cdots + m\mathbb{E}[\|v_{m-1}\|_{w_{m-1}}^2]] - (1 - L\alpha) \sum_{t=0}^{m} \mathbb{E}[\|v_t\|_{w_t}^2] \\
\leq \; & \left[2(L_l^2 + \theta^2 C_g^2)\alpha^2 m - (1 - L\alpha)\right] \sum_{t=1}^{m} \mathbb{E}[\|v_{t-1}\|_{w_{t-1}}^2],
\end{aligned}$$

where the last inequality holds due to $\alpha \leq \dfrac{1}{L}$, which clearly holds according to the condition of α in the statement.

Addressing the first terms inside the square brackets, the following α is a larger root of $2(L_l^2 + \theta^2 C_g^2)\alpha^2 m - (1 - L\alpha) = 0$, i.e.,

$$\alpha \quad = \quad \frac{2}{L + \sqrt{L^2 + 8m(L_l^2 + C_g^2\theta^2)}},$$

and the smaller root is less than zero. Therefore, from the assumption, the right-hand side of the equation above is not greater than zero. Consequently, we obtain

$$\sum_{t=0}^{m} \mathbb{E}[\|v_t - \mathrm{grad} f(w_t)\|_{w_t}^2] - (1 - L\alpha) \sum_{t=0}^{m} \mathbb{E}[\|v_t\|_{w_t}^2] \quad \leq \quad 0.$$

Finally, according to Lemma B.1, we have

$$\begin{aligned} \sum_{t=0}^{m} \mathbb{E}[\|\mathrm{grad} f(w_t)\|_{w_t}^2] \quad &\leq \quad \frac{2}{\alpha}\mathbb{E}[f(w_0) - f(w^*)] + \sum_{t=0}^{m} \mathbb{E}[\|\mathrm{grad} f(w_t) - v_t\|_{w_t}^2] - (1 - L\alpha) \sum_{t=0}^{m} \mathbb{E}[\|v_t\|_{w_t}^2] \\ &\leq \quad \frac{2}{\alpha}\mathbb{E}[f(w_0) - f(w^*)]. \end{aligned}$$

If we select $\tilde{w} = w_t$, where t is selected randomly from $\{0, 1, \ldots, m\}$, we obtain

$$\mathbb{E}[\|\mathrm{grad} f(\tilde{w}^s)\|_{\tilde{w}^s}^2] = \frac{1}{m+1} \sum_{t=0}^{m} \mathbb{E}[\|\mathrm{grad} f(w_t)\|_{w_t}^2] \leq \frac{2}{\alpha(m+1)}[f(w_0) - f(w^*)].$$

This completes the proof. □

B.3.3 Proof of Theorem 4.6

Proof. As this proof is identical to that of Theorem 2 in (Nguyen et al., 2017b), it is omitted. □

When $\alpha = \dfrac{2}{L + \sqrt{L^2 + 8m(L_l^2 + C_g^2\theta^2)}}$, we obtain the value of $\dfrac{\alpha(m+1)}{2}$ as

