Learning-Rate-Free Stochastic Optimization over Riemannian Manifolds

Daniel Dodd [1] **Louis Sharrock** [1] **Christopher Nemeth** [1]

Abstract

In recent years, interest in gradient-based optimization over Riemannian manifolds has surged. However, a significant challenge lies in the reliance on hyperparameters, especially the learning rate, which requires meticulous tuning by practitioners to ensure convergence at a suitable rate. In this work, we introduce innovative learning-rate-free algorithms for stochastic optimization over Riemannian manifolds, eliminating the need for hand-tuning and providing a more robust and user-friendly approach. We establish high probability convergence guarantees that are optimal, up to logarithmic factors, compared to the best-known optimally tuned rate in the deterministic setting. Our approach is validated through numerical experiments, demonstrating competitive performance against learning-rate-dependent algorithms.

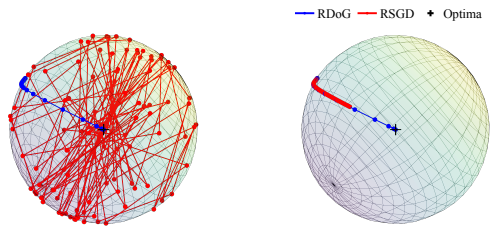
1. Introduction

We study Riemannian optimization problems of the form

$$\min_{x \in \mathcal{M}} f(x), \tag{1}$$

where f is a geodesically convex function, and $\mathcal{M}$ is a Riemannian manifold. In recent years, there has been a growing interest within the machine learning community in addressing optimization challenges on such geometric spaces. These problems manifest in diverse applications, including principal component analysis (Edelman et al., 1998), dictionary learning (Sun et al., 2017), low-rank matrix completion (Boumal & Absil, 2011), tensor factorization (Ishteva et al., 2011), Gaussian mixture models (Hosseini & Sra, 2015) and metric learning (Zadeh et al., 2016).

One of the prominent hurdles in applying Riemannian gradient-based optimization is the requirement for careful tuning of the learning rate or step size parameter. Selecting an appropriate learning rate is imperative to the algorithm's performance, impacting the convergence rate, final solution quality, and overall algorithm stability. To illustrate, Figure 1 showcases the impact of inadequate learning rates on the convergence rate of Riemannian stochastic gradient descent (RSGD, Bonnabel, 2013).



(a) Learning rate too large. (b) Learning rate too small.

Figure 1. Rayleigh quotient maximization on the unit sphere. Our algorithm, RDoG, converges without tuning, while RSGD shows sensitivity to the learning rate, leading to (a) overshooting or (b) slow convergence.

Recently, the expense and lack of robustness associated with learning rate tuning have spurred substantial research of *learning-rate-free* methods for Euclidean optimization. These aim to automate tuning by crafting algorithms that achieve near-optimal convergence rates with minimal knowledge of the function's properties and do not have any tuning parameters. Notable examples include online learning schemes like coin betting (Orabona & Pál, 2016) and exponentiated gradients (McMahan & Orabona, 2014) and bisection subroutines (Carmon & Hinder, 2022). Our paper addresses the absence of comparable tools for Riemannian optimization with the first comprehensive study of learning-rate-free algorithms in this setting.

Contributions: Building upon the recently proposed Distance over Gradients (DoG, Ivgi et al., 2023) and Distance over Weighted Gradients (DoWG, Khaled et al., 2023) Euclidean optimization approaches, we introduce dynamic learning-rate-scheduler algorithms for stochastic Riemannian optimization. Our results establish high probability convergence guarantees, achieving optimal convergence rates with logarithmic factors in smooth and Lipschitz settings, rendering them a robust solution for geodesically convex stochastic optimization on Riemannian manifolds.

[1]Department of Mathematics and Statistics, Lancaster University, UK. Correspondence to: Daniel Dodd <d.dodd1@lancaster.ac.uk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

2. Preliminaries

2.1. Riemannian Geometry

In this section, we recall some fundamental definitions from Riemannian geometry (e.g., Petersen, 2006; Lee, 2012; Boumal, 2023).

Riemannian manifold, tangent space, metric. A *Riemannian manifold* $\mathcal{M}$ is a smooth, locally Euclidean space. At each point x on $\mathcal{M}$, there is a corresponding *tangent space* $\mathcal{T}_x\mathcal{M}$ representing all possible tangential directions, endowed with a smoothly varying inner product $\langle\cdot,\cdot'\rangle_x : \mathcal{T}_x \times \mathcal{T}_x \to \mathbb{R}$ termed the *Riemannian metric*, that induces a norm $\|\cdot\|_x = \sqrt{\langle\cdot,\cdot\rangle_x}$. The metric measures angles, curve lengths, surface areas, and volumes locally, with global quantities obtained by integrating these contributions.

Geodesics and distances. The length of a curve $c : [0,1] \mapsto c(t) \in \mathcal{M}$ is $L(c) = \int_0^1 \|c'(t)\|_{c(t)} \mathrm{d}t$. Generalizing straight lines leads to *geodesics*, constant speed curves representing the shortest path between points x and y on the manifold: $\gamma = \arg\min_c L(c)$ with $\gamma(0) = x$, $\gamma(1) = y$, and $\|\gamma'(t)\|_{\gamma(t)} = 1$, establishing a metric space structure with *geodesic distance* $d(x,y) = \inf_c L(c)$.

Exponential maps. The concept of moving along a "straight" curve with constant velocity is given by the *exponential map*. As such, for any point x on $\mathcal{M}$, and any tangent vector $v \in \mathcal{T}_x\mathcal{M}$, there is a unique unit speed geodesic γ satisfying $\gamma(0) = x$ and $\gamma'(0) = v$. The corresponding exponential map $\exp_x : \mathcal{T}_x\mathcal{M} \to \mathcal{M}$ is defined as $\exp_x(v) = \gamma(1)$. When $\exp_x$ is well-defined on $\mathcal{T}_x\mathcal{M}$ for all $x \in \mathcal{M}$, the geodesic distance $d(x,y)$ is given by $\|\exp_x^{-1}(y)\|_x$.

Parallel transport. *Parallel transport* $\Gamma_x^y : \mathcal{T}_x\mathcal{M} \to \mathcal{T}_y\mathcal{M}$ provides a means to move tangent vectors from one tangent space to another while preserving their norm, and roughly speaking, "direction," analogous to translation in Euclidean space.

Curvature. The *curvature* of a Riemannian manifold is determined by its metric at each point. The *sectional curvature* at a point x on the manifold is the Gauss curvature of a two-dimensional submanifold formed as the image of a two-dimensional subspace of the tangent space $\mathcal{T}_x\mathcal{M}$ under the exponential map.

Trigonometric bound. The law of cosines in Euclidean space is fundamental for analyzing optimization algorithms,

$$a^2 = b^2 + c^2 - 2bc\cos(A),$$

where a, b, c are the sides of a Euclidean triangle with A the angle between sides b and c. Trigonometric geometry behaves differently in manifolds compared to Euclidean spaces. While the equality does not hold for nonlinear spaces, a trigonometric distance bound can be established for manifolds with sectional curvature bounded below.

Lemma 2.1. *(Zhang & Sra, 2016, Lemma 5) Suppose a, b, c are the side lengths of a geodesic triangle Δ in a Riemannian manifold with sectional curvature lower bounded by $\kappa > -\infty$ and A is the angle between sides b and c (defined through the inverse exponential map and inner product in tangent space). Then*

$$a^2 \le \zeta_\kappa(c) b^2 + c^2 - 2bc\cos(A),$$

where $\zeta_\kappa : \mathbb{R}_+ \to \mathbb{R}$ is the geometric curvature function

$$\zeta_\kappa(d) = \begin{cases} \frac{\sqrt{|\kappa|}\cdot d}{\tanh\left(\sqrt{|\kappa|}\cdot d\right)}, & \text{if } \kappa < 0, \\ 1, & \text{if } \kappa \ge 0. \end{cases}$$

Proof. Given by Lemma 3.12 of (Cordero-Erausquin et al., 2001) and by Lemma 5 of (Zhang & Sra, 2016). □

2.2. Function Classes

Geodesic convexity. $\mathcal{M}$ is *geodesically convex* if every two points are connected by a geodesic. A function $f : \mathcal{M} \to \mathbb{R}$ is *geodesically convex* if, for any geodesic $\gamma \subset \mathcal{M}$,

$$f(\gamma(t)) \le (1-t)f(\gamma(0)) + tf(\gamma(1)), \quad \forall t \in [0,1].$$

Equivalently, $\mathcal{M}$ is geodesically convex if, for any $x, y \in \mathcal{M}$, there exists a tangent vector $\partial f(x) \in \mathcal{T}_x\mathcal{M}$ such that

$$f(y) \ge f(x) + \langle \partial f(x), \exp_x^{-1}(y)\rangle_x,$$

where $\partial f(x)$ is a *Riemannian subgradient of f at x*. When f is differentiable, $\{\partial f(x)\} = \operatorname{grad} f(x)$, the *Riemannian gradient of f at x*, defined as the tangent vector in $\mathcal{T}_x\mathcal{M}$ satisfying

$$\langle \operatorname{grad} f(x), v\rangle_x = df(x)[v],$$

where $df(x) : \mathcal{T}_x\mathcal{M} \to \mathbb{R}$ denotes the differential of f at x.

Geodesic Lipschitz. A function $f : \mathcal{M} \to \mathbb{R}$ is said to be *geodesically L-Lipschitz* if, for all $x, y \in \mathcal{M}$, there exists a constant $L > 0$ such that,

$$|f(y) - f(x)| \le L \cdot \|\exp_x^{-1}(y)\|_x.$$

When f is differentiable, the geodesically L-Lipschitzness is equivalent to $\|\operatorname{grad} f(x)\|_x \le L$ for all $x \in \mathcal{M}$.

Geodesic smoothness. A differentiable function f is *geodesically S-smooth* if its gradient is geodesically S-Lipschitz. That is, if for all $x, y \in \mathcal{M}$,

$$\|\operatorname{grad} f(x) - \Gamma_y^x \operatorname{grad} f(y)\|_x \le S \cdot \|\exp_x^{-1}(y)\|_x,$$

where Γ_y^x is the parallel transport from y to x. One can show in this case that

$$f(y) \le f(x) + \langle \operatorname{grad} f(x), \exp_x^{-1}(y)\rangle_x + \frac{S}{2} \cdot \|\exp_x^{-1}(y)\|_x^2.$$

3. Algorithms and Theory

We are interested in solving optimization problems of the form (1). We proceed under the following standard regularity conditions (Zhang & Sra, 2016; Alimisis et al., 2020a).

Assumption 3.1 (Geodesic convexity)**.** The geodesically convex function $f\colon \mathcal{M} \to \mathbb{R}$ attains its minimum at $x_\star$ within its closed and geodesically convex domain $\mathcal{M}$, which includes a well-defined exponential map.

Assumption 3.2 (Lower bounded sectional curvature)**.** $\mathcal{M}$ exhibits sectional curvature bounded from below: $\kappa > -\infty$.

To minimize f, we will assume access to a *stochastic gradient oracle* $\mathcal{G}$. When queried at $x \in \mathcal{M}$, the oracle returns a stochastic (sub)gradient estimator $\mathcal{G}(x)$ which satisfies $\mathbb{E}\left[\mathcal{G}(x)|x\right] \in \partial f(x)$. In a slight abuse of notation, we will henceforth write $\operatorname{grad} f(x) \coloneqq \mathbb{E}\left[\mathcal{G}(x)|x\right]$. We consider the following additional assumptions.

Assumption 3.3 (Locally bounded stochastic gradients)**.** There exists some continuous function $\ell\colon \mathcal{M} \to \mathbb{R}_+$ such that $\|\mathcal{G}(x)\|_x \leq \ell(x)$ almost surely.

Assumption 3.4 (Locally smooth stochastic gradients)**.** There exists some continuous function $s\colon \mathcal{M} \to \mathbb{R}_+$ such that $\|\mathcal{G}(x) - \Gamma_y^x \mathcal{G}(y)\|_x \leq s(x)\|\exp_x^{-1}(y)\|_x$, almost surely.

Assumption 3.3 corresponds to the Riemannian analog of Ivgi et al. (2023)'s locally bounded gradient assumption, whereas Assumption 3.4 introduces a novel condition.

3.1. Riemannian Stochastic Gradient Descent

Our work centers on the Riemannian stochastic gradient descent algorithm (RSGD) introduced by Bonnabel (2013), which from an initial point $x_0 \in \mathcal{M}$ iterates the following update rule:

$$x_{t+1} = \exp_{x_t}(-\eta_t g_t).$$

Here $t \geq 0$ denotes the iteration index, $g_t \coloneqq \mathcal{G}(x_t)$ represents the stochastic gradient oracle, and $\eta_t > 0$ is a user-chosen learning rate or step size parameter.

