

## A Proofs

### A.1 Proof of Lemma 2.8

*Proof.* Without loss of generality, we assume  $\|v\| = 1$ , otherwise conduct the proof for  $v/\|v\|$ . Let  $\gamma(t)$  consists of unit speed geodesic segments. Denote the parallel transported vector  $v$  along  $\gamma$  as  $w(t)$ , i.e.  $w(0) = v$ . For the extrinsic geometry we denote  $v = v^T(t) + v^\perp(t)$  where  $v^T(t) \in T_{\gamma(t)}\mathcal{M}$  is the projection of  $v$  onto  $T_{\gamma(t)}\mathcal{M}$  and  $v^\perp(t)$  is the normal component of  $v$  which is orthogonal to  $T_{\gamma(t)}\mathcal{M}$ . Note that,  $v^T(t)$  is independent of the path  $\gamma$  and only depends on the point  $\gamma(t)$ . Eventually, we want to bound  $\|w(t) - v^T(t)\|$ .

Since  $w(t)$  is a parallel transport of  $v$ , the tangent component of its derivative must be zero, i.e.  $(w'(t))^T = 0$ . Now consider any unit parallel vector field  $z(t) \in T_{\gamma(t)}\mathcal{M}$  along  $\gamma$ , we have  $\langle v^\perp(t), z(t) \rangle = 0$ , then by taking the derivative with respect to  $t$ , we obtain  $\langle (v^\perp)'(t), z(t) \rangle = -\langle v^\perp(t), z'(t) \rangle = -\langle v^\perp(t), \Pi(\gamma'(t), z(t)) \rangle$  where  $\Pi$  is the second fundamental form. We also have  $(v^T)'(t) = -(v^\perp)'(t)$  since  $v^T(t) + v^\perp(t) = v$  is fixed. We get  $\langle (v^T)'(t), z(t) \rangle = \langle v^\perp(t), \Pi(\gamma'(t), z(t)) \rangle$ . Now the right hand side has a uniform upper bound of  $C$ , and by the arbitrarily chosen  $z(t) \in T_{\gamma(t)}\mathcal{M}$ , we get  $\|((v^T)'(t))^T\| \leq C$ .

We can now bound the derivative of  $\|w(t) - v^T(t)\|$  as

$$\begin{aligned} (\|w(t) - v^T(t)\|^2)' &= (1 - 2\langle w(t), v^T(t) \rangle + \|v^T(t)\|^2)' \\ &= -2\langle v^T(t), w'(t) \rangle - 2\langle w(t), (v^T(t))' \rangle + 2\langle v^T(t), (v^T(t))' \rangle. \end{aligned}$$

The first term is 0 since  $w'(t) \in T_{\gamma(t)}^\perp\mathcal{M}$  and  $v^T(t) \in T_{\gamma(t)}\mathcal{M}$  are orthogonal. Then we have

$$(\|w(t) - v^T(t)\|^2)' = 2\langle v^T(t) - w(t), (v^T(t))' \rangle \leq 2C\|v^T(t) - w(t)\|.$$

This means that  $\|w(t) - v^T(t)\|' \leq C$ . Now integrating the above inequality on the geodesic segments of  $\gamma$  where the initial value,  $\|w(0) - v^T(0)\| = 0$ , we obtain  $\|\mathcal{P}_{0,t}^\gamma(v) - \text{Proj}_{T_{\gamma(t)}\mathcal{M}}(v)\| = \|w(t) - v^T(t)\| \leq C\|v\|\text{length}(\gamma)$  which completes the proof.  $\square$

### A.2 Proof of Lemma 2.9

*Proof.* It is given that  $\text{dist}(w_t, y) \leq \delta$ , so  $P_{w_t, y}^g(\nabla_t) = P_{w_t, y}^g(\text{grad}f(w_t)) \in \partial_\delta f(y)$ . Also, the average  $\frac{1}{T} \sum_{t=0}^{T-1} P_{w_t, y}^g(\nabla_t) \in \partial_\delta f(y)$ . Then, we have by linearity of parallel transport operation that

$$\begin{aligned} \|\text{grad}f(y)\|_\delta &\leq \left\| \frac{1}{T} \sum_{t=0}^{T-1} P_{w_t, y}^g(\nabla_t) \right\| \\ &\leq \left\| P_{x_0, y}^g \left( \frac{1}{T} \sum_{t=0}^{T-1} (\mathcal{P}_{S^{t+1}}^s)^{-1} \circ P_{w_t, x_{t+1}}^g(\nabla_t) \right) \right\| + \left\| \frac{1}{T} \sum_{t=0}^{T-1} P_{w_t, y}^g(\nabla_t) - P_{x_0, y}^g \circ (\mathcal{P}_{S^{t+1}}^s)^{-1} \circ P_{w_t, x_{t+1}}^g(\nabla_t) \right\| \\ &\leq \left\| \frac{1}{T} \sum_{t=0}^{T-1} (\mathcal{P}_{S^{t+1}}^s)^{-1} \circ P_{w_t, x_{t+1}}^g(\nabla_t) \right\| + \left\| \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla_t - P_{y, w_t}^g \circ P_{x_0, y}^g \circ (\mathcal{P}_{S^{t+1}}^s)^{-1} \circ P_{w_t, x_{t+1}}^g(\nabla_t)\| \right\|. \end{aligned}$$

In the last term, we have a difference of a vector and its parallel transport along the sequence  $S'_t := \{w_t, x_{t+1}, x_t, \dots, x_0, y, w_t\}$ . In this sequence, based on the assumptions of the lemma, we know that  $\text{dist}(x_s, x_{s+1}) \leq D$  for  $s \in \{0, \dots, t\}$ ,  $\text{dist}(w_t, x_{t+1}) \leq D$  and  $\text{dist}(y, w_t) \leq \delta$ . To bound  $\text{dist}(x_0, y)$ , we can use triangle inequality to derive

$$\text{dist}(x_0, y) \leq \text{dist}(x_0, x_1) + \text{dist}(x_1, w_0) + \text{dist}(w_0, y) \leq \delta + 2D.$$

Then, we have  $\text{length}(S'_t) \leq D(t+4) + 2\delta$ . Now, by using Lemma 2.8 we have

$$\left\| \nabla_t - P_{y, w_t}^g \circ P_{x_0, y}^g \circ (\mathcal{P}_{S^{t+1}}^s)^{-1} \circ P_{w_t, x_{t+1}}^g(\nabla_t) \right\| \leq CL(D(t+4)) + 2C\delta L. \quad (3)$$

Since  $D = \frac{\delta}{T}$ , taking average over  $t$ , we have

$$\frac{1}{T} \sum_{t=0}^{T-1} \left\| \nabla_t - P_{y, w_t}^g \circ P_{x_0, y}^g \circ (\mathcal{P}_{S^{t+1}}^s)^{-1} \circ P_{w_t, x_{t+1}}^g(\nabla_t) \right\| \leq 3L\delta C, \quad (4)$$

if  $T > 7$ . As a result, we obtain

$$\|\text{grad}f(y)\|_\delta \leq \left\| \frac{1}{T} \sum_{t=0}^{T-1} (\mathcal{P}_{S_{t+1}}^s)^{-1} \circ P_{w_t, x_{t+1}}^g(\nabla_t) \right\| + 3LC\delta. \quad (5)$$

□

### A.3 Proof of Theorem 3.1

*Proof.* We first find a bound on the term  $\left\| \sum_{t=0}^{T-1} (\mathcal{P}_{S_{t+1}}^s)^{-1} \circ P_{w_t, x_{t+1}}^g(\nabla_t) \right\|$  for an epoch consisting of  $T$  iterations. Though we take the average for  $K$  epochs, we drop subscript  $k$  for convenience.

Now, consider a length  $T$  period where the online algorithm works without restart. Using the retraction operator  $\text{Retr}_{x_t}(\cdot)$ , we have the update  $x_{t+1} = \text{Retr}_{x_t}(\Delta_t)$ . Let us also define  $\gamma(s) = \text{Retr}_{x_t}(s\Delta_t)$  for  $s \in [0, 1]$  and  $g(s) = f(\gamma(s))$ . We then have

$$g(1) - g(0) = f(x_{t+1}) - f(x_t) = \int_0^1 g'(s) ds = \int_0^1 \langle \text{grad}f(\gamma(s)), \gamma'(s) \rangle ds = \mathbb{E}_s[\langle \text{grad}f(\gamma(s)), \gamma'(s) \rangle],$$

for  $s \sim \text{unif}[0, 1]$ .

Let  $g_t = \text{grad}F(w_t, \nu_t)$  and  $\nabla_t = \mathbb{E}[g_t] \in T_{w_t}\mathcal{M}$ . Since parallel transport preserves the inner product, we have the following relationship:

$$\begin{aligned} \mathbb{E}[f(x_{t+1}) - f(x_t)] &= \mathbb{E}[\langle P_{w_t, x_{t+1}}^g(g_t), P_{w_t, x_{t+1}}^g(\gamma'(s_t)) \rangle] \\ &= \mathbb{E}[\langle P_{w_t, x_{t+1}}^g(g_t), P_{x_t, x_{t+1}}^g(\Delta_t) - u_{t+1} \rangle] + \mathbb{E}[\langle P_{w_t, x_{t+1}}^g(g_t), u_{t+1} \rangle] \\ &\quad + \mathbb{E}[\langle P_{w_t, x_{t+1}}^g(g_t), P_{w_t, x_{t+1}}^g(\gamma'(s_t)) - P_{x_t, x_{t+1}}^g(\Delta_t) \rangle], \end{aligned}$$

where the expectation is over  $s$  and  $\nu$ .