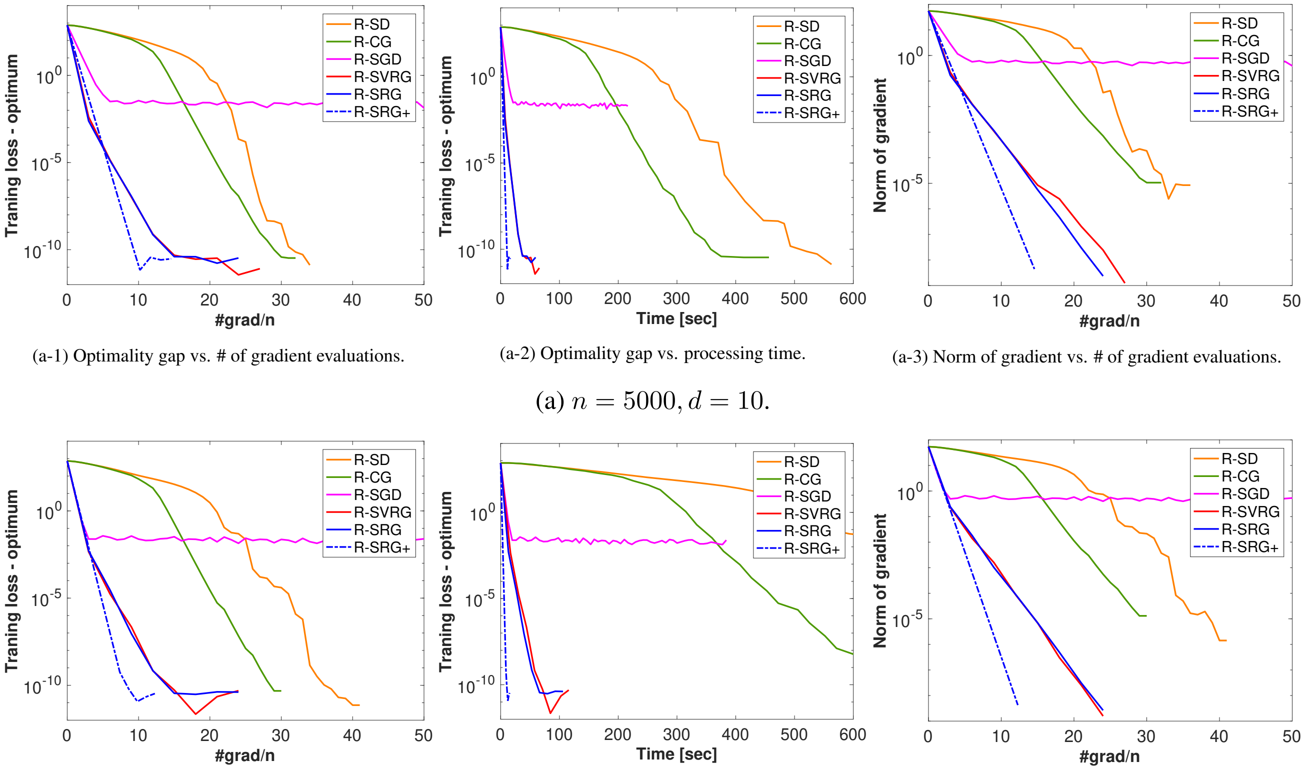
$$\begin{aligned} \frac{\alpha(m+1)}{2} \quad &= \quad \frac{m+1}{L + \sqrt{L^2 + 8m\left(L_l^2 + C_g^2\theta^2\right)}} > \frac{m}{2\sqrt{L^2 + 8m\left(L_l^2 + C_g^2\theta^2\right)}} = \frac{m}{2L\sqrt{1 + 8m\left(\frac{L_l^2}{L^2} + \frac{C_g^2\theta^2}{L^2}\right)}} \\ &> \quad \frac{m}{2L\sqrt{16m\left(\rho_l^2 + \frac{C_g^2\theta^2}{L^2}\right)}} \quad = \quad \frac{\sqrt{m}}{8L\sqrt{\rho_l^2 + \frac{C_g^2\theta^2}{L^2}}}. \end{aligned}$$

Therefore, we conclude that it is necessary to choose m such that $m > 64L^2\left(\rho_l^2 + \dfrac{C_g^2\theta^2}{L^2}\right)\tau^2 = 64\tau^2\left(L^2\rho_l^2 + C_g^2\theta^2\right)$ to satisfy $\dfrac{\alpha(m+1)}{2} > \tau$.

C Additional evaluations

C.1 Riemannian centroid problem on SPD manifold

We present additional evaluation results for the Riemannian centroid problem on the SPD manifold. The sample size n and dimension d are varied. Figures A.1 (a) and (b) show the results obtained for $n = 5000$ and $d = 10$, and $n = 10000$ and $d = 10$, respectively. The results indicate that the proposed R-SRG is competitive with R-SVRG, while R-SRG+ outperforms the others (especially when n and d are larger), in terms of the number of gradient evaluations and processing time.



(a-1) Optimality gap vs. # of gradient evaluations.

(a-2) Optimality gap vs. processing time.

(a-3) Norm of gradient vs. # of gradient evaluations.

(a) $n = 5000, d = 10$.

(b-1) Optimality gap vs. # of gradient evaluations.

(b-2) Optimality gap vs. processing time.