Our analysis commences by characterizing the "ideal step size" in the deterministic gradient setting, an extension of Theorem 9 from Zhang & Sra (2016).

Theorem 3.5. *Under noiseless conditions and Assumption 3.1, and 3.2, RSGD with a constant step size $\eta_t = \eta > 0$, for a geodesically L-Lipschitz function, satisfies*

$$\min_{t \leq T}\left[f(x_t) - f(x_\star)\right] \leq \left[\frac{\bar{d}_T^2}{2\eta T} + \frac{\eta \zeta_\kappa(\bar{d}_T)\sum_{t=0}^T \|g_t\|_{x_t}^2}{2T}\right],$$

where $\bar{d}_t \coloneqq \max_{s \leq t} d_s, d_s = d(x_s, x_\star)$. Minimizing this bound with respect to η, gives a convergence rate of $O\left(L\bar{d}_T\sqrt{\frac{\zeta_\kappa(\bar{d}_T)}{T}}\right)$ with corresponding "ideal step size"

$$\eta_\star = \frac{\bar{d}_T}{\sqrt{\zeta_\kappa(\bar{d}_T)\sum_{t=0}^T \|g_t\|_{x_t}^2}}.$$

Proof. See Appendix B.1. □

3.2. Riemannian Distance Over Gradients

In practice, determining the "ideal step size" $\eta_\star$, even in hindsight, is challenging due to its dependence on the unknown maximum distance $\bar{d}_T$. In this section, we introduce an adaptive algorithm that estimates this whilst attaining the optimal convergence rate up to a logarithmic factor.

Learning-rate-free schedule for RSGD. Our key proposal, inspired by Ivgi et al. (2023), is to estimate $\bar{d}_T$ via a proxy,

$$\bar{r}_t \coloneqq \max_{s \leq t} r_s, \quad r_s \coloneqq \max(d(x_0, x_s), \epsilon),$$

where $\epsilon > 0$ is an initial estimate. Intuitively, the maximum deviation from the starting point should reflect the maximum deviation from the optimum, assuming the RSGD iterations converge to the optimum. Integrating this estimation into the "ideal step size," we establish an adaptive sequence of step sizes

$$\eta_t = \frac{\bar{r}_t}{\sqrt{\zeta_\kappa(\bar{r}_t)\sum_{s=0}^t \|g_s\|_{x_s}^2}}.$$

We term this step size schedule as *Riemannian Distance over Gradients* (RDoG, Algorithm 1). Observe that the initial step gives a step size of $\epsilon/\|g_0\|_{x_0}$, a normalized gradient step of size ϵ. We demonstrate that, provided ϵ is chosen sufficiently small, the specific value is insensitive.

Algorithm 1 RDoG

Input: initial point x_0, initial estimate $\epsilon > 0$, $G_{-1} = 0$.
for $t = 0$ **to** $T - 1$ **do**
  $g_t = \mathcal{G}(x_t)$
  $\bar{r}_t = \max\left(\epsilon, \max_{s \leq t} d(x_s, x_0)\right)$
  $G_t = G_{t-1} + ||g_t||_{x_t}^2$
  $\eta_t = \frac{\bar{r}_t}{\sqrt{\zeta_\kappa(\bar{r}_t) G_t}}$
  $x_{t+1} = \exp_{x_t}\left(-\eta_t g_t\right)$
end for

Optimality gap bounds assuming bounded iterates. We bound the error of the weighted average sequence

$$\tilde{x}_{t+1} = \exp_{\tilde{x}_t}\left(\frac{\bar{r}_t/\sqrt{\zeta_\kappa(\bar{r}_t)}}{\sum_{s=0}^t \bar{r}_s/\sqrt{\zeta_\kappa(\bar{r}_s)}} \exp_{\tilde{x}_t}^{-1}(x_t)\right), \ \tilde{x}_1 = x_0.$$

To simplify our analysis, we write $\log_+(\cdot) \coloneqq 1 + \log(\cdot)$ where the logarithm has a base of e, and introduce the following quantities

$$G_t \coloneqq \sum_{s=0}^{t} \|g_s\|_{x_s}^2, \ \theta_{t,\delta} \coloneqq \log\left(\frac{60\log(6t)}{\delta}\right).$$

Our first result establishes a bound on the optimality gap under bounded iterates.

Theorem 3.6. *Suppose that Assumption 3.1, 3.2, and 3.3 hold. Then, for all $\delta \in (0,1)$ and $L > 0$, and for all $t \leq T$, RDoG (Algorithm 1) satisfies the optimality gap $f(\tilde{x}_t) - f(x_\star)$ of*

$$O\left(\frac{(d_0 + \bar{r}_t)\sqrt{\zeta_\kappa(d_0 + \bar{r}_t)}\sqrt{G_{t-1} + \theta_{t,\delta}G_{t-1} + \theta_{t,\delta}^2 L^2}}{\sum_{s=0}^{t-1} \frac{\bar{r}_s/\sqrt{\zeta_\kappa(\bar{r}_s)}}{\bar{r}_t/\sqrt{\zeta_\kappa(\bar{r}_t)}}}\right),$$

with probability at least $1 - \delta - \mathbb{P}(\bar{\ell}_T > L)$, where $\bar{\ell}_T \coloneqq \max_{s\leq T} \ell(x_s)$.

Proof. See Appendix D.2. □

This theorem yields a corollary for bounded manifolds.

Corollary 3.7. *Under Assumption 3.1, 3.2, and 3.3, for any $D \geq d_0$, let $L_D \coloneqq \max_{x\in\mathcal{M}:d(x,x_0)\leq D} \ell(x)$. Then, for all $\delta \in (0,1)$ and for $\tau \in \arg\max_{t\leq T} \sum_{s=0}^{t-1} \frac{\bar{r}_s/\sqrt{\zeta_\kappa(\bar{r}_s)}}{\bar{r}_t/\sqrt{\zeta_\kappa(\bar{r}_t)}}$, RDoG (Algorithm 1) satisfies the optimality gap $f(\tilde{x}_\tau) - f(x_\star)$ of*

$$O\left(\frac{D\sqrt{\zeta_\kappa(D)}\sqrt{G_{\tau-1}\theta_{\tau,\delta} + L_D^2\theta_{\tau,\delta}^2}}{T}\log_+\left(\frac{D\sqrt{\zeta_\kappa(\epsilon)}}{\epsilon\sqrt{\zeta_\kappa(D)}}\right)\right),$$

with probability at least $1 - \delta - \mathbb{P}(\bar{\ell}_T > L)$.

Proof. See Appendix D.2. □

Unlike prior work on bounded domains (e.g., Zhang & Sra, 2016; Wang et al., 2021), our approach adapts without knowledge of the domain width to set the learning rate, achieving optimality up to a logarithmic factor.

Remark 3.8. We enhance this result to a high probability convergence guarantee of $O(1/T)$ under uniformly averaged iterates, following Assumption 3.4 in Appendix D.3.

Remark 3.9. In Appendix D.4, we ensure bounded iterates with high probability by slightly reducing step sizes.

Remark 3.10. Omitting the geometric curvature term $\zeta_\kappa(\cdot)$ from RDoG's step sizes and weighted averaging results in an additional cost of $O(\sqrt{\zeta_\kappa(D)})$ in the optimality gap. Further details are available in Appendix D.5.

3.3. Normalized Riemannian Stochastic Gradient Descent

We consider extending standard Euclidean normalized gradient descent (Shor, 2012; Levy, 2016; Konnov, 2003; Hazan et al., 2015) to Riemannian manifolds, providing scale-free adaptability, with updates of the form

$$x_{t+1} = \exp_{x_t}\left(-\eta_t \frac{\operatorname{grad} f(x_t)}{\|\operatorname{grad} f(x_t)\|_{x_t}}\right), \quad x_0 \in \mathcal{M}.$$

We term this algorithm *Normalized Riemannian Stochastic Gradient Descent* (NRSGD). In the deterministic Euclidean setting, normalized gradient descent automatically adjusts to the Lipschitz constant in non-smooth optimization (Nesterov, 2018, Theorem 3.2.2) and the smoothness constant(s) in smooth optimization (Grimmer, 2019, Corollary 2.2). This adaptability extends to NRSGD, as we will demonstrate.

Theorem 3.11. *Under noiseless conditions and Assumption 3.1, and 3.2, NRSGD with a constant step size $\eta_t = \eta > 0$, for a geodesically L-Lipschitz function, satisfies*

$$\min_{t\leq T}\left[f(x_t) - f(x_\star)\right] \leq L\left[\frac{\bar{d}_T^2}{2\eta T} + \frac{\eta}{2}\zeta_\kappa(\bar{d}_T)\right].$$

While for a geodesically S-smooth function, we have

$$\min_{t\leq T}\left[f(x_t) - f(x_\star)\right] \leq 2S\left[\frac{\bar{d}_T^2}{2\eta T} + \frac{\eta}{2}\zeta_\kappa(\bar{d}_T)\right]^2.$$

Minimizing these give respective convergence rates $O\left(L\bar{d}_T\sqrt{\frac{\zeta_\kappa(\bar{d}_T)}{T}}\right)$ and $O\left(\frac{2S\bar{d}_T^2\zeta_\kappa(\bar{d}_T)}{T}\right)$ with corresponding "ideal step size"

$$\eta_\star = \frac{\bar{d}_T}{\sqrt{T\zeta_\kappa(\bar{d}_T)}}.$$

Proof. See Appendix C.1. □

Learning-rate-free schedule for NRSGD. Normalization brings adaptivity to Lipschitz and smoothness settings, using a common *universal* "ideal step size". However, like RSGD, this "ideal step size" relies on the intractable maximum distance quantity $\bar{d}_T$. Our solution is to substitute this with our proxy $\bar{r}_t$, resulting in our second algorithm, *Normalized Riemannian Distance over Gradients* (NRDoG) algorithm, summarized in Algorithm 4 in Appendix F.

3.4. Riemannian Distance Over Weighted Gradients

Weighted learning-rate-free schedule for RSGD. We introduce a third algorithm *Riemannian Distance over Weighted Gradients* (RDoWG, Algorithm 2) that extends the recently proposed Distance over Weighted Gradients

(DoWG) (Khaled et al., 2023) to the Riemannian setting. Like RDoG and NRDoG, RDoWG estimates the intractable maximum distance quantity $\bar{d}_T$ by utilizing the maximum distance deviation from the initial point, $\bar{r}_t$. However, in RDoWG, the normalization is based on the square root of the *weighted* gradient sum, $v_t = \sum_{s=0}^{t} \bar{r}_s^2 ||g_s||_{x_s}^2$, rather than simply the square root of the gradient sum $G_t = \sum_{s=0}^{t} ||g_s||_{x_s}^2$.

The motivation for this normalization choice, as discussed in Khaled et al. (2023), lies in its improved adaptation to the problem geometry, especially in regions far from the initialization at x_0. Specifically, as the distances $\{\bar{r}_t\}_{t\geq 0}$ monotonically increase, later gradients receive greater weights than earlier gradients. This choice aligns with the practice in previous Riemannian optimization schemes such as RADAM (Becigneul & Ganea, 2019), which also utilizes weighted gradient sums. However, unlike RADAM, where weights are determined by fixed user-selected hyperparameters, RDoWG adaptively estimates these weights.

Algorithm 2 RDoWG

Input: initial point x_0, initial estimate $\epsilon > 0$, $v_{-1} = 0$.
for $t = 0$ **to** $T-1$ **do**
  $g_t = \mathcal{G}(x_t)$
  $\bar{r}_t = \max\left(\epsilon, \max_{s\leq t} d(x_s, x_0)\right)$
  $v_t = v_{t-1} + \bar{r}_t^2 \|g_t\|_{x_t}^2$
  $\eta_t = \frac{\bar{r}_t}{\sqrt{\zeta_\kappa(\bar{r}_t) v_t}}$
  $x_{t+1} = \exp_{x_t}(-\eta_t g_t)$
end for

Optimality gap bounds assuming bounded iterates. We bound the error of the weighted average sequence

$$\tilde{x}_{t+1} = \exp_{\tilde{x}_t}\left(\frac{\bar{r}_t^2/\zeta_\kappa(\bar{r}_t)}{\sum_{s=0}^{t} \bar{r}_s^2/\zeta_\kappa(\bar{r}_s)} \exp_{\tilde{x}_t}^{-1}(x_t)\right),\ \tilde{x}_1 = x_0.$$

We initiate our analysis in the non-smooth setting before transitioning to the smooth setting. Our initial result, assuming bounded iterates, provides the optimality gap achieved by RDoWG.