We next introduce a series of vectors  $u_t \in T_{x_t}\mathcal{M}$  to facilitate the analysis. We can then decompose  $\mathbb{E}[f(x_{t+1}) - f(x_t)]$  into three parts and analyze each part separately.

Let us choose  $u_0 = -D \frac{\sum_{\tau=0}^{T-1} (\mathcal{P}_{S_{\tau+1}}^s)^{-1} \circ P_{w_\tau, x_{\tau+1}}^g(\nabla_\tau)}{\left\| \sum_{\tau=0}^{T-1} (\mathcal{P}_{S_{\tau+1}}^s)^{-1} \circ P_{w_\tau, x_{\tau+1}}^g(\nabla_\tau) \right\|}$  and  $u_t = \mathcal{P}_{S_t}^s(u_0)$ . Summing above over  $t = 0, \dots, T-1$ , we derive

$$\begin{aligned} \mathbb{E}[f(x_T) - f(x_0)] &= \underbrace{\sum_{t=0}^{T-1} \mathbb{E}[\langle P_{w_t, x_{t+1}}^g(g_t), P_{x_t, x_{t+1}}^g(\Delta_t) - u_{t+1} \rangle]}_{\text{Term A}} \\ &\quad + \underbrace{\sum_{t=0}^{T-1} \mathbb{E}[\langle P_{w_t, x_{t+1}}^g(g_t), u_{t+1} \rangle]}_{\text{Term B}} \\ &\quad + \underbrace{\sum_{t=0}^{T-1} \mathbb{E}[\langle P_{w_t, x_{t+1}}^g(g_t), P_{w_t, x_{t+1}}^g(\gamma'(s_t)) - P_{x_t, x_{t+1}}^g(\Delta_t) \rangle]}_{\text{Term C}}. \end{aligned}$$

We next provide analytical bounds for each of the three terms respectively.

#### Bound on Term A.

Recall that the ball of radius  $D$  in  $T_{x_t}\mathcal{M}$  is denoted by  $\mathbb{B}_{T_{x_t}\mathcal{M}}(D)$ . We denote the projection operator to  $\mathbb{B}_{T_{x_t}\mathcal{M}}(D)$  by  $\text{Proj}_{\mathcal{D}_{x_t}}$ . Then, the update rule with parallel transport can be written as  $\Delta_{t+1} = \text{Proj}_{\mathcal{D}_{x_{t+1}}}(P_{x_t, x_{t+1}}^g(\Delta_t) - \eta P_{w_t, x_{t+1}}^g(g_t))$ . We have  $u_{t+1} = P_{x_t, x_{t+1}}^g(u_t)$  and due to the projection properties of the convex sets, we get

$$\|\Delta_{t+1} - u_{t+1}\|^2 \leq \|P_{x_t, x_{t+1}}^g(\Delta_t) - u_{t+1} - \eta P_{w_t, x_{t+1}}^g(g_t)\|^2 = \|P_{x_t, x_{t+1}}^g(\Delta_t - u_t) - \eta P_{w_t, x_{t+1}}^g(g_t)\|^2.$$

Rearranging the terms for any  $\eta > 0$ , we obtain

$$\langle P_{w_t, x_{t+1}}^g(g_t), P_{x_t, x_{t+1}}^g(\Delta_t - u_t) \rangle \leq \frac{1}{2\eta} (\|P_{x_t, x_{t+1}}^g(\Delta_t - u_t)\|^2 - \|\Delta_{t+1} - u_{t+1}\|^2) + \frac{\eta}{2} \|P_{w_t, x_{t+1}}^g(g_t)\|^2.$$

Since parallel transport is isometric, we have  $\|P_{x_t, x_{t+1}}^g(\Delta_t - u_t)\| = \|\Delta_t - u_t\|$ . Summing above over  $t = 0, \dots, T-1$  gives

$$\begin{aligned} \sum_{t=0}^{T-1} \langle P_{w_t, x_{t+1}}^g(g_t), P_{x_t, x_{t+1}}^g(\Delta_t - u_t) \rangle &\leq \sum_{t=0}^{T-1} \frac{1}{2\eta} (\|\Delta_t - u_t\|^2 - \|\Delta_{t+1} - u_{t+1}\|^2) + \frac{\eta}{2} \sum_{t=0}^{T-1} \|P_{w_t, x_{t+1}}^g(g_t)\|^2 \\ &\leq \frac{1}{2\eta} (\|\Delta_0 - u_0\|^2 - \|\Delta_T - u_T\|^2) + \frac{\eta}{2} \sum_{t=0}^{T-1} \|P_{w_t, x_{t+1}}^g(g_t)\|^2. \end{aligned}$$

Noting that  $\eta = \frac{D}{G\sqrt{T}}$ , we take expectation from above and use the bound  $G^2$  on the stochastic gradient to get

$$\text{Term A} \leq \frac{1}{2\eta} D^2 + \frac{\eta}{2} G^2 T = DG\sqrt{T}. \quad (6)$$

#### Bound on Term B.

For the ease of notation, we write the stochastic component of the gradient as  $\varepsilon_t := g_t - \nabla_t$ . Due to the isometry of parallel transport operation, we have

$$\begin{aligned} \sum_{t=0}^{T-1} \langle P_{w_t, x_{t+1}}^g(g_t), u_{t+1} \rangle &= \sum_{t=0}^{T-1} \langle (\mathcal{P}_{S_{t+1}}^s)^{-1} \circ P_{w_t, x_{t+1}}^g(g_t), (\mathcal{P}_{S_{t+1}}^s)^{-1}(u_{t+1}) \rangle \\ &= \langle u_0, \sum_{t=0}^{T-1} (\mathcal{P}_{S_{t+1}}^s)^{-1} \circ P_{w_t, x_{t+1}}^g(g_t) \rangle \\ &= -DT \left\| \frac{1}{T} \sum_{t=0}^{T-1} (\mathcal{P}_{S_{t+1}}^s)^{-1} \circ P_{w_t, x_{t+1}}^g(\nabla_t) \right\| + \langle u_0, \sum_{t=0}^{T-1} (\mathcal{P}_{S_{t+1}}^s)^{-1} \circ P_{w_t, x_{t+1}}^g(\varepsilon_t) \rangle. \end{aligned}$$

For the second term in the last line, since parallel transport operation is linear and the noise has mean zero, we have  $\mathbb{E}[P_{x,y}^g(\varepsilon)] = P_{x,y}^g(\mathbb{E}[\varepsilon]) = 0$ . Then, we can use the independence of noise and isometry of parallel transport, such that

$$\begin{aligned} \mathbb{E} \left[ \left\| \sum_{t=0}^{T-1} (\mathcal{P}_{S_{t+1}}^s)^{-1} \circ P_{w_t, x_{t+1}}^g(\varepsilon_t) \right\|^2 \right] &= \mathbb{E} \left[ \sum_{t=0}^{T-1} \|(\mathcal{P}_{S_{t+1}}^s)^{-1} \circ P_{w_t, x_{t+1}}^g(\varepsilon_t)\|^2 \right] \\ &= \sum_{t=0}^{T-1} \mathbb{E} [\|\varepsilon_t\|^2] = T\sigma^2. \end{aligned}$$

Therefore, applying Cauchy inequality and using  $\|u_0\| = D$ , we get

$$\begin{aligned} \mathbb{E} \left[ \langle u_0, \sum_{t=0}^{T-1} (\mathcal{P}_{S_{t+1}}^s)^{-1} \circ P_{w_t, x_{t+1}}^g(\varepsilon_t) \rangle \right] &\leq D \mathbb{E} \left[ \left\| \sum_{t=0}^{T-1} (\mathcal{P}_{S_{t+1}}^s)^{-1} \circ P_{w_t, x_{t+1}}^g(\varepsilon_t) \right\| \right] \\ &\leq D \sqrt{\mathbb{E} \left[ \left\| \sum_{t=0}^{T-1} (\mathcal{P}_{S_{t+1}}^s)^{-1} \circ P_{w_t, x_{t+1}}^g(\varepsilon_t) \right\|^2 \right]} \\ &\leq D\sigma\sqrt{T}. \end{aligned}$$

Combining the previous equations, we provide the following bound for the second term as

$$\text{Term B} \leq -DT \mathbb{E} \left[ \left\| \frac{1}{T} \sum_{t=0}^{T-1} (\mathcal{P}_{S_{t+1}}^s)^{-1} \circ P_{w_t, x_{t+1}}^g(\nabla_t) \right\| \right] + D\sigma\sqrt{T}.$$