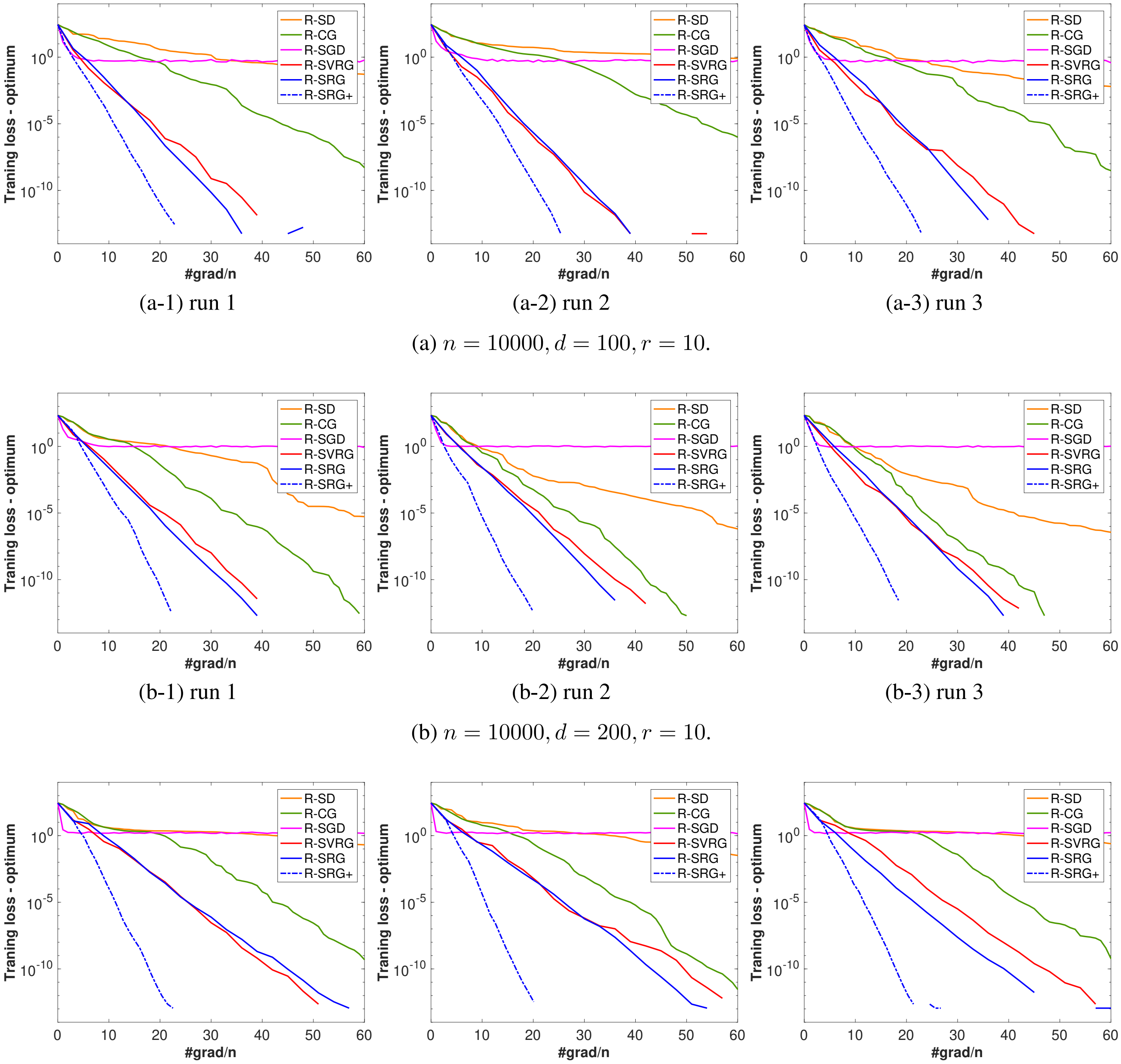
(b-3) Norm of gradient vs. # of gradient evaluations.

(b) $n = 10000, d = 10$.

Figure A.1. Riemannian centroid problem on SPD manifold.

C.2 PCA problem on Grassmann manifold

We present additional evaluation results for the PCA problem on the Grassmann manifold. Again, n and d are varied. Figures A.2(a)–(c) show the optimality gap results of three trials, for $(n, d, r) = (10000, 100, 10)$, $(10000, 200, 10)$, and $(50000, 200, 10)$, respectively. Overall, the proposed R-SRG is competitive with R-SVRG, and R-SRG+ outperforms the others, especially when n is larger.



(a-1) run 1

(a-2) run 2

(a-3) run 3

(a) $n = 10000, d = 100, r = 10$.

(b-1) run 1

(b-2) run 2

(b-3) run 3

(b) $n = 10000, d = 200, r = 10$.

(c-1) run 1

(c-2) run 2

(c-3) run 3

(c) $n = 50000, d = 200, r = 10$.

Figure A.2. PCA problem on Grassmann manifold.

References

Absil, P.-A., Mahony, R., and Sepulchre, R. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

Huang, W., Absil, P.-A., and Gallivan, K. A. A Riemannian symmetric rank-one trust-region method. *Math. Program., Ser. A*, 150:179–216, 2015a.

Huang, W., Gallivan, K. A., and Absil, P.-A. A Broyden class of quasi-Newton methods for Riemannian optimization. *SIAM J. Optim.*, 25(3):1660–1685, 2015b.

Jeuris, B., Vandebril, R., and Vandereycken, B. A survey and comparison of contemporary algorithms for computing the matrix geometric mean. *ETNA*, 2012.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takac, M. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *ICML*, 2017a.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takac, M. Stochastic recursive gradient algorithm for nonconvex optimization. *arXiv preprint arXiv:1705.07261*, 2017b.

Pennec, X., Fillard, P., and Ayache, N. A Riemannian framework for tensor computing. *Int. Jornal of Computer Vision*, 66 (1):41–66, 2006.

Yuan, X., Huang, W. Absil, P.-A., and Gallivan, K. A. A Riemannian limited-memory BFGS algorithm for computing the matrix geometric mean. In *ICCS*, 2016.