Theorem 3.12. *Suppose that Assumption 3.1, 3.2, and 3.3 hold. Then, for all $\delta \in (0,1)$ and $L > 0$, and for all $t \leq T$, RDoWG (Algorithm 2) satisfies the optimality gap $f(\tilde{x}_t) - f(x_\star)$ of*

$$O\left(\frac{(d_0 + \bar{r}_t)\sqrt{\zeta_\kappa(d_0+\bar{r}_t)}\sqrt{G_{t-1} + \theta_{t,\delta} G_{t-1} + \theta_{t,\delta}^2 L^2}}{\sum_{s=0}^{t-1} \frac{\bar{r}_s^2/\zeta_\kappa(\bar{r}_s)}{\bar{r}_t^2/\zeta_\kappa(\bar{r}_t)}}\right),$$

with probability at least $1 - \delta - \mathbb{P}(\bar{\ell}_T > L)$.

Proof. See Appendix E.3 □

We obtain a result on bounded domains which is optimal up to a logarithmic factor.

Corollary 3.13. *Suppose that Assumption 3.1, 3.2, and 3.3 hold. In addition, for any $D \geq d_0$, let $L_D := \max_{x\in\mathcal{M}: d(x,x_0)\leq D} \ell(x)$. Then, for all $\delta \in (0,1)$ and for $\tau \in \arg\max_{t\leq T} \sum_{s=0}^{t-1} \frac{\bar{r}_s^2/\zeta_\kappa(\bar{r}_s)}{\bar{r}_t^2/\zeta_\kappa(\bar{r}_t)}$, RDoWG (Algorithm 2) satisfies the optimality gap $f(\tilde{x}_\tau) - f(x_\star)$ of*

$$O\left(\frac{D\sqrt{\zeta_\kappa(D)}\sqrt{G_{\tau-1}\theta_{\tau,\delta} + L_D^2\theta_{\tau,\delta}^2}}{T} \log_+\left(\frac{D\sqrt{\zeta_\kappa(\epsilon)}}{\epsilon\sqrt{\zeta_\kappa(D)}}\right)\right),$$

with probability at least $1 - \delta - \mathbb{P}(\bar{\ell}_T > L)$.

Proof. See Appendix E.3 □

We proceed with analyzing the smooth setting. Our initial result yields an optimality gap for bounded iterates. It is worth noting that RDoG achieves similar results via uniform averaging, albeit with an additional cost (see Appendix D.3 for further details).

Theorem 3.14. *Suppose that Assumption 3.1, 3.2, and 3.4 hold and write $\bar{s}_T := \max_{t\leq T} s(x_t)$. Then, for all $\delta \in (0,1)$ and $S > 0$, and for all $t \leq T$, RDoWG (Algorithm 2) satisfies the optimality gap $f(\tilde{x}_t) - f(x_\star)$ of*

$$O\left(\frac{(d_0+\bar{r}_t)^2\zeta_\kappa(d_0+\bar{r}_t)(S\theta_{t,\delta}^2)}{\sum_{s=0}^{t-1}\frac{\bar{r}_s^2/\zeta_\kappa(\bar{r}_s)}{\bar{r}_t^2/\zeta_\kappa(\bar{r}_t)}}\right),$$

with probability at least $1 - \delta - \mathbb{P}(\bar{s}_T > S)$.

Proof. See Appendix E.4 □

This result achieves the optimal rate, aligning with the smooth analysis of Zhang & Sra (2016), with an additional logarithmic factor on bounded domains.

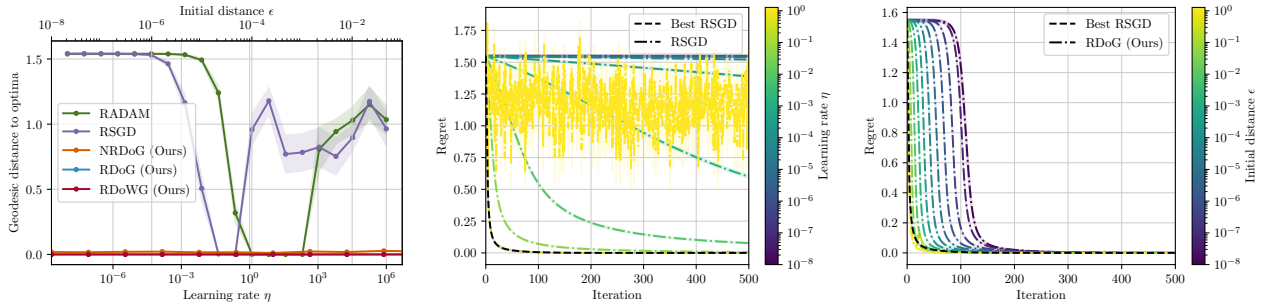
Corollary 3.15. *Suppose that Assumption 3.1, 3.2, and 3.4 hold. In addition, for any $D \geq d_0$, let $S_D := \max_{x\in\mathcal{M}: d(x,x_0)\leq D} s(x)$. Then, for all $\delta \in (0,1)$ and for $\tau \in \arg\max_{t\leq T} \sum_{s=0}^{t-1} \frac{\bar{r}_s^2/\zeta_\kappa(\bar{r}_s)}{\bar{r}_t^2/\zeta_\kappa(\bar{r}_t)}$, RDoWG (Algorithm 2) satisfies the optimality gap $f(\tilde{x}_\tau) - f(x_\star)$ of*

$$O\left(\frac{D^2\zeta_\kappa(D)S_D\theta_{\tau,\delta}^2}{T}\log_+\left(\frac{D\sqrt{\zeta_\kappa(\epsilon)}}{\epsilon\sqrt{\zeta_\kappa(D)}}\right)\right),$$

with probability at least $1 - \delta - \mathbb{P}(\bar{s}_T > S)$.

Proof. See Appendix E.4 □

Stability analysis. While RDoWG is generally stable in practice, in theory, the algorithm trajectories can diverge. Drawing inspiration from Ivgi et al. (2023), we now introduce a variant of RDoWG that guarantees iterates remain bounded with high probability. The concept involves using step sizes that are smaller by a polylogarithmic factor. Following the taxonomy introduced in Ivgi et al. (2023), we



(a) Initial distance sensitivity. (b) Regret trace plot RSGD. (c) Regret trace plot RDoG.

Figure 2. **Results for Rayleigh quotient maximization on the sphere.** (a) Geodesic distance between the final iterate and the numerical solution after $T = 5000$ iterations as a function of the learning rate for RADAM and RSGD and as a function of the initial distance estimate for RDoG, RDoWG, and NRDoG. (b) Shows the regret (the function value of each iterate minus the function value of the numerical solution) for RSGD for a selection of learning rates. (c) Shows the regret for RDoG for a selection of different initial distance estimates. Results are averaged over ten replications.

Algorithm 3 T-RDoWG

Input: initial point x_0, initial estimate $\epsilon > 0$, $v_{-1} = 0$.
for $t = 0$ **to** $T - 1$ **do**
$g_t = \mathcal{G}(x_t)$
$\bar{r}_t = \max\left(\epsilon, \max_{s \le t} d(x_s, x_0)\right)$
$v_t = v_{t-1} + \bar{r}_t^2 \|g_t\|_{x_t}^2$
$v_t' = 8^4 \theta_{T,\delta}^2 \log_+^2\left(\frac{(1+t)\bar{r}_t^2 \bar{\ell}_t^2/\zeta_\kappa(\bar{r}_t)}{\bar{r}_0^2 \bar{\ell}_0^2/\zeta_\kappa(\bar{r}_0)}\right)(v_{t-1} + 16\frac{\bar{r}_t^2}{\zeta_\kappa(\bar{r}_t)}\bar{\ell}_t^2)$
$\eta_t = \frac{\bar{r}_t}{\sqrt{\zeta_\kappa(\bar{r}_t) v_t'}}$
$x_{t+1} = \exp_{x_t}(-\eta_t g_t)$
end for

refer to this scheme as *Tamed Riemannian Distance over Weighted Gradients* (T-RDoWG, Algorithm 3).

Our first result characterizes the key property of T-RDoWG: bounded iterates with high probability.

Theorem 3.16. *Suppose that Assumption 3.1, 3.2, and 3.3 hold, and $\epsilon \le 3d_0$. Then, for any $\delta \in (0,1)$, and for any $t \in \mathbb{N}$, the iterations of T-RDoWG (Algorithm 3) satisfy $\mathbb{P}(\bar{r}_t > 3d_0) \le \delta$.*

Proof. See Appendix E.5 □

Using this result, we can now obtain the convergence rate of T-RDoWG.

Corollary 3.17. *Suppose that Assumption 3.1, 3.2, and 3.3 hold, and $\epsilon \le 3d_0$. For any $\delta \in (0, 1/2)$, and for any $t \in \mathbb{N}$, let $\tau \in \arg\max_{t \le T} \sum_{s=0}^{t-1} \frac{\bar{r}_s^2/\zeta_\kappa(\bar{r}_s)}{\bar{r}_t^2/\zeta_\kappa(\bar{r}_t)}$. Then T-RDoWG (Algorithm 3) satisfies the optimality gap $f(\tilde{x}_\tau) - f(x_\star)$ of*

$$O\left(c\frac{d_0\sqrt{\zeta_\kappa(d_0)(G_\tau + L_\star^2)}}{T}\right) = O\left(c\frac{d_0\sqrt{\zeta_\kappa(d_0)}L_\star}{\sqrt{T}}\right),$$

with probability at least $1 - 2\delta$, where $L_\star := \max_{x \in \mathcal{M}: d(x,x_0) \le 3d(x_\star,x_0)} \ell(x)$ and $c = \log_+(T\frac{d_0 L_\star}{f(x_0) - f(x_\star)}) \log_+(\frac{d_0}{\epsilon}) \log(\frac{\log_+(T)}{\delta})$.

Proof. See Appendix E.5 □

Remark 3.18. We can also extend the analysis to obtain a similar optimality gap in the smooth setting. For brevity, we omit the details here.

4. Related Work

Riemannian optimization. Numerous authors have studied optimization on Riemannian manifolds. Earlier works on this topic established the asymptotic convergence of first-order methods in both the deterministic (Udrişte, 1994; Absil et al., 2008) and the stochastic (Liu et al., 2004; Bonnabel, 2013) settings. More recently, Zhang & Sra (2016) obtained the first non-asymptotic analysis for Riemannian stochastic gradient descent, assuming geodesic convexity. Subsequently, other authors have obtained iteration complexity results for Riemannian proximal-point methods (Bento et al., 2017), Frank-Wolfe schemes (Weber & Sra, 2021), variance reduced methods (Zhang et al., 2016; Kasai et al., 2017; Sato et al., 2019; Zhou et al., 2021), trust-region methods (Boumal et al., 2018; Agarwal et al., 2021), amongst others. In parallel, there has also been growing interest in obtaining Riemannian counterparts of accelerated (Liu et al., 2017; Alimisis et al., 2020b; Zhang & Sra, 2018; Ahn & Sra, 2020) and adaptive (Becigneul & Ganea, 2019; Kasai et al., 2019; Cho & Lee, 2017; Roy et al., 2018) methods used in Euclidean optimization. No existing works, however, consider learning-rate-free Riemannian optimization algorithms.

Learning-rate-free Euclidean optimization. On the other hand, learning-rate-free methods for (stochastic) optimization on Euclidean spaces are substantial; see, e.g., Orabona & Cutkosky (2020); Carmon & Hinder (2022) and references therein. Most relevant to our work, Carmon & Hinder (2022) recently introduced a learning-rate-free algorithm for stochastic convex optimization based on interval bisection. Building on this work, Ivgi et al. (2023), Defazio &

Mishchenko (2023) and Khaled et al. (2023) have since obtained learning-rate-free (stochastic) convex optimization algorithms which, under varying assumptions, achieve the optimal convergence rate of (stochastic) gradient descent up to a logarithmic factor. Many other learning-rate-free optimization algorithms originate in the online learning literature. These include methods based on coin betting (Orabona & Pál, 2016; Orabona & Tommasi, 2017), exponentiated gradients (Streeter & McMahan, 2012; Orabona, 2013), amongst others (e.g., McMahan & Orabona, 2014; Orabona & Cutkosky, 2020). Recently, coin betting ideas have demonstrated effectiveness on Wasserstein spaces (Sharrock & Nemeth, 2023; Sharrock et al., 2023; 2024), that heuristically follow a Riemannian interpretation (Villani, 2003).