#### Bound on Term C.

For the third term, given the bound on the stochastic gradient, we have

$$\mathbb{E} [\langle P_{w_t, x_{t+1}}^g(g_t), P_{w_t, x_{t+1}}^g(\gamma'(s_t)) - P_{x_t, x_{t+1}}^g(\Delta_t) \rangle] \leq G \mathbb{E} [\|P_{w_t, x_{t+1}}^g(\gamma'(s_t)) - P_{x_t, x_{t+1}}^g(\Delta_t)\|].$$

We write  $\gamma'(s_t) = P_{x_t, w_t}^g(\Delta_t) + v$ , where  $v \in T_{w_t}\mathcal{M}$ . Then, applying Lemma 2.8, since  $\|\Delta_t\| \leq D$  we get

$$\begin{aligned} \|P_{w_t, x_{t+1}}^g(\gamma'(s_t)) - P_{x_t, x_{t+1}}^g(\Delta_t)\| &= \|P_{w_t, x_{t+1}}^g \circ P_{x_t, w_t}^g(\Delta_t) - P_{x_t, x_{t+1}}^g(\Delta_t) + P_{w_t, x_{t+1}}^g(v)\| \\ &\leq \|\Delta_t - P_{x_{t+1}, x_t}^g \circ P_{w_t, x_{t+1}}^g \circ P_{x_t, w_t}^g(\Delta_t)\| + \|v\| \\ &\leq CD \times \text{length}(x_t, w_t, x_{t+1}, x_t) + \|v\| \end{aligned}$$

The length of the path  $\{x_t, w_t, x_{t+1}, x_t\}$  is less than  $3D$ . So the first term is bounded by  $3CD^2$ .

We now proceed to bounding  $\|v\|$ . We know that  $\gamma(s) = \text{Retr}_{x_t}(s\Delta_t)$ ,  $\gamma'(0) = \Delta_t$ , and  $\|\Delta_t\| \leq D$ . Since we have  $\|\gamma''(s)\| \leq C'D^2$  for  $s \in [0, 1]$  with  $C'$  defined in Assumption 2.5, we obtain

$$\|\Delta_t - \gamma'(s_t)\| \leq \int_0^{s_t} \|\gamma''(\tau)\| d\tau \leq C'D^2. \quad (7)$$

On the other hand, consider the unit-speed geodesic  $\gamma_2(s)$  connecting  $x_t$  and  $w_t$  and let  $z(s)$  be the parallel transported vector field with  $z(0) = \Delta_t$  along  $\gamma_2(s)$ . Since  $z(s)$  is parallel transported it has 0 intrinsic acceleration. The difference  $\Delta_t - P_{x_t, w_t}^g(\Delta_t)$  comes from the extrinsic acceleration. We have

$$\|\Delta_t - P_{x_t, w_t}^g(\Delta_t)\| = \left\| \int_0^{\text{dist}(x_t, w_t)} \Pi(z(s), \gamma_2'(s)) ds \right\| \leq C \text{dist}(x_t, w_t) \|z(0)\| \leq CD^2. \quad (8)$$

Combining the previous two inequalities, we get

$$\|v\| = \|\gamma'(s_t) - P_{x_t, w_t}^g(\Delta_t)\| \leq \|\gamma'(s_t) - \Delta_t\| + \|\Delta_t - P_{x_t, w_t}^g(\Delta_t)\| \leq (C + C')D^2.$$

Therefore, for the third term we have the following bound

$$\text{Term C} \leq (GC' + 4GC)D^2T. \quad (9)$$

When we sum the bounds for three terms we obtain

$$\begin{aligned} \mathbb{E} \left[ \left\| \frac{1}{T} \sum_{t=0}^{T-1} (\mathcal{P}_{S_{t+1}}^s)^{-1} \circ P_{w_t, x_{t+1}}^g(\nabla_t) \right\| \right] &\leq \frac{1}{DT} \mathbb{E}[f(x_0) - f(x_T)] \\ &\quad + \frac{1}{DT} (D\sigma\sqrt{T} + DG\sqrt{T} + (GC' + 4GC)D^2T). \end{aligned} \quad (10)$$

This holds for any  $k \in [K]$  and we take the average over  $K$  epochs, where summation of  $f(x_0) - f(x_T)$  terms telescope due to initialization at each epoch. Then

$$\begin{aligned} \mathbb{E}_k \left[ \left\| \frac{1}{T} \sum_{t=0}^{T-1} (\mathcal{P}_{S_{k,t+1}}^s)^{-1} \circ P_{w_{k,t}, x_{k,t+1}}^g(\nabla_{k,t}) \right\| \right] &= \frac{1}{K} \sum_{k=1}^K \mathbb{E} \left[ \left\| \frac{1}{T} \sum_{t=0}^{T-1} (\mathcal{P}_{S_{k,t+1}}^s)^{-1} \circ P_{w_{k,t}, x_{k,t+1}}^g(\nabla_{k,t}) \right\| \right] \\ &\leq \frac{f(x_{1,0}) - \mathbb{E}[f(x_{K,T})]}{DTK} + \frac{\sigma + G}{\sqrt{T}} + (GC' + 4GC)D. \end{aligned}$$

By using Lemma 2.9, defining  $\theta := f(x_{1,0}) - \mathbb{E}[f(x_{K,T})]$ , and recalling that  $\delta = DT$  and  $N = KT$ , we obtain

$$\begin{aligned} \mathbb{E}[\|\text{grad}f(w_{out})\|_\delta] &\leq \mathbb{E}_k \left[ \left\| \frac{1}{T} \sum_{t=0}^{T-1} (\mathcal{P}_{S_{k,t+1}}^s)^{-1} \circ P_{w_{k,t}, x_{k,t+1}}^g(\nabla_{k,t}) \right\| \right] + 3CL\delta \\ &\leq \frac{\theta}{DTK} + \frac{\sigma + G}{\sqrt{T}} + (GC' + 4GC)D + 3CL\delta \\ &\leq \frac{\theta T}{\delta N} + \frac{\sigma + G}{\sqrt{T}} + (GC' + 4GC)D + 3CL\delta. \end{aligned}$$

We now optimize the above bound over  $T$  by choosing  $T = (\delta N)^{\frac{2}{3}}$ . Since  $\delta \leq 1$  we also have  $D \leq (\delta N)^{-\frac{1}{3}}$ . As a result, we get the final bound

$$\mathbb{E}[\|\text{grad}f(w_{out})\|_\delta] \leq (\theta + \sigma + G + GC' + 4GC)(\delta N)^{-\frac{1}{3}} + 3CL\delta = C_1(\delta N)^{-\frac{1}{3}} + 3C_2L\delta. \quad (11)$$

where  $C_1 := \theta + \sigma + G + GC' + 4GC$  and  $C_2 := C$ .  $\square$

#### A.4 Proof of Theorem 3.3

*Proof.* In this part, we consider the case where inner products belong to the ambient space, so we can write the inner products without transporting the vectors to the same tangent space. For  $T$  iterations in an epoch, we omit the subscripts  $k$ . Similar to the proof of Theorem 3.1, we can decompose the difference between consecutive iterates as follows

$$\mathbb{E}[f(x_{t+1}) - f(x_t)] = \mathbb{E}[\langle g_t, \gamma'(s_t) \rangle] = \mathbb{E}[\langle g_t, \Delta_t - u_t \rangle] + \mathbb{E}[\langle g_t, u_t \rangle] + \mathbb{E}[\langle g_t, \gamma'(s_t) - \Delta_t \rangle].$$

Summing above over  $t \in \{0, \dots, T-1\}$ , we get

$$\mathbb{E}[f(x_T) - f(x_0)] = \underbrace{\mathbb{E}\left[\sum_{t=0}^{T-1} \langle g_t, \Delta_t - u_t \rangle\right]}_{\text{Term D}} + \underbrace{\mathbb{E}\left[\sum_{t=0}^{T-1} \langle g_t, u_t \rangle\right]}_{\text{Term E}} + \underbrace{\mathbb{E}\left[\sum_{t=0}^{T-1} \langle g_t, \gamma'(s_t) - \Delta_t \rangle\right]}_{\text{Term F}}.$$

Let us recall that  $\mathcal{D}_x := \mathbb{B}_{T_x \mathcal{M}}(D)$  represents a ball with radius  $D$  in  $T_x \mathcal{M}$  and  $\text{Proj}_{\mathcal{D}_x} : T_x \mathcal{M} \rightarrow T_x \mathcal{M}$  denotes the projection operator onto  $\mathcal{D}_x$ . Also, recall that for the projection case, the update rule of  $\Delta_t$  is given as  $\Delta_{t+1} = \text{Proj}_{\mathcal{D}_{x_{t+1}}} \text{Proj}_{T_{x_{t+1}} \mathcal{M}}(\Delta_t - \eta g_t)$ . We next bound the three terms above respectively.