5. Experiments

In this section, we assess the numerical performance of RDoG (Algorithm 1), RDoWG (Algorithm 2), and NRDoG (Algorithm 4) against manually tuned RSGD (Bonnabel, 2013) and RADAM (Becigneul & Ganea, 2019). Implementing all algorithms in Python 3 with JAX (Bradbury et al., 2018), our experiments run on a MacBook Pro 16" (2021) with an Apple M1 Pro chip and 16GB of RAM. Detailed manifold descriptions and required operations for the experiments are provided in Appendix G. Code to reproduce the experiments is available at `https://github.com/daniel-dodd/riemannian_dog`.

5.1. Rayleigh Quotient Maximization on the Sphere

We seek to find the dominant eigenvector of a symmetric matrix A in $\mathbb{R}^{d\times d}$ by minimizing $-\frac{1}{2}x^TAx$ on the unit sphere $\mathbb{S}^{d-1}$. This is challenging for high-dimensional and ill-conditioned A in the Euclidean case. We consider $A = \frac{1}{d}BB^T$, with $B \in \mathbb{R}^{d\times q}$ having standard Gaussian entries.

For illustration purposes, we first consider $d = 3$ and $q = 5$ in Figure 1, underscoring the pivotal role of selecting an optimal learning rate for RSGD, as deviations, whether too small or too large, adversely affect performance.

In a higher-dimensional scenario with $d = 1000$ and $q = 1100 \cong d$, resulting in a high condition number, we employ RADAM and RSGD with a grid of twenty logarithmically spaced learning rates $\eta \in [10^{-8}, 10^{6}]$. On the other hand, we investigate RDoG and RDoWG with ten logarithmically spaced initial distance values $\epsilon \in [10^{-8}, 10^{0}]$. Here, we initialize ten starting points $x_0 \in \mathbb{R}^d$ by drawing their entries independently from a standard Gaussian distribution, then projecting them onto the sphere through normalizing, a shared procedure for each optimizer.

Our results show that RDoG, RDoWG, and NRDoG are insensitive to initial distance, consistently achieving robust performance in recovering negligible geodesic distance to the numerical solution via the eigendecomposition. In contrast, the effectiveness of RADAM and RSGD depends on selecting an appropriate learning rate. Notably, as seen in Figure 2, RSGD is highly sensitive to the learning rate, while RDoG rapidly adapts to optimal regret within a few hundred iterations, irrespective of the initial distance estimate's magnitude. Additional regret trace plots for other optimizers are available in Appendix H.1, along with similar plots for geodesic distance to the optima. These underscore that the algorithms quickly adapt within a few hundred iterations without prior knowledge of the function.

5.2. PCA on the Grassmann Manifold

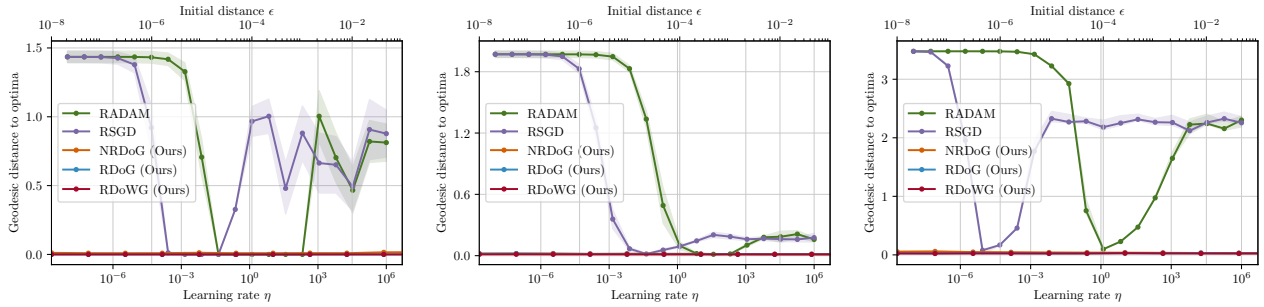
We investigate principal component analysis (PCA) on the Grassmann manifold $\mathbb{G}(d, r)$, where points are represented as equivalence classes with an orthogonal matrix $x \in \mathbb{R}^{d\times r}$ having orthonormal columns ($x^Tx = \mathrm{I}$). The PCA problem minimizes the sum of squared residual errors between projected data points and the original data, $\min_{x\in\mathbb{G}(d,r)} \frac{1}{n}\sum_{i=1}^{n}\|z_i - xx^Tz_i\|_2^2$, with each z_i represented as a d-dimensional data point. We consider datasets `Wine`, `Waveform-5000`, and `Tiny ImageNet`. The numerical solution is computed using the scikit-learn implementation (Pedregosa et al., 2011). The geodesic distances of final iterates (using weighted averages for RDoG and RDoWG) are compared against learning-rate-dependent algorithms, as shown in Figure 3.

In training, `Wine` uses the full batch for $T = 5000$ iterations, and `Waveform-5000` and `Tiny ImageNet` use batch sizes of 64 for $T = 2000$ iterations. Each dataset has an 80:20 train-test split per replication. Following Pymanopt (Townsend et al., 2016), initial points $x_0 \in \mathbb{R}^{d\times r}$ are drawn from a standard Gaussian distribution and projected onto the manifold using vectorized QR decomposition.

Results in Figure 3 from five random train-test splits show RDoG, RDoWG, and NRDoG are insensitive to initial distance estimates across magnitudes, with ten logarithmically spaced values in $\epsilon \in [10^{-8}, 10^{0}]$. In contrast, RADAM and RSGD require a narrower tuning range of optimal learning rates, exploring twenty logarithmically spaced values in $\eta \in [10^{-8}, 10^{6}]$. Additional results in Appendix H.2 further highlight the robust adaptation of RDoG, RDoWG, and NRDoG.

5.3. Embedding Graphs in the Poincaré Ball

The WordNet noun hierarchy (Miller et al., 1990) is a lexical database of English words organized into a hierarchical structure, where each word is categorized based on its semantic relationships with other words. Moreover, the *hypernymy relation*, often termed *Is-A relation*, signifies that one concept (the hypernym) encompasses another (the



(a) Wine ($d = 13, r = 1$). (b) Waveform ($d = 40, r = 2$). (c) Tiny ImageNet ($d = 12,288, r = 5$).

Figure 3. **Results for PCA on the Grassmann manifold**. (a)-(c) Geodesic distance between the final iterate and the numerical solution after $T = 2000$ iterations as a function of the learning rate for RADAM and RSGD and as a function of the initial distance estimate for RDoG, RDoWG, and NRDoG. (b)-(c) Uses the final iterate of the *weighted average* sequence for RDoG, RDoWG, and NRDog. Results are averaged over five replications.

hyponym). For instance, `mammal` is a hypernym of `dog` and `cat`. Following Nickel & Kiela (2017), we consider representing the transitive closure of the mammals' subtree that involves 1,180 nouns denoted as $\mathcal{N}$ (of which `mammal` is a hypernym) and 6,450 hypernymy relations, represented as $\mathcal{R} = \{(u, v)\} \subset \mathcal{N} \times \mathcal{N}$.

The embedding is performed in the Poincaré ball of hyperbolic geometry which is well-known to be better suited to embed tree-like graphs than the Euclidean space (Gromov, 1987; Sala et al., 2018). As such, the Poincaré ball model is defined as $\mathbb{B}_d = \{x \in \mathbb{R}^d : \|x\| < 1\}$ equipped with the Riemannian metric $\langle \cdot, \cdot' \rangle_x = 4/(1 - \|x\|^2)^2 \langle \cdot, \cdot' \rangle$. We adopt the loss function from the official code of Nickel & Kiela (2017), deviating from the one described in the paper:

$$\min_{\theta \colon \mathcal{N} \to \mathbb{B}_d} \sum_{(u,v) \in \mathcal{R}} -\log \left(\frac{e^{-d(\theta(u), \theta(v))}}{\sum_{v' \in \mathrm{Neg}(u,v)} e^{-d(\theta(u), \theta(v'))}} \right),$$

where each noun pair $(u, v) \in \mathcal{R}$ has associated embeddings $\theta(u), \theta(v) \in \mathbb{B}_d$, and $\mathrm{Neg}(u, v) = \{v' : (u, v') \notin \mathcal{R}\} \cup \{v\}$ is the set of negative examples for u, including v, and

$$d(\cdot, \cdot') = \operatorname{arcosh} \left(1 + 2 \frac{\|\cdot - \cdot'\|^2}{(1 - \|\cdot\|^2)(1 - \|\cdot'\|^2)} \right),$$

is the geodesic distance measuring the dissimilarity between the embeddings of two nouns in the Poincaré ball. Intuitively, minimizing this loss function encourages closely related mammals to be positioned closer together in the embedding space and less similar pairs to be farther apart.

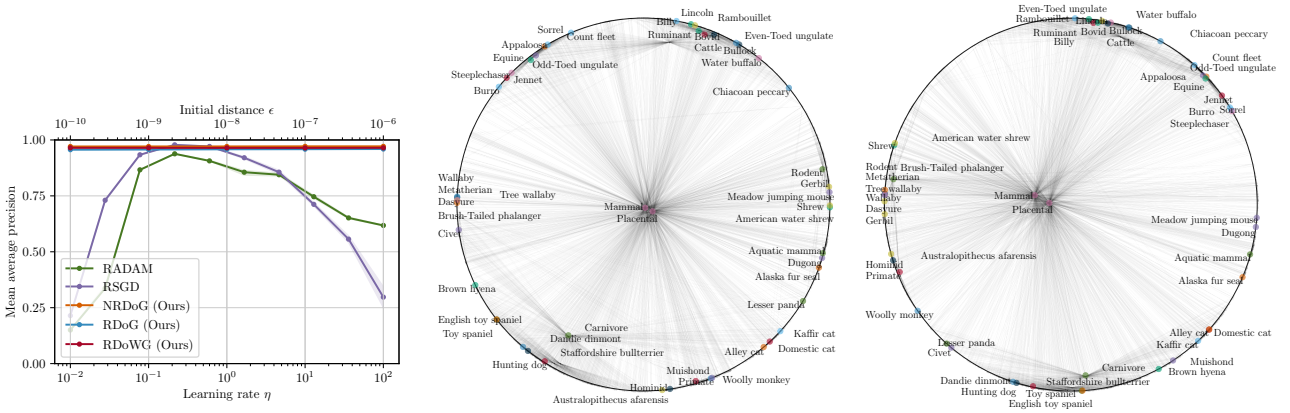
For initialization, following Nickel & Kiela (2017), we uniformly initialize the embeddings in $[-10^{-3}, 10^{-3}]^d$ and consider ten logarithmically spaced learning rates $\eta \in [10^{-2}, 10^2]$ and five logarithmically spaced initial distance estimates $\epsilon \in [10^{-10}, 10^{-6}]$. In the first ten epochs, we use RSGD with a reduced learning rate of $\eta/10$ for RSGD and RADAM. During this *burn-in* phase, negative word sampling is based on the graph degree raised to the power of $3/4$, leading to numerical improvements. No burn-in heuristic is applied for RDoG, RDoWG, or NRDoG. Thereafter, we run the optimizers on the initialized embeddings for one thousand epochs, with each iteration having a batch size of ten and fifty uniformly sampled negative samples. We repeat this experiment over five replications.

To measure the quality of the embeddings obtained from each optimizer, we follow Nickel & Kiela (2017) and compute, for each observed edge $(u, v) \in \mathcal{R}$, the corresponding distance $d(u, v)$ in the embedding space and rank it among the distances of all unobserved edges for u, i.e., $\{d(u, v') : (u, v') \notin \mathcal{R}\}$. Subsequently, we calculate the mean average precision of this ranking.

In Figure 4, embeddings of dimension five are presented. RDoG and RDoWG demonstrate competitive performance, while RADAM and RSGD require careful tuning. The performance significantly degrades for RADAM and RSGD without a burn-in heuristic, as exemplified in Appendix H.3. Visualizing two-dimensional embeddings between RDoG and RSGD trained for two thousand epochs, with burn-in applied only for RSGD and using the optimal learning rate selected from ten logarithmically spaced values $\eta \in [10^{-2}, 10^2]$, we observe meaningful groupings across various categories without employing burn-in heuristics for RDoG. Additional embedding plots for the other optimizers are presented in Appendix H.3.

6. Discussion

We have introduced new learning-rate-free optimizers for Riemannian manifolds and have highlighted significant numerical improvements over learning-rate-dependent algorithms. Our theoretical results provide high probability convergence guarantees that are optimal, up to a logarithmic factor, compared to the theoretically, yet practically unavailable, optimal deterministic algorithms.



(a) Mean average precision. (b) RDoG embeddings. (c) RSGD embeddings.

Figure 4. **Results for Poincaré word embeddings**. (a) The mean average precision of the embeddings is assessed against the ground truth after 1000 training epochs. Results are averaged over five replications, with the embedding dimension set to five. (b)-(c) Two-dimensional embeddings after 2000 training epochs are visualized and annotated for the first 50 nouns of the mammal's subtree for RDoG and RSGD.

Many existing Riemannian optimization methods rely on a retraction map, which serves as a cost-effective approximation of the exponential map on manifolds and is a reasonable choice in numerous real-world scenarios. Incorporating this into our framework is paramount for enhancing the effectiveness of our algorithms, particularly in large-scale optimization problems. Moreover, certain Riemannian manifolds, such as the Stiefel or multivariate Gaussian Fisher-Rao manifolds, pose challenges due to the intractability of the geodesic distance. Recognizing the argument underlying our convergence guarantees (though potentially less robust) holds for upper bounds on geodesic distance, exploring tractable or more economical approximations in these situations is essential.

Additionally, it is crucial to explore integrating these methods with recent proven advances in momentum acceleration (Liu et al., 2017; Alimisis et al., 2020b; Zhang & Sra, 2018; Ahn & Sra, 2020) — a challenge both in theory and practice. Furthermore, developing practical algorithms that offer guarantees on iterate boundedness is a consideration for future research.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback. DD was supported by the EPSRC-funded STOR-i Centre for Doctoral Training, grant number EP/S022252/1. LS and CN were supported by the Engineering and Physical Sciences Research Council (EPSRC), grant number EP/V022636/1. CN acknowledges further support from the EPSRC, grant numbers EP/S00159X/1 and EP/Y028783/1.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

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

Agarwal, N., Boumal, N., Bullins, B., and Cartis, C. Adaptive regularization with cubics on manifolds. *Mathematical Programming*, 188(1):85–134, 2021.

Ahn, K. and Sra, S. From Nesterov's estimate sequence to Riemannian acceleration. In *Proceedings of the 33rd Conference on Learning Theory (COLT 2020)*, 2020.

Alimisis, F., Orvieto, A., Becigneul, G., and Lucchi, A. A continuous-time perspective for modeling acceleration in Riemannian optimization. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1297–1307. PMLR, 2020a.

Alimisis, F., Orvieto, A., Becigneul, G., and Lucchi, A. A continuous-time perspective for modeling acceleration in Riemannian optimization. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, 2020b.

Becigneul, G. and Ganea, O.-E. Riemannian adaptive optimization methods. In *International Conference on Learning Representations*, 2019.

Bento, G. C., Ferreira, O. P., and Melo, J. G. Iteration-complexity of gradient, subgradient and proximal point methods on Riemannian manifolds. *Journal of Optimization Theory and Applications*, 173(2):548–562, 2017.

Bonnabel, S. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58 (9):2217–2229, 2013.

Boumal, N. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023.

Boumal, N. and Absil, P.-A. RTRMC: A Riemannian trust-region method for low-rank matrix completion. In *Proceedings of the 24th Conference on Neural Information Processing Systems (NIPS 2011)*, volume 24, 2011.

Boumal, N., Absil, P.-A., and Cartis, C. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2018.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. JAX: composable transformations of Python+NumPy programs, 2018.

Carmon, Y. and Hinder, O. Making SGD parameter-free. In *Conference on Learning Theory*, pp. 2360–2389. PMLR, 2022.

Cho, M. and Lee, J. Riemannian approach to batch normalization. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

Cordero-Erausquin, D., McCann, R. J., and Schmuckenschläger, M. A Riemannian interpolation inequality à la Borell, Brascamp and Lieb. *Inventiones mathematicae*, 146(2):219–257, 2001.

Defazio, A. and Mishchenko, K. Learning-rate-free learning by D-adaptation. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, Honolulu, HI, 2023.

Edelman, A., Arias, T. A., and Smith, S. T. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Grimmer, B. Convergence rates for deterministic and stochastic subgradient methods without Lipschitz continuity. *SIAM Journal on Optimization*, 29(2):1350–1365, 2019.

Gromov, M. Hyperbolic groups. In *Essays in group theory*, pp. 75–263. Springer, 1987.

Hazan, E., Levy, K., and Shalev-Shwartz, S. Beyond convexity: Stochastic quasi-convex optimization. In *Proceedings of the 28th Conference on Neural Information Processing Systems (NIPS 2015)*, volume 28, 2015.

Hosseini, R. and Sra, S. Matrix manifold optimization for Gaussian mixtures. In *Proceedings of the 28th Conference on Neural Information Processing Systems (NeurIPS 2015)*, volume 28, 2015.

Ishteva, M., Absil, P.-A., Van Huffel, S., and De Lathauwer, L. Best low multilinear rank approximation of higher-order tensors, based on the Riemannian trust-region scheme. *SIAM Journal on Matrix Analysis and Applications*, 32(1):115–135, 2011.

Ivgi, M., Hinder, O., and Carmon, Y. DoG is SGD's best friend: A parameter-free dynamic step size schedule. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 14465–14499. PMLR, 2023.

Kasai, H., Sato, H., and Mishra, B. Riemannian stochastic variance reduced gradient on Grassmann manifold. *arXiv preprint arXiv:1605.07367*, 2017.

Kasai, H., Jawanpuria, P., and Mishra, B. Riemannian adaptive stochastic gradient algorithms on matrix manifolds. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3262–3271. PMLR, 2019.

Khaled, A., Mishchenko, K., and Jin, C. DoWG unleashed: An efficient universal parameter-free gradient descent method. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023.

Konnov, I. V. On convergence properties of a subgradient method. *Optimization Methods and Software*, 18(1):53–62, 2003.

Lee, J. M. *Smooth manifolds*. Springer, 2012.

Levy, K. Y. The power of normalization: Faster evasion of saddle points. *arXiv preprint arXiv:1611.04831*, 2016.

Liu, X., Srivastava, A., and Gallivan, K. Optimal linear representations of images for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):662–666, 2004.

Liu, Y., Shang, F., Cheng, J., Cheng, H., and Jiao, L. Accelerated first-order methods for geodesically convex optimization on Riemannian manifolds. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

McMahan, H. B. and Orabona, F. Unconstrained online linear learning in Hilbert spaces: Minimax algorithms and normal approximations. In *Proceedings of the 27th Conference on Learning Theory (COLT 2014)*, Barcelona, Spain, 2014.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4): 235–244, 12 1990.

Nesterov, Y. *Lectures on Convex Optimization*. Number 978-3-319-91578-4 in Springer Optimization and Its Applications. Springer, 2018.

Nickel, M. and Kiela, D. Poincaré embeddings for learning hierarchical representations. In *Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS 2017)*, volume 30, 2017.

Orabona, F. Dimension-free exponentiated gradient. In *Proceedings of the 26th Conference on Neural Information Processing Systems*, 2013.

Orabona, F. and Cutkosky, A. Tutorial on Parameter-Free Online Learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, 2020.

Orabona, F. and Pál, D. Coin betting and parameter-free online learning. In *Proceedings of the 29th Conference on Neural Information Processing Systems (NeurIPS 2016)*, volume 29, 2016.

Orabona, F. and Tommasi, T. Training Deep Networks without Learning Rates Through Coin Betting. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, 2017.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Petersen, P. *Riemannian geometry*, volume 171. Springer, 2006.

Roy, S. K., Mhammedi, Z., and Harandi, M. Geometry aware constrained optimization techniques for deep learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4460–4469, 2018.

Sala, F., De Sa, C., Gu, A., and Ré, C. Representation tradeoffs for hyperbolic embeddings. In *International Conference on Machine Learning (ICML 2018)*, pp. 4460–4469. PMLR, 2018.

Sato, H., Kasai, H., and Mishra, B. Riemannian stochastic variance reduced gradient algorithm with retraction and vector transport. *SIAM Journal on Optimization*, 29(2): 1444–1472, 2019.

Sharrock, L. and Nemeth, C. Coin Sampling: Gradient-Based Bayesian Inference without Learning Rates. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, Honolulu, HI, 2023.

Sharrock, L., Mackey, L., and Nemeth, C. Learning Rate Free Sampling in Constrained Domains. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS 2023)*, New Orleans, LA, 2023.

Sharrock, L., Dodd, D., and Nemeth, C. Tuning-free maximum likelihood training of latent variable models via coin betting. In *International Conference on Artificial Intelligence and Statistics*, pp. 1810–1818. PMLR, 2024.

Shor, N. Z. *Minimization methods for non-differentiable functions*, volume 3. Springer Science & Business Media, 2012.

Streeter, M. and McMahan, H. B. No-regret algorithms for unconstrained online convex optimization. In *Proceedings of the 25th Conference on Neural Information Processing Systems (NIPS 2012)*, 2012.

Sun, J., Qu, Q., and Wright, J. Complete dictionary recovery over the sphere ii: Recovery by Riemannian trust-region method. *IEEE Transactions on Information Theory*, 63 (2):885–914, February 2017.

Townsend, J., Koep, N., and Weichwald, S. Pymanopt: A Python toolbox for optimization on manifolds using automatic differentiation. *Journal of Machine Learning Research*, 17(137):1–5, 2016.

Udrişte, C. *Convex Functions and Optimization Methods on Riemannian Manifolds*. Springer Dordrecht, 1994.

Ungar, A. A. A gyrovector space approach to hyperbolic geometry. *Synthesis Lectures on Mathematics and Statistics*, 1(1):1–194, 2008.

Villani, C. *Topics in Optimal Transportation*. American Mathematical Society, Providence, Rhode Island, 2003.

Wang, X., Tu, Z., Hong, Y., Wu, Y., and Shi, G. No-regret online learning over Riemannian manifolds. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, 2021.

Weber, M. and Sra, S. Projection-free nonconvex stochastic optimization on Riemannian manifolds. *IMA Journal of Numerical Analysis*, 42(4):3241–3271, 09 2021.

Zadeh, P., Hosseini, R., and Sra, S. Geometric mean metric learning. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of the 33rd International Conference on Machine Learning (ICML 2016)*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2464–2471, New York, New York, USA, 2016. PMLR.

Zhang, H. and Sra, S. First-order methods for geodesically convex optimization. In Feldman, V., Rakhlin, A., and Shamir, O. (eds.), *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 1617–1638, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

Zhang, H. and Sra, S. Towards Riemannian accelerated gradient methods. *arXiv preprint arXiv:1806.02812*, 2018.

Zhang, H., Reddi, S. J., and Sra, S. Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. In *Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016)*, 2016.

Zhou, P., Yuan, X.-T., Yan, S., and Feng, J. Faster first-order methods for stochastic non-convex optimization on Riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):459–472, 2021.

A. Useful Results

We begin by introducing essential lemmas for the establishment of our theory.

A.1. Trigonometric Distance Bounds for Manifolds

The law of cosines in Euclidean space is fundamental for analyzing optimization algorithms,

$$a^2 = b^2 + c^2 - 2bc\cos(A), \tag{2}$$

where a, b, c are the sides of a Euclidean triangle with A the angle between sides b and c.

Trigonometric geometry behaves differently in manifolds compared to Euclidean spaces. While the equality does not hold for nonlinear spaces, a trigonometric distance bound can be established for manifolds with curvature bounded below.

Lemma A.1. *(Zhang & Sra, 2016, Lemma 5) If a, b, c are the side lengths of a geodesic triangle Δ in a Riemannian manifold with sectional curvature lower bounded by $\kappa > -\infty$ and A is the angle between sides b and c (defined through the inverse exponential map and inner product in tangent space), then*

$$a^2 \leq \zeta_\kappa(c)b^2 + c^2 - 2bc\cos(A). \tag{3}$$

Proof. Given by Lemma 3.12 of (Cordero-Erausquin et al., 2001) and by Lemma 5 of (Zhang & Sra, 2016). □

This lemma holds profound implications for our analysis of geodesically convex functions f. Specifically, the property of geodesic convexity allows us to bound $f(x_t) - f(x_\star)$ by the inner product $\langle -\operatorname{grad} f(x_t), \exp_{x_t}^{-1}(x_\star)\rangle_{x_t}$. The above trigonometric inequality empowers us to bound this inner product to devise tractable optimization algorithms.

To streamline future analysis, we expand our perspective to encompass bounding the inner product $\langle -g_t, \exp_{x_t}^{-1}(x_\star)\rangle_{x_t}$ for any tangent vector $g_t \in \mathcal{T}_{x_t}\mathcal{M}$.