##### Bound on Term D.

First, define  $u_t = \text{Proj}_{T_{x_t} \mathcal{M}}(u)$  where  $u = -D \sum_{t=0}^{T-1} \nabla_t / \|\sum_{t=0}^{T-1} \nabla_t\|$ . We start with the update equation where  $\Delta_{t+1} = \text{Proj}_{\mathcal{D}_{x_{t+1}}} \text{Proj}_{T_{x_{t+1}} \mathcal{M}}(\Delta_t - \eta g_t)$ . Subtracting  $u_t$  from  $\Delta_{t+1}$ , we obtain

$$\Delta_{t+1} - u_t = (\text{Proj}_{\mathcal{D}_{x_{t+1}}} \text{Proj}_{T_{x_{t+1}} \mathcal{M}}(\Delta_t - \eta g_t) - \text{Proj}_{T_{x_{t+1}} \mathcal{M}}(u_t)) + (\text{Proj}_{T_{x_{t+1}} \mathcal{M}}(u_t) - u_t). \quad (12)$$

We know that  $(v - \text{Proj}_{T_x \mathcal{M}} v) \perp T_x \mathcal{M}$  for any  $x \in \mathcal{M}$ , so two terms in the above equation are perpendicular to each other. By taking square on both sides, we get

$$\|\Delta_{t+1} - u_t\|^2 = \|\text{Proj}_{\mathcal{D}_{x_{t+1}}} \text{Proj}_{T_{x_{t+1}} \mathcal{M}}(\Delta_t - \eta g_t) - \text{Proj}_{T_{x_{t+1}} \mathcal{M}}(u_t)\|^2 + \|\text{Proj}_{T_{x_{t+1}} \mathcal{M}}(u_t) - u_t\|^2. \quad (13)$$

Notice that  $\|\text{Proj}_{T_{x_{t+1}} \mathcal{M}}(u_t)\| \leq \|u_t\| \leq \|u\| = D$ , so  $\text{Proj}_{T_{x_{t+1}} \mathcal{M}}(u_t) \in \mathcal{D}_{x_{t+1}}$ .

On the other hand, since  $\mathcal{D}_{x_{t+1}}$  is convex, we have  $\|\text{Proj}_{\mathcal{D}_{x_{t+1}}}(a) - b\|^2 \leq \|a - b\|^2$  for any  $a \in T_{x_{t+1}} \mathcal{M}$  and  $b \in \mathcal{D}_{x_{t+1}}$ . Applying this in Equation 13 and using the contraction property of projection, we have

$$\|\Delta_{t+1} - u_t\|^2 \leq \|\Delta_t - \eta g_t - u_t\|^2 + \|\text{Proj}_{T_{x_{t+1}} \mathcal{M}}(u_t) - u_t\|^2. \quad (14)$$

Rearranging the terms, for any  $\eta > 0$ , we obtain

$$\langle g_t, \Delta_t - u_t \rangle \leq \frac{1}{2\eta} (\|\Delta_t - u_t\|^2 - \|\Delta_{t+1} - u_t\|^2) + \frac{\eta}{2} \|g_t\|^2 + \frac{1}{2\eta} \|u_t - \text{Proj}_{T_{x_{t+1}} \mathcal{M}}(u_t)\|^2. \quad (15)$$

Summing above over  $t$  gives

$$\begin{aligned} \sum_{t=0}^{T-1} \langle g_t, \Delta_t - u_t \rangle &\leq \frac{1}{2\eta} (\|\Delta_0 - u_0\|^2 - \|\Delta_T - u_{T-1}\|^2) \\ &\quad + \frac{1}{2\eta} \sum_{t=1}^{T-1} (\|\Delta_t - u_t\|^2 - \|\Delta_t - u_{t-1}\|^2) \\ &\quad + \frac{\eta}{2} \sum_{t=0}^{T-1} \|g_t\|^2 + \frac{1}{2\eta} \sum_{t=0}^{T-1} \|u_t - \text{Proj}_{T_{x_{t+1}} \mathcal{M}}(u_t)\|^2. \end{aligned}$$

Since  $\|u_0\| \leq D$  and  $\Delta_0 = 0$ , we have  $\|\Delta_0 - u_0\|^2 - \|\Delta_T - u_{T-1}\|^2 \leq D^2$ . Since  $\|u_t\| \leq D$ , we can also write

$$\|\Delta_t - u_t\|^2 - \|\Delta_t - u_{t-1}\|^2 = \langle u_{t-1} - u_t, 2\Delta_t - u_t - u_{t-1} \rangle \leq 4D\|u_t - u_{t-1}\|. \quad (16)$$

According to Lemma A.1, for  $v \in T_x \mathcal{M}$  and  $\|v - \text{Proj}_{T_y \mathcal{M}}(v)\| \leq 2mC\|v\|\text{dist}(x, y)$ , so we have

$$\|u_t - \text{Proj}_{T_{x_{t+1}} \mathcal{M}}(u_t)\| \leq 2mC\|u_t\|\text{dist}(x_t, x_{t+1}) \leq 2mCD^2, \quad (17)$$

given that  $\text{dist}(x_t, x_{t+1}) \leq D$  and  $\|u_t\| \leq D$ . By the same token, we have  $\|u_t - u_{t-1}\| \leq 2mCD^2$ .

Combining all the previous inequalities, we get

$$\mathbb{E} \left[ \sum_{t=0}^{T-1} \langle g_t, \Delta_t - u_t \rangle \right] \leq \frac{D^2}{2\eta} + \frac{4mCD^3T}{\eta} + \frac{\eta}{2}G^2T + \frac{1}{2\eta}4m^2C^2D^4T, \quad (18)$$

since  $\mathbb{E}[\|g_t\|^2] \leq G$ . Again, we choose  $\eta = \frac{D}{G\sqrt{T}}$ . Since  $\delta \leq 1$ ,  $T > 1$ , we have  $DT = \delta \leq 1$  and  $D^2T \leq 1$ . Therefore, we can simplify above to derive

$$\begin{aligned} \mathbb{E} \left[ \sum_{t=0}^T \langle g_t, \Delta_t - u_t \rangle \right] &\leq \frac{DG\sqrt{T}}{2} + 4mCD^2GT\sqrt{T} + \frac{DG\sqrt{T}}{2} + 2m^2C^2D^3GT\sqrt{T} \\ &\leq DG(1 + 4Cm)\sqrt{T} + 2m^2D^2TGC^2. \end{aligned} \quad (19)$$

#### Bound on Term E.

Recall that  $u = -D \sum_{t=0}^{T-1} \nabla_t / \|\sum_{t=0}^{T-1} \nabla_t\|$  and  $u_t = \text{Proj}_{T_{x_t} \mathcal{M}}(u)$ , and define  $u'_t := \text{Proj}_{T_{w_t} \mathcal{M}}(u)$ . We know that  $\|u_t - u'_t\| \leq 2mC\|u\|\text{dist}(x_t, w_t) \leq 2mCD^2$  using Lemma A.1. Also  $\nabla_t \in T_{w_t} \mathcal{M}$  and  $u - u'_t \perp \nabla_t$  so  $\langle \nabla_t, u'_t \rangle = \langle \nabla_t, u \rangle$ . Given that  $\varepsilon_t = g_t - \nabla_t$  and  $\langle \varepsilon_t, u'_t \rangle = \langle \varepsilon_t, u \rangle$ , the second term can be written as

$$\begin{aligned} \mathbb{E} \left[ \sum_{t=0}^{T-1} \langle g_t, u_t \rangle \right] &= \mathbb{E} \left[ \sum_{t=0}^{T-1} \langle \nabla_t, u'_t \rangle \right] + \mathbb{E} \left[ \sum_{t=0}^{T-1} \langle \varepsilon_t, u'_t \rangle \right] + \mathbb{E} \left[ \sum_{t=0}^{T-1} \langle g_t, u_t - u'_t \rangle \right] \\ &\leq -DT \mathbb{E} \left[ \left\| \frac{1}{T} \sum_{t=0}^{T-1} \nabla_t \right\| \right] + D\sigma\sqrt{T} + 2mGCD^2T. \end{aligned}$$

#### Bound on Term F.

Recalling equation 7 and using the bound on the stochastic gradient, the third term be bounded such that

$$\mathbb{E} \left[ \sum_{t=0}^{T-1} \langle g_t, \gamma'(s_t) - \Delta_t \rangle \right] \leq C'GD^2T.$$

When we sum all the inequalities for the terms in the decomposition of  $\mathbb{E}[f(x_T) - f(x_0)]$  we obtain the following inequality

$$\mathbb{E} \left[ \left\| \frac{1}{T} \sum_{t=0}^{T-1} \nabla_t \right\| \right] \leq \frac{f(x_0) - \mathbb{E}[f(x_T)]}{DT} + \frac{1}{\sqrt{T}}(\sigma + G + 4mGC) + DG(2Cm + C' + 2m^2C^2).$$