Lemma A.2. *(Zhang & Sra, 2016, Corollary 8) For any Riemannian manifold $\mathcal{M}$ where the sectional curvature is lower bounded by $\kappa > -\infty$ and any point $x_\star, x_t \in \mathcal{M}$ and any tangent vector $g_t \in \mathcal{T}_{x_t}\mathcal{M}$, scalar $\eta_t > 0$ consider the RSGD update $x_{t+1} = \exp_{x_t}(-\eta_t g_t)$. Then by Lemma A.1, we have*

$$\langle -g_t, \exp_{x_t}^{-1}(x_\star)\rangle_{x_t} \leq \frac{1}{2\eta_t}\left(d_t^2 - d_{t+1}^2\right) + \frac{\eta_t}{2}\zeta_\kappa(d_t)\|g_t\|_{x_t}^2. \tag{4}$$

Proof. Consider the geodesic triangle Δ with vertices x_{t+1}, x_t, and $x_\star$. Then we have the side lengths of Δ are given by

$$a = d(x_{t+1}, x_\star) = d_{t+1}, \quad b = d(x_{t+1}, x_t) = \eta_t\|g_t\|_{x_t}, \quad c = d(x_t, x_\star) = d_t. \tag{5}$$

Recalling that the angle between two tangent vectors u and v at $x \in \mathcal{M}$ is given by $\arccos \frac{\langle u, v\rangle_x}{\|u\|_x\|v\|_x}$. Now, considering the angle, A, between side lengths b and c, we have,

$$2bc\cos(A) = 2bc\cos\left(\arccos\left(\frac{\langle \exp_{x_t}^{-1}(x_{t+1}), \exp_{x_t}^{-1}(x_\star)\rangle_{x_t}}{\|\exp_{x_t}^{-1}(x_{t+1})\|_{x_t}\|\exp_{x_t}^{-1}(x_\star)\|_{x_t}}\right)\right) = \langle -\eta_t g_t, \exp_{x_t}^{-1}(x_\star)\rangle_{x_t}. \tag{6}$$

Substituting these terms in Lemma A.1 and rearranging yields the result as required. □

A.2. Jensen's Inequality for Geodesically Convex Functionals

We present an analog for Jensen's inequality for geodesically convex functions on Riemannian manifolds. This will allow us to leverage innovative weighted averaging strategies in the regret analysis of our algorithms.

Lemma A.3. *Let f be geodesically convex. For any sequence of iterates $x_0, \ldots, x_t \in \mathcal{M}$ and positive weights $w_0, \ldots, w_t \in \mathbb{R}_+$, define the online weighted average sequence by*

$$\tilde{x}_{t+1} = \exp_{\tilde{x}_t}\left(\frac{w_t}{\sum_{s=0}^t w_s}\exp_{\tilde{x}_t}^{-1}(x_t)\right), \qquad \tilde{x}_1 = x_0. \tag{7}$$

Then we have

$$f(\tilde{x}_t) \leq \frac{1}{\sum_{s=0}^{t-1} w_s} \sum_{s=0}^{t-1} w_s f(x_s). \tag{8}$$

Proof. We prove this by induction. The base case for $t = 1$ holds by definition. Now for $t \geq 2$, for the inductive step, assume the statement is true for $t-1$ and consider t. We have,

$$\frac{1}{\sum_{s=0}^{t-1} w_s} \sum_{s=0}^{t-1} w_s f(x_s) = \frac{w_{t-1}}{\sum_{s=0}^{t-1} w_s} f(x_{t-1}) + \frac{1}{\sum_{s=0}^{t-1} w_s} \sum_{s=0}^{t-2} w_s f(x_s) \tag{9}$$

$$= \frac{w_{t-1}}{\sum_{s=0}^{t-1} w_s} f(x_{t-1}) + \frac{\sum_{s=0}^{t-2} w_s}{\sum_{s=0}^{t-1} w_s} \frac{1}{\sum_{s=0}^{t-2} w_s} \sum_{s=0}^{t-2} w_s f(x_s) \tag{10}$$

$$\geq \frac{w_{t-1}}{\sum_{s=0}^{t-1} w_s} f(x_{t-1}) + \frac{\sum_{s=0}^{t-2} w_s}{\sum_{s=0}^{t-1} w_s} f(\tilde{x}_{t-1}). \tag{11}$$

In the final line, we have exploited the inductive assumption. Finally, we note that $\gamma(s) = \exp_x\left((1-s)\exp_x^{-1}(x) + s\exp_x^{-1}(y)\right)$ for $s \in [0,1]$ defines a geodesic between any two points x and y in $\mathcal{M}$. Moreover, by geodesic convexity we have

$$f(\gamma(s)) \leq (1-s)f(\gamma(0)) + sf(\gamma(1)) = (1-s)f(x) + sf(y). \tag{12}$$

Thus applying this to Equation (11) with $x = \tilde{x}_{t-1}$, $y = x_{t-1}$ and $s = \frac{w_{t-1}}{\sum_{s=0}^{t-1} w_s}$ and noting that for this choice,

$$\gamma\left(\frac{w_{t-1}}{\sum_{s=0}^{t-1} w_s}\right) = \exp_{\tilde{x}_{t-1}}\left(\left(1 - \frac{w_{t-1}}{\sum_{s=0}^{t-1} w_s}\right)\exp^{-1}_{\tilde{x}_{t-1}}(\tilde{x}_{t-1}) + \frac{w_{t-1}}{\sum_{s=0}^{t-1} w_s}\exp^{-1}_{\tilde{x}_{t-1}}(x_{t-1})\right) \tag{13}$$

$$= \exp_{\tilde{x}_{t-1}}\left(\frac{w_{t-1}}{\sum_{s=0}^{t-1} w_s}\exp^{-1}_{\tilde{x}_{t-1}}(x_{t-1})\right) \tag{14}$$

$$= \tilde{x}_t, \tag{15}$$

yields the result as required. □

A.3. Smoothness Bounds

We present smoothness results that establish bounds on individual gradient norms, that we will use in our later analysis to yield tighter regret bounds under the geodesic smoothness assumption.

Lemma A.4. *Suppose f is S-smooth and lower bounded by $f(x_\star)$. Then, for all $x \in \mathcal{M}$ we have*

$$\|\operatorname{grad} f(x)\|_x \leq \sqrt{2S(f(x) - f(x_\star))}. \tag{16}$$

Proof. This is a trivial consequence of e.g., Proposition 4.7 and 4.8 of (Boumal, 2023). We include the proof for completeness. Let $x \in M$ and define $y = \exp_x\left(-\frac{1}{S}\operatorname{grad} f(x)\right)$. Then geodesic smoothness provides,

$$f(y) \leq f(x) + \langle \operatorname{grad} f(x), \exp_x^{-1}(y)\rangle_x + \frac{S}{2}\|\exp_x^{-1}(y)\|_x^2 \tag{17}$$

$$= f(x) - \frac{1}{S}\|\operatorname{grad} f(x)\|_x^2 + \frac{1}{2S}\|\operatorname{grad} f(x)\|_x^2 \tag{18}$$

$$= f(x) - \frac{1}{2S}\|\operatorname{grad} f(x)\|_x^2. \tag{19}$$

Now since f is lower bounded by $f(x_\star)$ we thus have

$$f(x_\star) \leq f(y) \leq f(x) - \frac{1}{2S}\|\operatorname{grad} f(x)\|_x^2. \tag{20}$$

Rearranging gives the result. □

Using the above argument, we provide a bound on the norm of the stochastic error.

Lemma A.5. *Under locally smooth stochastic gradients (Assumption 3.4), for the stochastic error* $\Delta(x) \coloneqq \mathcal{G}(x) - \operatorname{grad} f(x)$ *we almost surely have that*

$$\|\Delta(x)\|_x \leq (\sqrt{s(x)} + \sqrt{S})\sqrt{2(f(x) - f(x_\star))}. \tag{21}$$

Proof. Noting that Assumption 3.4 implies that for any $x, y \in \mathcal{M}$ we almost surely have that

$$f(s) \leq f(x) + \langle \mathcal{G}(x), \exp_x^{-1}(y)\rangle_x + \frac{s(x)}{2}\|\exp_x^{-1}(y)\|_x^2. \tag{22}$$

We follow the same argument as in Lemma A.4 to deduce that almost surely,

$$\|\mathcal{G}(x)\|_x \leq \sqrt{2s(x)(f(x) - f(x_\star))}. \tag{23}$$

While the triangle inequality and applying Lemma A.4 to $\|\operatorname{grad} f(x)\|_x$ gives,

$$\|\Delta(x)\|_x \leq \|\mathcal{G}(x)\|_x + \|\operatorname{grad} f(x)\|_x \leq \sqrt{2s(x)(f(x) - f(x_\star))} + \sqrt{2S(f(x) - f(x_\star))}. \tag{24}$$

□

A.4. Bounds for Real-Valued Series

Lemma A.6. *(Ivgi et al., 2023, Lemma 3) Let* $a_0, a_1, \ldots, a_T$ *be a positive increasing sequence. Then*

$$\max_{t \leq T} \sum_{s=0}^{t-1} \frac{a_s}{a_t} \geq e^{-1}\left(\frac{T}{1 + \log(a_T/a_0)} - 1\right). \tag{25}$$

Proof. Lemma 3 of (Ivgi et al., 2023). Define $K \coloneqq \lceil \log(a_T/a_0) \rceil$, and $n \coloneqq \lfloor T/K \rfloor$. Then, given the sequence is increasing we have

$$\log\left(\frac{a_T}{a_0}\right) \geq \sum_{k=0}^{K-1} \log\left(\frac{a_{n(k+1)}}{a_{nk}}\right) \geq K \min_{k<K} \log\left(\frac{a_{n(k+1)}}{a_{nk}}\right). \tag{26}$$

Rearranging gives,

$$\min_{k<K} \log\left(\frac{a_{n(k+1)}}{a_{nk}}\right) \leq \log\left(\frac{a_T}{a_0}\right)/K \leq 1 \implies \min_{k<K} \frac{a_{n(k+1)}}{a_{nk}} \leq e. \tag{27}$$

Thus,

$$\max_{t \leq T} \sum_{s=0}^{t-1} \frac{a_s}{a_t} \geq \max_{t \in [n,T]} n\frac{a_{t-n}}{a_t} = \max_{k \leq K} n\frac{a_{n(k-1)}}{a_{nk}} \geq ne^{-1} \tag{28}$$

$$= e^{-1}\left\lfloor \frac{T}{\lceil \log(a_T/a_0) \rceil} \right\rfloor \geq e^{-1}\left(\frac{T}{1 + \log(a_T/a_0)} - 1\right). \tag{29}$$

□

Lemma A.7. *(Ivgi et al., 2023, Lemma 4). Let* $a_0, \ldots, a_t$ *be a nondecreasing sequence of nonnegative numbers. Then*

$$\sum_{k=1}^{t} \frac{a_k - a_{k-1}}{\sqrt{a_k}} \leq 2\left(\sqrt{a_t} - \sqrt{a_0}\right). \tag{30}$$

Proof. This is a well-known result e.g., Lemma 4 of (Ivgi et al., 2023). We have

$$\sum_{k=1}^{t} \frac{a_k - a_{k-1}}{\sqrt{a_k}} = \sum_{k=1}^{t} \frac{(\sqrt{a_k} + \sqrt{a_{k-1}})(\sqrt{a_k} - \sqrt{a_{k-1}})}{\sqrt{a_k}} \tag{31}$$

$$\leq 2 \sum_{k=1}^{t} (\sqrt{a_k} - \sqrt{a_{k-1}}) = 2(\sqrt{a_t} - \sqrt{a_0}). \tag{32}$$

□

Lemma A.8. *(Ivgi et al., 2023, Lemma 6). Recall* $\log_+(z) \coloneqq 1 + \log(z)$. *Consider a non-decreasing sequence of nonnegative numbers,* $a_{-1}, a_0, a_1, \ldots, a_t$, *then*

$$\sum_{k=0}^{t} \frac{a_k - a_{k-1}}{a_k \log_+^2(a_k/a_{-1})} \leq 1. \tag{33}$$

Proof. Lemma 6 of (Ivgi et al., 2023). We have

$$\sum_{k=0}^{t} \frac{a_k - a_{k-1}}{a_k \log_+^2(a_k/a_{-1})} \leq \sum_{k=0}^{t} \int_{a_{k-1}/a_0}^{a_k/a_{-1}} \frac{\mathrm{d}\alpha}{\alpha \log_+^2(\alpha)} = \int_1^{a_t/a_{-1}} \frac{\mathrm{d}\alpha}{\alpha \log_+^2(\alpha)} \tag{34}$$

$$\leq \int_1^{\infty} \frac{\mathrm{d}\alpha}{\alpha \log_+^2(\alpha)} = \left[\frac{1}{1 + \log(\alpha)}\right]_1^{\infty} = 1. \tag{35}$$

□

A.5. Martingale Concentration bound

Lemma A.9. *(Ivgi et al., 2023, Lemma 7). Consider a filtration process* $\mathcal{F}_t$ *and let* $\mathbb{S}$ *be the set of nonnegative and nondecreasing sequences. Let* $C_t \in \mathcal{F}_{t-1}$ *and let* X_t *be a martingale difference sequence adapted to* $\mathcal{F}_{t-1}$ *such that* $|X_t| \leq C_t$ *with probability* 1 *for all* t. *Recalling that* $\theta_{t,\delta} \coloneqq \log(60 \log(6t)/\delta)$. *Then, for all* $\delta \in (0, 1)$, $c > 0$, *and* $\hat{X}_t \in \mathcal{F}_{t-1}$ *such that* $|\hat{X}_t| \leq C_t$ *with probability* 1,

$$\mathbb{P}\left(\exists t \leq T, \exists \{y_s\}_{s=1}^{\infty} \in \mathbb{S} : \left|\sum_{s=1}^{t} y_s X_s\right| \geq 8 y_t \sqrt{\theta_{t,\delta} \sum_{s=1}^{t} \left(X_s - \hat{X}_s\right)^2 + c^2 \theta_{t,\delta}^2}\right) \leq \delta + \mathbb{P}\left(\exists t \leq T : C_t > c\right). \tag{36}$$

Proof. See Lemma 7 of (Ivgi et al., 2023). □

B. RGD "Ideal Step Size" Analysis

B.1. Proof of Theorem 3.5

Proof. Using geodesic convexity and applying Lemma A.2, we have

$$\min_{t\le T}\left[f(x_t)-f(x_\star)\right] \le \frac{1}{T}\sum_{t=0}^{T}\left[f(x_t)-f(x_\star)\right] \tag{37}$$

$$\le \frac{1}{T}\sum_{t=0}^{T}\langle -g_t, \exp_{x_t}^{-1}(x_\star)\rangle_{x_t} \tag{38}$$

$$\le \frac{1}{T}\sum_{t=0}^{T}\left[\frac{1}{2\eta}\left(d_t^2-d_{t+1}^2\right)+\frac{\eta}{2}\zeta_\kappa(d_t)\|g_t\|_{x_t}^2\right] \tag{39}$$

$$= \frac{d_0^2}{2\eta T}+\frac{\eta\sum_{t=0}^{T}\zeta_\kappa(d_t)\|g_t\|_{x_t}^2}{2T} \tag{40}$$

$$\le \frac{\bar{d}_T^2}{2\eta T}+\frac{\eta\zeta_\kappa(\bar{d}_T)\sum_{t=0}^{T}\|g_t\|_{x_t}^2}{2T}. \tag{41}$$

Now, setting $\eta = \frac{\bar{d}_T}{\sqrt{\zeta_\kappa(\bar{d}_T)\sum_{t=0}^{T}\|g_t\|_{x_t}^2}}$, we have

$$\min_{t\le T}\left[f(x_t)-f(x_\star)\right] \le \frac{\bar{d}_T\sqrt{\zeta_\kappa(\bar{d}_T)\sum_{t=0}^{T}\|g_t\|_{x_t}^2}}{2T}+\frac{\zeta_\kappa(\bar{d}_T)\sum_{t=0}^{T}\|g_t\|_{x_t}^2}{2T\zeta_\kappa(\bar{d}_T)\sqrt{\sum_{t=0}^{T}\|g_t\|_{x_t}^2}} \tag{42}$$

$$= \frac{\bar{d}_T\sqrt{\zeta_\kappa(\bar{d}_T)\sum_{t=0}^{T}\|g_t\|_{x_t}^2}}{T} \tag{43}$$

$$\le \frac{L\bar{d}_T\sqrt{\zeta_\kappa(\bar{d}_T)}}{\sqrt{T}}. \tag{44}$$

Where we have bounded $\|g_t\|_{x_t}\le L$ due to the Lipschitz assumption, and $d_\infty \ge d_t$ for all $t\ge 0$.

□

C. NRGD "Ideal Step Size" Analysis

C.1. Proof of Theorem 3.11

Proof of Theorem 3.11. Using Lemma A.2 we have

$$\left\langle \frac{-\operatorname{grad} f(x_t)}{\|\operatorname{grad} f(x_t)\|_{x_t}}, \exp_{x_t}^{-1}(x_\star)\right\rangle_{x_t} \le \frac{1}{2\eta}\left(d_t^2-d_{t+1}^2\right)+\frac{\eta}{2}\zeta_\kappa(d_t). \tag{45}$$

Averaging the above, we have

$$\frac{1}{T}\sum_{t=0}^{T}\left\langle \frac{-\operatorname{grad} f(x_t)}{\|\operatorname{grad} f(x_t)\|_{x_t}}, \exp_{x_t}^{-1}(x_\star)\right\rangle_{x_t} \le \frac{d_0^2}{2\eta T}+\frac{\eta}{2T}\sum_{t=0}^{T}\zeta_\kappa(d_t). \tag{46}$$

Now for the Lipshitz setting, we have $\|\operatorname{grad} f(x_t)\|_{x_t}\le L$ thus,

$$\frac{1}{T}\sum_{t=0}^{T}\left\langle \frac{-\operatorname{grad} f(x_t)}{L}, \exp_{x_t}^{-1}(x_\star)\right\rangle_{x_t} \le \frac{1}{T}\sum_{t=0}^{T}\left\langle \frac{-\operatorname{grad} f(x_t)}{\|\operatorname{grad} f(x_t)\|_{x_t}}, \exp_{x_t}^{-1}(x_\star)\right\rangle_{x_t} \le \frac{d_0^2}{2\eta T}+\frac{\eta}{2T}\sum_{t=0}^{T}\zeta_\kappa(d_t). \tag{47}$$

Multiplying through by L and using definition of geodesic convexity yields,

$$\min_{t\le T}\left[f(x_t)-f(x_\star)\right] \le \frac{1}{T}\sum_{t=0}^{T}\left[f(x_t)-f(x_\star)\right] \le L\left[\frac{d_0^2}{2\eta T}+\frac{\eta}{2T}\sum_{t=0}^{T}\zeta_\kappa(d_t)\right] \le \left[\frac{\bar{d}_T^2}{2\eta T}+\frac{\eta}{2}\zeta_\kappa(\bar{d}_T)\right]. \tag{48}$$

Now substituting $\eta = \frac{\bar{d}_T}{\sqrt{T\zeta_\kappa(\bar{d}_T)}}$, gives

$$\min_{t\le T}\left[f(x_t) - f(x_\star)\right] \le \frac{L\bar{d}_T\sqrt{T\zeta_\kappa(\bar{d}_T)}}{T} \le \frac{L\bar{d}_T\sqrt{\zeta_\kappa(\bar{d}_T)}}{\sqrt{T}}, \tag{49}$$

which completes the proof for the Lipshitz case.

Now we proceed to consider the smooth setting. By convexity we have

$$\langle -\operatorname{grad} f(x_t), \exp_{x_t}^{-1}(x_\star)\rangle_{x_t} \ge f(x_t) - f(x_\star) \ge 0. \tag{50}$$

And by smoothness (Lemma A.4), we have

$$\|\operatorname{grad} f(x_t)\|_{x_t} \le \sqrt{2S(f(x_t) - f(x_\star))}. \tag{51}$$

Now if $f(x_t) = f(x_\star)$ the theorem holds trivially, suppose not. Then combining the above expressions, we have

$$\left\langle \frac{-\operatorname{grad} f(x_t)}{\|\operatorname{grad} f(x_t)\|_{x_t}}, \exp_{x_t}^{-1}(x_\star)\right\rangle_{x_t} \ge \frac{f(x_t) - f(x_\star)}{\sqrt{2S(f(x_t) - f(x_\star))}} = \frac{\sqrt{(f(x_t) - f(x_\star))}}{\sqrt{2S}}. \tag{52}$$

Thus we have

$$\min_{t\le T}\left[\sqrt{f(x_t) - f(x_\star)}\right] \le \frac{1}{T}\sum_{t=0}^{T}\sqrt{f(x_t) - f(x_\star)} \le \sqrt{2S}\left[\frac{d_0^2}{2\eta T} + \frac{\eta}{2T}\sum_{t=0}^{T}\zeta_\kappa(d_t)\right] \le \sqrt{2S}\left[\frac{\bar{d}_T^2}{2\eta T} + \frac{\eta}{2}\zeta_\kappa(\bar{d}_T)\right]. \tag{53}$$

Squaring gives us the first result. Now, plugging in $\eta = \frac{\bar{d}_T}{\sqrt{T\zeta_\kappa(\bar{d}_T)}}$, gives

$$\min_{t\le T}\left[f(x_t) - f(x_\star)\right] \le \frac{2S\bar{d}_T^2 T\zeta_\kappa(\bar{d}_T)}{T^2} \le \frac{2S\bar{d}_T^2\zeta_\kappa(\bar{d}_T)}{T}. \tag{54}$$

□

D. RDoG Theoretical Analysis

D.1. Overview

In this section, we analyze RDoG (Algorithm 1). Thus we consider RSGD with step sizes given by,

$$\eta_t = \frac{\bar{r}_t}{\sqrt{\zeta_\kappa(\bar{r}_t)\sum_{s=0}^{t}\|g_s\|_{x_s}^2}}, \tag{55}$$

We consider bounding the error of the weighted average sequence,

$$\tilde{x}_{t+1} = \exp_{\tilde{x}_t}\left(\frac{\bar{r}_t/\sqrt{\zeta_\kappa(\bar{r}_t)}}{\sum_{s=0}^{t}\bar{r}_s/\sqrt{\zeta_\kappa(\bar{r}_s)}}\exp_{\tilde{x}_t}^{-1}(x_t)\right), \quad \tilde{x}_1 = x_0.$$

For a geodesically convex function $f\colon \mathcal{M}\to\mathbb{R}$, we have by Lemma A.3 that $\tilde{x}_t$ satisfies,

$$f(\tilde{x}_t) - f(x_\star) \le \frac{1}{\sum_{s=0}^{t-1}(\bar{r}_s/\sqrt{\zeta_\kappa(\bar{r}_s)})}\sum_{s=0}^{t-1}(\bar{r}_s/\sqrt{\zeta_\kappa(\bar{r}_s)})\langle -\operatorname{grad} f(x_s), \exp_{x_s}^{-1}(x_\star)\rangle_{x_s}. \tag{56}$$

Recalling that g_s represents the stochastic oracle evaluation at x_s, denoted as $\mathcal{G}(x_s)$, we can decompose the numerator into two components:

$$\underbrace{\sum_{s=0}^{t-1}(\bar{r}_s/\sqrt{\zeta_\kappa(\bar{r}_s)})\langle -g_s, \exp_{x_s}^{-1}(x_\star)\rangle_{x_s}}_{\text{weighted regret}} + \underbrace{\sum_{s=0}^{t-1}(\bar{r}_s/\sqrt{\zeta_\kappa(\bar{r}_s)})\langle \Delta_s, \exp_{x_s}^{-1}(x_\star)\rangle_{x_s}}_{\text{noise}}, \tag{57}$$

with $\Delta_s \coloneqq g_s - \operatorname{grad} f(x_s)$.

D.2. Non-Smooth Analysis

We give deterministic bounds for the weighted regret (Lemma D.1) and high probability bounds for the noise term (Lemma D.2).