After taking average over  $K$  epochs, we can bound each term with the previous results. We note that the first term will telescope (due to initialization) while the other terms can be bounded independent of  $k$ . Therefore,

$$\begin{aligned} \mathbb{E} \left[ \frac{1}{K} \sum_{k=1}^K \left\| \frac{1}{T} \sum_{t=0}^{T-1} \nabla_{k,t} \right\| \right] &\leq \frac{f(x_{1,0}) - \mathbb{E}[f(x_{K,T})]}{DTK} + \frac{1}{\sqrt{T}}(\sigma + G + 4mGC) \\ &\quad + DG(2mC + C' + 2m^2C^2) \\ &\leq \frac{\theta T}{\delta N} + \frac{1}{\sqrt{T}}(\sigma + G + 4mGC) + DG(2mC + C' + 2m^2C^2), \end{aligned}$$

where the last line is by recalling that  $\delta = DT$  and  $N = KT$ . Next, we choose  $T = (\delta N)^{\frac{2}{3}}$  to get

$$\mathbb{E} \left[ \frac{1}{K} \sum_{k=1}^K \left\| \frac{1}{T} \sum_{t=0}^{T-1} \nabla_{k,t} \right\| \right] \leq (\delta N)^{-\frac{1}{3}}(\theta + \sigma + G + 4mGC) + \delta^{\frac{1}{3}}N^{-\frac{2}{3}}(2mGC + GC' + 2m^2GC^2).$$

To link  $\|\text{grad}f(\bar{w})\|_\delta$  and  $\|\frac{1}{T} \sum_{t=0}^{T-1} \nabla_t\|$  we have the following inequality

$$\begin{aligned} \|\text{grad}f(\bar{w})\|_\delta &\leq \left\| \frac{1}{T} \sum_{t=0}^{T-1} P_{w_t, \bar{w}}^g(\text{grad}f(w_t)) \right\| = \left\| \frac{1}{T} \sum_{t=0}^{T-1} P_{w_t, \bar{w}}^g(\nabla_t) \right\| \\ &\leq \left\| \frac{1}{T} \sum_{t=0}^{T-1} \nabla_t \right\| + \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla_t - P_{w_t, \bar{w}}^g(\nabla_t)\| \\ &\leq \left\| \frac{1}{T} \sum_{t=0}^{T-1} \nabla_t \right\| + CL\delta, \end{aligned}$$

where the last line is derived similar to equation 8. As a result, we have

$$\mathbb{E}[\|\text{grad}f(w_{out})\|_\delta] \leq C_3(\delta N)^{-\frac{1}{3}} + C_4\delta^{\frac{1}{3}}N^{-\frac{2}{3}} + \delta LC.$$

where  $C_3 := \theta + \sigma + G + 4mGC$  and  $C_4 := 2mGC + GC' + 2m^2GC^2$ .

□

**Lemma A.1.** *Let  $v$  be a vector in the ambient space and  $x, y \in \mathcal{M}$ . Assume that the second fundamental form is bounded by  $C$  for unit vectors. Then,*

$$\|\text{Proj}_{T_x\mathcal{M}}(v) - \text{Proj}_{T_y\mathcal{M}}(v)\| \leq 2mC\|v\|\text{dist}(x, y),$$

where  $m = n - d$  is the dimension of the normal space.

*Proof.* Let  $\gamma(t)$  be the unit-speed minimizing geodesic connecting  $x$  to  $y$ . Let  $\phi(t) = \text{Proj}_{T_{\gamma(t)}\mathcal{M}}$  be the orthogonal projection operator to  $T_{\gamma(t)}\mathcal{M}$ . Choose an orthonormal frame  $\{n_i(t)\}_{i=1}^m$  from the normal space  $N_{\gamma(t)}\mathcal{M}$ , where  $m = n - d$  is the dimension of the normal space. Then, we can represent  $\phi(t) = I - \sum_{i=1}^m n_i(t)n_i(t)^\top$ .

Differentiating  $\phi(t)$  gives  $\phi'(t) = -\sum_{i=1}^m (n_i'(t)n_i(t)^\top + n_i(t)n_i'(t)^\top)$ . For any  $v$  in the ambient space, we have

$$\|\phi'(t)v\| \leq \sum_{i=1}^m |\langle n_i(t), v \rangle| \|n_i'(t)\| + |\langle n_i'(t), v \rangle| \|n_i(t)\| \leq 2\|v\| \sum_{i=1}^m \|n_i'(t)\|.$$

Hence, we have  $\|\phi'(t)\|_{op} \leq 2 \sum_{i=1}^m \|n_i'(t)\|$ . Next, we relate  $\|n_i'(t)\|$  to the second fundamental form. For each  $n_i(t)$ , the shape (Weingarten) operator  $S_{n_i(t)} : T_{\gamma(t)}\mathcal{M} \rightarrow T_{\gamma(t)}\mathcal{M}$  satisfies  $\langle S_{n_i(t)}(u), w \rangle = \langle \Pi(u, w), n_i \rangle$  and  $\|S_{n_i(t)}\|_{op} \leq C$ . We also have

$$\|n_i'(t)\| = \|S_{n_i(t)}(\gamma'(t))\| \leq \|S_{n_i(t)}\|_{op} \|\gamma'(t)\| \leq C.$$

Thus,  $\sum_{i=1}^m \|n_i'(t)\| \leq mC$ , which results in

$$\|\text{Proj}_{T_x\mathcal{M}} - \text{Proj}_{T_y\mathcal{M}}\|_{op} \leq \int_0^{\text{dist}(x, y)} \|\phi'(t)\|_{op} dt \leq 2mC\text{dist}(x, y).$$

□

## A.5 Proof of Theorem 4.2

Before the proof, we propose some lemmas to streamline the analysis. First, we define the function  $h_\delta$  such that  $\text{grad}h_\delta(x) = \mathbb{E}[g_\delta(x)]$ , and we replace our goal of finding  $(\delta, \epsilon)$ -stationary point of  $f$  with  $(\frac{\delta'}{2}, \epsilon)$ -stationary point of  $h_{\frac{\delta'}{2}}$ , where  $\delta'$  can be chosen to compensate approximation errors. We compute an upper bound on the Lipschitz constant of  $h_\delta$  using Lemmas A.2 and A.3. Next, we compute an upper bound on the distance between  $\text{grad}h_\delta(x)$  and  $\partial_\delta f(x)$  with Lemmas A.4, A.5 and 4.1. Combining these findings, we complete the proof of Theorem 4.2.

Recall the following gradient estimator in the zeroth order setting

$$g_\delta(x) = \frac{d}{2\delta} (F(\text{Exp}_x(\delta u), \nu) - F(\text{Exp}_x(-\delta u), \nu))u,$$

where  $u$  is sampled uniformly from a sphere in  $T_x\mathcal{M}$ . Let us define  $h_\delta : \mathcal{M} \rightarrow \mathbb{R}$  such that  $h_\delta(x) := \int f \circ \text{Exp}_x(u) dp_x(u)$ , where  $p_x$  is a uniform measure over  $\mathbb{B}_{T_x\mathcal{M}}(\delta) \subset T_x\mathcal{M}$ . Considering the function  $f \circ \text{Exp}_x : T_x\mathcal{M} \rightarrow \mathbb{R}$  which is defined over a Euclidean space we have  $\text{grad} h_\delta(x) = \mathbb{E}_u[g_\delta(x)]$  [22, Lemma 1]. We start with the derivation of analytical properties of  $h_\delta(x)$ .

**Lemma A.2** (Lemma 5 in [58], Lemma 1 in [47]). *Let the sectional curvature be bounded by  $K_s$ . For  $x, y \in \mathcal{M}$  and  $w \in T_x\mathcal{M}$  we have*

$$\text{dist}(\text{Exp}_x(w), \text{Exp}_y(P_{x,y}^g(w))) \leq C_7 \text{dist}(x, y), \quad (20)$$

where  $C_7$  depends on  $K_s$ .

By using the above properties we can find the Lipschitz constant of  $h_\delta$  in terms of the Lipschitz constant of  $f$  and  $C_7$ , which depends on the curvature bound  $K_s$ .

**Lemma A.3.** *Let  $f : \mathcal{M} \rightarrow \mathbb{R}$  be an  $L$ -Lipschitz function and define  $h_\delta(x) = \int f \circ \text{Exp}_x(u) dp_x(u)$ , where  $p_x$  is a uniform measure over  $\mathbb{B}_{T_x\mathcal{M}}(\delta) \subset T_x\mathcal{M}$ . Then,  $h_\delta$  is Lipschitz continuous on  $\mathcal{M}$  with a Lipschitz constant bounded by  $LC_7$ .*

*Proof.*

$$\begin{aligned} |h_\delta(x) - h_\delta(y)| &= \left| \int f \circ \text{Exp}_x(u_x) dp_x(u_x) - \int f \circ \text{Exp}_y(P_{x,y}^g(u_x)) dp_x(u_x) \right| \\ &\leq \int L \text{dist}(\text{Exp}_x(u_x), \text{Exp}_y(P_{x,y}^g(u_x))) dp_x(u_x) \\ &\leq LC_7 \text{dist}(x, y), \end{aligned}$$

where in the last inequality we used Lemma A.2. □

We now state the relation between Riemannian gradient of  $h_\delta$  and  $\partial_\delta f$  in terms of Lipschitz constant  $L$ , sectional curvature bound  $K_s$  and  $\delta$ .

**Lemma A.4** (Proposition A.3 in [12]). *Let  $\mathcal{M}$  be a Riemannian manifold whose sectional curvatures are in the interval  $[K_{\min}, K_{\max}]$ , and let  $K_s = \max(|K_{\min}|, |K_{\max}|)$ . Consider  $(x, s) \in T\mathcal{M}$  and the geodesic  $\gamma(t) = \text{Exp}_x(ts)$ . If  $\gamma$  is defined and has no interior conjugate point on the interval  $[0, 1]$ , then*

$$\forall v \in T_x\mathcal{M} \quad \|T_s(v) - P_{x,\gamma(t)}^g(v)\| \leq K_s f_{K_{\min}}(\|s\|) \|v_\perp\|, \quad (21)$$

where  $v_\perp = v - \frac{\langle s, v \rangle}{\langle s, s \rangle} s$  is the component of  $v$  orthogonal to  $s$ , and  $T_s = d\text{Exp}_x(s)$  is the differential of the exponential mapping. If it also holds that  $\|s\| \leq \frac{\pi}{\sqrt{|K_{\min}|}}$ , then

$$\forall v \in T_x\mathcal{M} \quad \|T_s(v) - P_{x,\gamma(t)}^g(v)\| \leq \frac{1}{3} K_s \|s\|^2 \|v_\perp\|. \quad (22)$$

**Lemma A.5.** *The distance between the Riemannian gradient of  $h_\delta$  and the Riemannian  $\delta$ -subdifferential of  $f$  can be bounded as*

$$\text{dist}(\text{grad} h_\delta(x), \partial_\delta f(x)) := \min_{z \in \partial_\delta f(x)} \left\{ \|\text{grad} h_\delta(x) - z\| \right\} \leq \frac{1}{3} K_s L \delta^2.$$

*Proof.* Let  $s, v \in T_x\mathcal{M}$ ,  $y(s) = \text{Exp}_x(s)$  and  $p_x$  be a uniform measure on  $\mathbb{B}_{T_x\mathcal{M}}(\delta)$ . Then, we can write

$$\begin{aligned} df(\text{Exp}_x(s))[d\text{Exp}_x(s)[v]] &= \langle \text{grad} f(y(s)), d\text{Exp}_x(s)[v] \rangle \\ &= \langle P_{y(s),x}^g(\text{grad} f(y(s))), P_{y(s),x}^g(d\text{Exp}_x(s)[v]) \rangle. \end{aligned}$$



Therefore,

$$\begin{aligned}
\langle v, \text{grad} h_\delta(x) \rangle &= dh_\delta(x)[v] = \int_{\mathbb{B}_{T_x \mathcal{M}}(\delta)} df(\text{Exp}_x(s)) [d\text{Exp}_x(s)[v]] dp_x(s) \\
&= \int_{\mathbb{B}_{T_x \mathcal{M}}(\delta)} \langle P_{y(s),x}^g(\text{grad} f(y(s))), P_{y(s),x}^g(d\text{Exp}_x(s)[v]) \rangle dp_x(s) \\
&= \int_{\mathbb{B}_{T_x \mathcal{M}}(\delta)} \langle P_{y(s),x}^g(\text{grad} f(y(s))), v \rangle dp_x(s) \\
&\quad + \int_{\mathbb{B}_{T_x \mathcal{M}}(\delta)} \langle P_{y(s),x}^g(\text{grad} f(y(s))), P_{y(s),x}^g(d\text{Exp}_x(s)[v]) - v \rangle dp_x(s).
\end{aligned}$$

This equality holds for any  $v \in T_x \mathcal{M}$ . So we can write

$$\begin{aligned}
\langle \text{grad} h_\delta(x) - \int_{\mathbb{B}_{T_x \mathcal{M}}(\delta)} P_{y(s),x}^g(\text{grad} f(y(s))) dp_x(s), v \rangle \\
&= \int_{\mathbb{B}_{T_x \mathcal{M}}(\delta)} \langle \text{grad} f(y(s)), d\text{Exp}_x(s)[v] - P_{x,y(s)}^g(v) \rangle dp_x(s) \\
&= \int_{\mathbb{B}_{T_x \mathcal{M}}(\delta)} \langle \text{grad} f(y(s)), (T_s - P_{x,y(s)}^g)[v] \rangle dp_x(s).
\end{aligned}$$

We can choose  $v$  as a unit vector in the direction of  $\text{grad} h_\delta(x) - \int_{\mathbb{B}_{T_x \mathcal{M}}(\delta)} P_{y(s),x}^g(\text{grad} f(y(s))) dp_x(s)$ . Then,

$$\begin{aligned}
&\left\| \text{grad} h_\delta(x) - \int_{\mathbb{B}_{T_x \mathcal{M}}(\delta)} P_{y(s),x}^g(\text{grad} f(y(s))) dp_x(s) \right\| \\
&= \int_{\mathbb{B}_{T_x \mathcal{M}}(\delta)} \langle \text{grad} f(y(s)), (T_s - P_{x,y(s)}^g)[v] \rangle dp_x(s) \\
&\leq \int_{\mathbb{B}_{T_x \mathcal{M}}(\delta)} \|\text{grad} f(y(s))\| \|(T_s - P_{x,y(s)}^g)[v]\| dp_x(s) \\
&\leq \frac{1}{3} L K_s \delta^2.
\end{aligned}$$

In the last inequality, we used the bound in Lemma A.4 and the facts that  $\|v\| = 1$  and  $\|s\| \leq \delta$  when  $dp_x(s) > 0$ . Since  $\int_{\mathbb{B}_{T_x \mathcal{M}}(\delta)} P_{y(s),x}^g(\text{grad} f(y(s))) dp_x(s)$  is an element of  $\partial_\delta f(x)$  the above implies that

$$\text{dist}(\text{grad} h_\delta(x), \partial_\delta f(x)) \leq \text{dist}\left(\text{grad} h_\delta(x), \int_{\mathbb{B}_{T_x \mathcal{M}}(\delta)} P_{y(s),x}^g(\text{grad} f(y(s))) dp_x(s)\right) \leq \frac{1}{3} L K_s \delta^2.$$

□

In the remaining part of the proof of Theorem 4.2, we establish the connection between  $\partial_{\frac{\delta}{2}} h_{\frac{\delta}{2}}(x)$  and  $\partial_\delta f(x)$ . By Definition 2.2 we have

$$\begin{aligned}
\partial_{\frac{\delta}{2}} h_{\frac{\delta}{2}}(x) &:= \text{cl conv}\{P_{y,x}^g(\partial h_{\frac{\delta}{2}}(y)) : y \in \text{cl } B(x, \frac{\delta}{2})\} \\
&\subseteq \text{cl conv}\{P_{y,x}^g(P_{z,y}^g(\partial f(z))) : y \in \text{cl } B(x, \frac{\delta}{2}) \text{ and } z \in \text{cl } B(y, \frac{\delta}{2})\} + \mathbb{B}_{T_x \mathcal{M}}(\frac{1}{12} K_s L \delta^2) \\
&\subseteq \text{cl conv}\{P_{y,x}^g(\partial f(y)) : y \in \text{cl } B(x, \delta)\} + \mathbb{B}_{T_x \mathcal{M}}(\frac{1}{12} K_s L \delta^2 + 2CL\delta).
\end{aligned}$$

In the first inclusion we used Lemma A.5 and in the second one we used Lemma 4.1. This result indicates that

$$\|\text{grad} f(w_{\text{out}})\|_\delta \leq \|\text{grad} h_{\frac{\delta}{2}}(w_{\text{out}})\|_{\frac{\delta}{2}} + 2CL\delta + \frac{1}{12} L K_s \delta^2. \quad (23)$$

Hence, we can work on  $h_{\frac{\delta}{2}}$ , which is  $LC_7$ -Lipschitz.

A crucial point in the proof of zeroth order estimator is the bound on the second moment  $\mathbb{E}[\|g_\delta(x)\|^2]$ . This is established by considering the function  $F(\text{Exp}_x(u), \nu)$ , which is  $L(\nu)$ -Lipschitz in its first argument, and  $u$  belongs to a Euclidean space. By applying [45] Lemma E.1] we have the bound  $\mathbb{E}[\|g_\delta(x)\|^2] \leq 16\sqrt{2\pi d}(L \frac{\sinh(\sqrt{K_s}\delta)}{\sqrt{K_s}\delta})^2$ .

To prove this statement, we will follow the same approach in the proof of [45] Lemma E.1]. We consider the function  $h(u) = F(\text{Exp}_x(\delta u), \nu)$  with  $\|u\| \leq 1$ . We have

$$|h(u) - h(v)| \leq L(\nu) \text{dist}(\text{Exp}_x(\delta u), \text{Exp}_x(\delta v)) \leq \frac{\sinh(\sqrt{K_s}\delta)}{\sqrt{K_s}\delta} L(\nu) \|\delta(u - v)\|,$$

where  $K_s$  is the bound on the sectional curvature, and the curvature related term arises due to Jacobi field comparison analysis [41] Theorem 11.9]. As we work on  $h_{\frac{\delta}{2}}$ , the term  $\frac{\sinh(\sqrt{K_s}\delta)}{\sqrt{K_s}\delta}$  can be bounded by  $\frac{\sinh(\sqrt{K_s}/2)}{\sqrt{K_s}/2}$  since  $\delta \leq 1$ , and  $\frac{\sinh(x)}{x}$  is monotonically increasing for  $x > 0$ .

Updating Theorem 3.1 with the parameters of  $h_{\frac{\delta}{2}}$ , i.e.,  $G \leftarrow \frac{\sinh(\sqrt{K_s}/2)}{\sqrt{K_s}/2} L \sqrt{16\sqrt{2\pi d}}$  and  $L \leftarrow LC_7$ , we have

$$\mathbb{E}[\|\text{grad} h_{\frac{\delta}{2}}(w_{out})\|_{\frac{\delta}{2}}] \leq C_5(\delta N)^{-\frac{1}{3}} + \frac{3}{2}\delta LC_7 C. \quad (24)$$

where  $C_5 = 2^{1/3}(\theta + L \frac{\sinh(\sqrt{K_s}/2)}{\sqrt{K_s}/2} \sqrt{16\sqrt{2\pi d}}(2 + C' + 4C))$ . Therefore, we get

$$\begin{aligned} \mathbb{E}[\|\text{grad} f(w_{out})\|_{\delta}] &\leq \mathbb{E}[\|\text{grad} h_{\frac{\delta}{2}}(w_{out})\|_{\frac{\delta}{2}}] + 2CL\delta + \frac{1}{12}LK_s\delta^2 \\ &\leq C_5(\delta N)^{-\frac{1}{3}} + \frac{3}{2}\delta LC_7 C + 2CL\delta + \frac{1}{12}LK_s\delta^2 \\ &\leq C_5(\delta N)^{-\frac{1}{3}} + \delta LC_6, \end{aligned}$$

where  $C_6 := C(2 + \frac{3}{2}C_7) + \frac{1}{12}K_s$  by using  $\delta < 1$ .

For a given  $\delta$ , we can choose

$$\delta' = \min \left\{ \delta, \frac{\epsilon}{2LC_6} \right\},$$

and we can find a  $(\delta', \epsilon)$ -stationary point in  $N = O(\delta'^{-1}\epsilon^{-3})$  iterations.

## A.6 Proof of Lemma 4.1

*Proof.* We know that  $\mathbb{E}_u[g_\delta(x)] = \text{grad} h_\delta(x)$ , and as we proved in Lemma A.5, we have  $\text{dist}(\text{grad} h_\delta(x), \partial_\delta f(x)) \leq \frac{1}{3}K_s L \delta^2$ . Hence,  $\mathbb{E}_u[g_\delta(x)] \in \partial_\delta f(x) + \mathbb{B}_{T_x \mathcal{M}}(\frac{1}{3}K_s L \delta^2)$ .

In the second part of the proof we want to show that  $P_{x_t, y}^g(\partial_{\frac{\delta}{2}} f(x_t)) \subset \partial_\delta f(y) + \mathbb{B}_{T_y \mathcal{M}}(2CL\delta)$ .

Consider an element  $v_{x_t} \in \partial_{\frac{\delta}{2}} f(x_t)$  such that it can be written as  $v_{x_t} = \sum_{i=1}^k \lambda_i P_{z_i, x_t}^g(v_{z_i})$  where  $0 \leq \lambda_i \leq 1$ ,  $\sum_{i=1}^k \lambda_i = 1$ ,  $\text{dist}(x_t, z_i) \leq \frac{\delta}{2}$  and  $v_{z_i} \in \partial f(z_i)$ . Now, consider  $v_y = \sum_{i=1}^k \lambda_i P_{z_i, y}^g(v_{z_i}) \in \partial_\delta f(y)$ . We can show that

$$\begin{aligned} \|P_{x_t, y}^g(v_{x_t}) - v_y\| &= \left\| \sum_{i=1}^k \lambda_i (P_{x_t, y}^g \circ P_{z_i, x_t}^g(v_{z_i}) - P_{z_i, y}^g(v_{z_i})) \right\| \\ &= \left\| \sum_{i=1}^k \lambda_i (P_{y, z_i}^g \circ P_{x_t, y}^g \circ P_{z_i, x_t}^g(v_{z_i}) - v_{z_i}) \right\| \\ &\leq \sum_{i=1}^k \lambda_i \|P_{y, z_i}^g \circ P_{x_t, y}^g \circ P_{z_i, x_t}^g(v_{z_i}) - v_{z_i}\| \\ &\leq \sum_{i=1}^k \lambda_i 2LC\delta = 2LC\delta, \end{aligned}$$

where we used Lemma 2.8 with  $\text{length}(z_i, x_t, y, z_i) \leq 2\delta$ .

As we show that for any  $v_{x_t} \in \partial_{\frac{\delta}{2}} f(x_t)$  there exists  $v_y \in \partial_{\delta} f(y)$  such that  $\|P_{x_t, y}^g(v_{x_t}) - v_y\| \leq 2L\delta C$ , we conclude that  $P_{x_t, y}^g(\partial_{\frac{\delta}{2}} f(x_t)) \subseteq \partial_{\delta} f(y) + \mathbb{B}_{T_y \mathcal{M}}(2CL\delta)$ .  $\square$

## B Discussion on Assumption 2.5

### B.1 Generality of the Assumption 2.5

In this section, we show that an extension of Assumption 2.5 holds for both smooth first-order and projection-based retraction curves. For a smooth first-order retraction curve  $\gamma(t) = \text{Retr}_x(t\xi)$ , the first and second derivatives are given by  $\gamma'(t) = d\text{Retr}_x(t\xi)[\xi]$  and  $\gamma''(t) = d^2\text{Retr}_x(t\xi)[\xi, \xi]$ , respectively. Hence, the bounds on  $\|\gamma'(t)\|$  and  $\|\gamma''(t)\|$  depend on the operator norm of  $d\text{Retr}_x(t\xi)$  and  $d^2\text{Retr}_x(t\xi)$ . Let  $B_1$  and  $B_2$  denote the uniform bounds such that  $\|d\text{Retr}_x(\eta)\|_{op} \leq B_1$  and  $\|d^2\text{Retr}_x(\eta)\|_{op} \leq B_2$  for all  $\eta$  in a ball with a finite radius in the tangent space  $T_x \mathcal{M}$ . Then, we have

$$\left\| \frac{d}{dt} \text{Retr}_x(t\xi) \right\| \leq B_1 \|\xi\|, \quad \text{and} \quad \left\| \frac{d^2}{dt^2} \text{Retr}_x(t\xi) \right\| \leq B_2 \|\xi\|^2.$$

In the implementation, it suffices to rescale the gradient clipping parameter  $D$  by a factor of  $\frac{1}{B_1}$  to maintain the same convergence rate.

A similar argument applies to projection-based retraction curves, e.g.,  $\gamma(t) = \text{Proj}_{\mathcal{M}}(x + t\xi)$ , where the constants  $B_1$  and  $B_2$  are determined by the operator norms of  $d\text{Proj}_{\mathcal{M}}(x + t\xi)[\xi]$  and  $d^2\text{Proj}_{\mathcal{M}}(x + t\xi)[\xi, \xi]$ .

### B.2 Example on Manifold-Retraction Pairs

In this part, we prove that for polar decomposition retractions  $\text{Retr}_X(t\xi) = (X + t\xi)(I_p + t^2\xi^\top \xi)^{-\frac{1}{2}}$  on the Stiefel manifold  $St(p, n) = \{X \in \mathbb{R}^{n \times p} : X^\top X = I_p\}$ , the conditions of Assumption 2.5 hold with  $C' = 1$ .

For the notational convenience, we denote  $A_t := I_p + t^2\xi^\top \xi$ ,  $B_t := \frac{d}{dt} \text{Retr}_x(t\xi)$  and  $C_t := \frac{d^2}{dt^2} \text{Retr}_x(t\xi)$ . Also,  $\|\cdot\|_F$  denotes the Frobenius norm. Let us start with

$$\begin{aligned} B_t &= \frac{d}{dt} \text{Retr}_x(t\xi) = \xi(I_p + t^2\xi^\top \xi)^{-\frac{1}{2}} - t(X + t\xi)(I_p + t^2\xi^\top \xi)^{-\frac{3}{2}}(\xi^\top \xi) \\ &= \xi(A_t)^{-\frac{1}{2}} - t(X + t\xi)A_t^{-\frac{3}{2}}(\xi^\top \xi) \\ &= \xi(A_t)^{-\frac{3}{2}}(A_t - t^2\xi^\top \xi) - tX(A_t)^{-\frac{3}{2}}(\xi^\top \xi) \\ &= \xi(A_t)^{-\frac{3}{2}} - tX(A_t)^{-\frac{3}{2}}(\xi^\top \xi). \end{aligned}$$

Since  $\xi \in T_X \mathcal{M}$ , we have  $X^\top \xi + \xi^\top X = 0$ . Now, suppose that the eigenvalue decomposition of  $\xi^\top \xi$  can be written as  $\xi^\top \xi = U\Sigma U^\top$ , where  $\Sigma = \text{diag}(\lambda_i)$ ,  $\lambda_i \geq 0$ . We have that  $A_t = U\text{diag}(1 + t^2\lambda_i)U^\top$  and  $\text{Tr}(A_t^{-3}(\xi^\top \xi)) = \sum_{i=1}^p \frac{\lambda_i}{(1+t^2\lambda_i)^3}$ . Then,  $\text{Tr}(B_t^\top B_t)$  can be bounded as

$$\begin{aligned} \text{Tr}(B_t^\top B_t) &= \text{Tr}(A_t^{-3}(\xi^\top \xi)) + \text{Tr}(t^2 A_t^{-3}(\xi^\top \xi)^2) \\ &= \sum_{i=1}^p \frac{\lambda_i}{(1+t^2\lambda_i)^3} + \frac{t^2 \lambda_i^2}{(1+t^2\lambda_i)^3} = \sum_{i=1}^p \frac{\lambda_i}{(1+t^2\lambda_i)^2} \\ &\leq \sum_{i=1}^p \lambda_i = \text{Tr}(\xi^\top \xi). \end{aligned}$$

So, we showed that the first condition holds, i.e.,  $\left\| \frac{d}{dt} \text{Retr}_x(t\xi) \right\|_F \leq \|\xi\|_F$ .

Similarly, we can write  $C_t$  as

$$C_t = \frac{d^2}{dt^2} \text{Retr}_x(t\xi) = -3t\xi(A_t)^{-\frac{5}{2}}(\xi^\top \xi) - X(A_t)^{-\frac{3}{2}}(\xi^\top \xi) + 3t^2 X(A_t)^{-\frac{5}{2}}(\xi^\top \xi)^2.$$

Then, the upper bound on the trace of  $C_t^\top C_t$  can be derived as follows

$$\begin{aligned} \text{Tr}(C_t^\top C_t) &= \sum_{i=1}^p \frac{9t^2 \lambda_i^3}{(1+t^2 \lambda_i)^5} + \frac{\lambda_i^2}{(1+t^2 \lambda_i)^3} + \frac{9t^4 \lambda_i^4}{(1+t^2 \lambda_i)^5} - \frac{6t^2 \lambda_i^3}{(1+t^2 \lambda_i)^4} \\ &= \sum_{i=1}^p \frac{\lambda_i^2 + 5t^2 \lambda_i^3 + 4t^4 \lambda_i^4}{(1+t^2 \lambda_i)^5} = \sum_{i=1}^p \frac{\lambda_i^2(1+4t^2 \lambda_i^2)}{(1+t^2 \lambda_i)^4} \\ &\leq \sum_{i=1}^p \lambda_i^2 \leq \left(\sum_{i=1}^p \lambda_i\right)^2 = \text{Tr}(\xi^\top \xi)^2. \end{aligned}$$

As a result we have  $\left\| \frac{d^2}{dt^2} \text{Retr}_x(t\xi) \right\|_F = \|C_t\|_F = \sqrt{\text{Tr}(C_t^\top C_t)} \leq \text{Tr}(\xi^\top \xi) = \|\xi\|_F^2$ . Hence, the second condition of Assumption 2.5 holds with  $C' = 1$ .

## C Discussion on Intrinsic and Extrinsic Curvature Measures

In our work, we used second fundamental form to analyze distortions caused by parallel transport and projection-based operations. As an extrinsic measure of curvature, the second fundamental form is required to analyze projection-based algorithms. Since parallel transport is an intrinsic operation, an intrinsic measure of curvature can be used to analyze the corresponding distortions.

In the proof of Lemmas 2.9 and 4.1, the main argument is to bound  $\|v - \mathcal{P}_S^s(v)\|$  with the length of a loop  $S$  and a bound on the second fundamental form. A similar result can be obtained under the weaker assumption of bounded sectional curvature; however, this approach requires estimating the area enclosed by the loop, which introduces additional geometric complexity.

In the proof of optimization with a zeroth order gradient estimator, we used intrinsic curvature bounds  $\|R(X, Y)Z\| \leq K_c$  and  $|\sec(X, Y)| \leq K_s$  for unit vectors  $X, Y, Z \in \mathfrak{X}(\mathcal{M})$ . To relate these to extrinsic geometry, let the second fundamental form satisfy  $\|\Pi(X, Y)\| \leq C$ . The covariant (Riemann) curvature tensor acts as  $\text{Rm}(X, Y, Z, W) = \langle R(X, Y)Z, W \rangle$ . By [41, Theorem 8.5] this yields  $\text{Rm}(W, X, Y, Z) \leq 2C^2$ , providing an upper bound on sectional curvature in terms of the second fundamental form.

From the definition of sectional curvature we have

$$\sec(u, v) = \frac{\text{Rm}(u, v, v, u)}{\|u\|^2 \|v\|^2 - \langle u, v \rangle^2},$$

for linearly independent vectors  $u$  and  $v$ .

Let us assume  $u$  and  $v$  are unit vectors and write  $v = u \sin(\theta) + w \cos(\theta)$ , where  $w$  is orthogonal to  $u$ . Then, we have

$$\begin{aligned} \sec(u, v) &= \frac{\text{Rm}(u, u \sin(\theta) + w \cos(\theta), u \sin(\theta) + w \cos(\theta), u)}{\|u\|^2 \|v\|^2 - \langle u, v \rangle^2} \\ &= \frac{\text{Rm}(u, w \cos(\theta), w \cos(\theta), u)}{\cos^2(\theta)} \\ &= \text{Rm}(u, w, w, u) \leq 2C^2. \end{aligned}$$

In the above equation we use linearity of covariant curvature tensor in each entry and the symmetry property that implies  $\text{Rm}(X, X, Y, Z) = \text{Rm}(X, Y, Z, Z) = 0$ . We can also follow this alternative explanation. Since the sectional curvature depends only on the plane spanned by the vectors and is independent of the choice of basis, we can, without loss of generality, consider any orthonormal pair  $u, v$  and obtain the same result. Then,  $\sec(u, v) = \text{Rm}(u, v, v, u) \leq 2C^2$ . As a result, we can use  $2C^2$  in place of  $K_s$  as well as  $K_c$ .

## D Comparison of Zeroth Order Riemannian Gradient Estimators

Two-point gradient estimators in the form of  $g_\delta^E(x) = \frac{d}{2\delta}(F(x + \delta u, \nu) - F(x - \delta u, \nu))u$  are commonly used in nonsmooth nonconvex optimization, where  $u$  is sampled on a unit sphere [45, 37].

A natural extension of the gradient estimator  $g_\delta^E(x)$  to the Riemannian setting is follows,

$$g_\delta^R(x) = \frac{S_\delta}{V_\delta} f(u) \frac{\text{Exp}_x^{-1}(u)}{\|\text{Exp}_x^{-1}(u)\|}, \quad (25)$$

where  $S_\delta$  denotes the surface area of a sphere centered at  $x$  with radius  $\delta$ , and  $V_\delta$  denotes the volume of a ball centered at  $x$  with radius  $\delta$  [62]. The main advantage of this estimator is its unbiasedness [62, Lemma 11] when  $u$  is uniformly sampled from the sphere centered at  $x$  with radius  $\delta$ . However, the gradient estimator requires the computation of  $S_\delta$  and  $V_\delta$  for each point  $x$ , which makes this approach computationally demanding in practical applications.

An alternative for the zeroth order Riemannian gradient estimator can be written based on sampling  $u$  on the tangent space  $T_x\mathcal{M}$  as follows

$$g_\delta(x) = \frac{d}{2\delta}(F(\text{Exp}_x(\delta u), \nu) - F(\text{Exp}_x(-\delta u), \nu))u, \quad (26)$$

which is used in our work. As this estimator does not involve computing a volume or a surface area, it is practically advantageous; however, establishing a clear connection with the Riemannian gradient remains challenging. In [42] the connection between a gradient estimator involving  $f(\text{Retr}_x(u))$  and  $\text{grad}f(x)$  is established, assuming smoothness of the objective and restricted Lipschitz-type gradient for the pullback function. However, for nonsmooth objectives, it is essential to define a smoothed objective  $h_\delta(x)$  such that  $h_\delta(x)$  is Lipschitz continuous, and the distance between  $\text{grad}h_\delta(x)$  and  $\partial_\delta f(x)$  can be controlled with  $\delta$ .

## E Constant Terms

Here, we provide the constant terms used throughout the proofs.

$$C_1 = \theta + \sigma + G + GC' + 4GC$$

$$C_2 = C$$

$$C_3 = \theta + \sigma + G + 4mGC$$

$$C_4 = 2mGC + GC' + 2m^2GC^2$$

$$C_5 = 2^{1/3} \left( \theta + L \frac{\sinh(\sqrt{K_s}/2)}{\sqrt{K_s}/2} \sqrt{16\sqrt{2\pi}d(2 + C' + 4C)} \right)$$

$$C_6 = C(2 + \frac{3}{2}C_7) + \frac{1}{12}K_s$$

For  $C_7$ , find the definition in [47, Lemma 1].