Lemma D.1. *Under Assumption 3.1 and 3.2, we have that the iterates of RDoG (Algorithm 1) satisfy*

$$\sum_{s=0}^{t-1}(\bar{r}_s/\sqrt{\zeta_\kappa(\bar{r}_s)})\langle -g_s, \exp_{x_s}^{-1}(x_\star)\rangle_{x_s} \leq \bar{r}_t\left(2\bar{d}_t + \frac{\bar{r}_t\zeta_\kappa(\bar{d}_t)}{\zeta_\kappa(\bar{r}_t)}\right)\sqrt{G_{t-1}}. \tag{58}$$

Proof. Applying Lemma A.2, we can bound the weighted average as

$$\sum_{s=0}^{t-1}\left(\bar{r}_s/\sqrt{\zeta_\kappa(\bar{r}_s)}\right)\langle -g_s, \exp_{x_s}^{-1}(x_\star)\rangle_{x_s} \leq \underbrace{\frac{1}{2}\sum_{s=0}^{t-1}\frac{\left(\bar{r}_s/\sqrt{\zeta_\kappa(\bar{r}_s)}\right)}{\eta_s}\left(d_s^2 - d_{s+1}^2\right)}_{(A)} + \underbrace{\frac{1}{2}\sum_{s=0}^{t-1}\left(\bar{r}_s/\sqrt{\zeta_\kappa(\bar{r}_s)}\right)\eta_s\zeta_\kappa(d_s)\|g_s\|_{x_s}^2}_{(B)}. \tag{59}$$

We bound the terms (A) and (B) in turn, beginning with the former:

$$(A) = \sum_{s=0}^{t-1}\sqrt{G_s}\left(d_s^2 - d_{s+1}^2\right) = d_0^2\sqrt{G_0} - d_t^2\sqrt{G_{t-1}} + \sum_{s=0}^{t-1} d_s^2\left(\sqrt{G_s} - \sqrt{G_{s-1}}\right) \tag{60}$$

$$\overset{(i)}{\leq} \bar{d}_t^2\sqrt{G_0} - d_t^2\sqrt{G_{t-1}} + \bar{d}_t^2\sum_{s=0}^{t-1}\left(\sqrt{G_s} - \sqrt{G_{s-1}}\right) = \sqrt{G}_{t-1}(\bar{d}_t^2 - d_t^2) \overset{(ii)}{\leq} 4\bar{r}_t\bar{d}_t\sqrt{G_{t-1}}. \tag{61}$$

Inequality (i) uses $d_s \leq \bar{d}_t$ and that G_t is nondecreasing. Inequality (ii) use that for $k \in \arg\max_{s\leq t} d_s$, we have $\bar{d}_t^2 - d_t^2 = d_k^2 - d_t^2 = (d_k - d_t)(d_k + d_t) \leq d(x_k, x_t)(d_k + d_t) \leq (\bar{r}_k + \bar{r}_t)(d_k + d_t) \leq 4\bar{r}_t\bar{d}_t$. Bounding the second term (B), we have for $\kappa < 0$:

$$(B) = \sum_{s=0}^{t-1}\frac{\bar{r}_s^2\zeta_\kappa(d_s)\|g_s\|_{x_s}^2}{\zeta_\kappa(\bar{r}_s)\sqrt{G_s}} = \sum_{s=0}^{t-1}\frac{\bar{r}_s^2\tanh(\sqrt{|\kappa|}\cdot\bar{r}_s)\zeta_\kappa(d_s)\|g_s\|_{x_s}^2}{\sqrt{|\kappa|}\cdot\bar{r}_s\sqrt{G_s}} = \sum_{s=0}^{t-1}\frac{\bar{r}_s\tanh(\sqrt{|\kappa|}\cdot\bar{r}_s)\zeta_\kappa(d_s)\|g_s\|_{x_s}^2}{\sqrt{|\kappa|}\cdot\sqrt{G_s}} \tag{62}$$

$$\leq \frac{1}{\sqrt{|\kappa|}}\bar{r}_t\tanh(\sqrt{|\kappa|}\cdot\bar{r}_t)\zeta_\kappa(\bar{d}_t)\sum_{s=0}^{t-1}\frac{\|g_s\|_{x_s}^2}{\sqrt{G_s}} \leq \frac{2}{\sqrt{|\kappa|}}\bar{r}_t\tanh(\sqrt{|\kappa|}\bar{r}_t)\zeta_\kappa(\bar{d}_t)\sqrt{G_{t-1}} \tag{63}$$

$$= \frac{2\bar{r}_t^2\tanh\left(\sqrt{|\kappa|}\cdot\bar{r}_t\right)}{\sqrt{|\kappa|}\cdot\bar{r}_t}\zeta_\kappa(\bar{d}_t)\sqrt{G_{t-1}} = \frac{2\bar{r}_t^2}{\zeta_\kappa(\bar{r}_t)}\zeta_\kappa(\bar{d}_t)\sqrt{G_{t-1}}. \tag{64}$$

While for $\kappa = 0$ the geometric curvature function $d \mapsto \zeta_\kappa(d)$ takes constant value one, thus the same bound above can be established trivially. Combining (A) and (B), gives the result. □

Lemma D.2. *For all $\delta \in (0,1), T \in \mathbb{N}$ and $L > 0$, if Assumption 3.1, 3.2, and 3.3 hold, the iterates of RDoG (Algorithm 1) satisfy*

$$\mathbb{P}\left(\exists t \leq T : \left|\sum_{s=0}^{t-1}\frac{\bar{r}_s}{\sqrt{\zeta_\kappa(\bar{r}_s)}}\langle -\Delta_s, \exp_{x_s}^{-1}(x_\star)\rangle_{x_s}\right| \geq b_t\right) \leq \delta + \mathbb{P}(\bar{\ell}_T > L), \tag{65}$$

where $b_t = 8\frac{\bar{r}_{t-1}}{\sqrt{\zeta_\kappa(\bar{r}_{t-1})}}\bar{d}_{t-1}\sqrt{\theta_{t,\delta}G_{t-1} + \theta_{t,\delta}^2L^2}$ and $\bar{\ell}_T \coloneqq \max_{s\leq T}\ell(x_s)$.

Proof. For $1 \leq s \leq T$ define the random variables

$$Y_s \coloneqq \frac{\bar{r}_{s-1}}{\sqrt{\zeta_\kappa(\bar{r}_{s-1})}}\bar{d}_{s-1}, \quad X_s \coloneqq \left\langle \Delta_{s-1}, \frac{\exp_{x_{s-1}}^{-1}(x_\star)}{\bar{d}_{s-1}}\right\rangle_{x_{s-1}}, \quad \hat{X}_s \coloneqq \left\langle -\operatorname{grad} f(x_{s-1}), \frac{\exp_{x_{s-1}}^{-1}(x_\star)}{\bar{d}_{s-1}}\right\rangle_{x_{s-1}}. \tag{66}$$

By the Cauchy-Schwartz inequality and Assumption 3.3, we have each $|X_s| \leq \ell(x)$, and each $|\hat{X}_s| \leq \ell(x)$ with probability 1. Moreover, and consider the filtration $\mathcal{F}_s = \sigma(\mathcal{G}(x_0) \ldots, \mathcal{G}(x_s))$. Then we have that X_s is a martingale difference sequence adapted to $\mathcal{F}_s$ and $\hat{X}_s \in \mathcal{F}_{s-1}$. By construction or any $t \leq T$, we have

$$\sum_{s=1}^{t} Y_s X_s = \sum_{s=0}^{t-1} \frac{\bar{r}_s}{\sqrt{\zeta_\kappa(\bar{r}_s)}} \langle \Delta_s, \exp_{x_s}^{-1}(x_\star) \rangle_{x_s}. \tag{67}$$

Therefore, applying Lemma A.9 yields the result as required. □

Combining the above results, we obtain the following.

Theorem D.3. *For all $\delta \in (0,1)$ and $L > 0$, if Assumption 3.1, 3.2, and 3.3 hold, then with probability at least $1 - \delta - \mathbb{P}(\bar{\ell}_T > L)$, for all $t \leq T$, the optimality gap on the weighted iterates $f(\tilde{x}_t) - f(x_\star)$ of RDoG (Algorithm 1) satisfy*

$$O\left(\frac{(d_0 + \bar{r}_t)\sqrt{\zeta_\kappa(d_0 + \bar{r}_t)}\sqrt{G_{t-1} + \theta_{t,\delta} G_{t-1} + \theta_{t,\delta}^2 L^2}}{\sum_{s=0}^{t-1} \frac{\bar{r}_s / \sqrt{\zeta_\kappa(\bar{r}_s)}}{\bar{r}_t / \sqrt{\zeta_\kappa(\bar{r}_t)}}} \right). \tag{68}$$

Proof. Combining Lemma D.1 and Lemma D.2, we have for the given probability that

$$f(\tilde{x}_t) - f(x_\star) \leq \frac{\left(2\bar{d}_t \sqrt{\zeta_\kappa(\bar{r}_t)} + \frac{\bar{r}_t}{\sqrt{\zeta_\kappa(\bar{r}_t)}} \zeta_\kappa(\bar{d}_t) \right) \sqrt{G_{t-1}} + 8\bar{d}_{t-1} \sqrt{\theta_{t,\delta} G_{t-1} + \theta_{t,\delta}^2 L^2}}{\sum_{s=0}^{t-1} \frac{\bar{r}_s / \sqrt{\zeta_\kappa(\bar{r}_s)}}{\bar{r}_t / \sqrt{\zeta_\kappa(\bar{r}_t)}}}. \tag{69}$$

Now using the fact $\bar{d}_t \leq d_0 + \bar{r}_t$ and that $d \mapsto \zeta_\kappa(d)$ and $d \mapsto \frac{d}{\sqrt{\zeta_\kappa(d)}}$ are increasing functions gives the result. □

We then have a useful result when the manifold is bounded but its exact diameter is unknown.

Corollary D.4. *Under Assumption 3.1, 3.2, and 3.3, for any $D \geq d_0$ let $L_D \coloneqq \max_{x \in \mathcal{M}: d(x, x_0) \leq D} \ell(x)$. Then, for all $\delta \in (0,1)$ and for $\tau \in \arg\max_{t \leq T} \sum_{s=0}^{t-1} \frac{\bar{r}_s / \sqrt{\zeta_\kappa(\bar{r}_s)}}{\bar{r}_t / \sqrt{\zeta_\kappa(\bar{r}_t)}}$, with probability at least $1 - \delta - \mathbb{P}(\bar{\ell}_T > L)$, iterates of RDoG (Algorithm 1) satisfy the optimality gap bound*

$$f(\tilde{x}_\tau) - f(x_\star) = O\left(\frac{D\sqrt{\zeta_\kappa(D)}\sqrt{G_{\tau-1}\theta_{\tau,\delta} + L_D^2 \theta_{\tau,\delta}^2}}{T} \log_+ \left(\frac{D / \sqrt{\zeta_\kappa(D)}}{\epsilon / \sqrt{\zeta_\kappa(\epsilon)}} \right) \right). \tag{70}$$

Proof. Apply Lemma A.6 to the denominator term of Theorem D.3. □

D.3. Smooth Guarantees via Uniform Averaging

Under the assumption of locally smooth stochastic gradients (Assumption 3.4), we can deduce an $O(1/T)$ convergence guarantee under uniformly averaged iterates,

$$\hat{x}_{t+1} = \exp_{\hat{x}_t}\left(\frac{1}{t} \exp_{\hat{x}_t}^{-1}(x_t) \right), \quad \hat{x}_1 = x_0.$$

We begin by presenting a theorem that shows a bound under uniform iterate averaging in the non-smooth setting.

Theorem D.5. *For all $\delta \in (0,1)$ and $L > 0$, if Assumption 3.1, 3.2, and 3.3 hold, then with probability at least $1 - \delta - \mathbb{P}(\bar{\ell}_T > L)$, for all $t \leq T$, the optimality gap on the uniformly averaged iterates $f(\hat{x}_T) - f(x_\star)$ of RDoG (Algorithm 1) satisfy:*

$$O\left(\frac{(d_0 \log_+ \frac{\bar{r}_T}{\epsilon} + \bar{r}_T)\sqrt{\zeta_\kappa(d_0 + \bar{r}_T)}\sqrt{G_{T-1} + \theta_{T,\delta} G_{T-1} + \theta_{T,\delta}^2 L^2}}{T} \right). \tag{71}$$

Proof. Define the times $\tau_s = \min\left\{\min\{k|\bar{r}_k \geq 2\bar{r}_{\tau_{k-1}}\}, T\right\}$, with $\tau_0 \coloneqq 0$. Moreover, let K be the first index such that $\tau_K = T$ and note that $K \leq 1 + \log_2 \frac{\bar{r}_T}{\epsilon}$ by construction. Now using the argument of Lemma D.13, we have that for $k \leq K$

$$\sum_{t=\tau_{k-1}}^{\tau_k - 1} \frac{\bar{r}_t}{\sqrt{\zeta_\kappa(\bar{r}_t)}} \langle -g_t, \exp_{x_t}^{-1}(x_\star)\rangle \leq \bar{r}_{\tau_k}\left(2\bar{d}_{\tau_k} + \frac{\bar{r}_{\tau_k}}{\zeta_\kappa(\bar{r}_{\tau_k})}\zeta_\kappa(\bar{d}_{\tau_k})\right)\sqrt{G_{\tau_k - 1}} \tag{72}$$

$$= O\left(\bar{r}_{\tau_k}\frac{(d_0 + \bar{r}_{\tau_k})}{\zeta_\kappa(d_0 + \bar{r}_{\tau_k})}\zeta_\kappa(d_0 + \bar{r}_{\tau_k})\sqrt{G_{T-1}}\right) \tag{73}$$

$$= O\left(\bar{r}_{\tau_k}(d_0 + \bar{r}_{\tau_k})\sqrt{G_{T-1}}\right). \tag{74}$$

Where the first equality holds due by the virtue of $d \mapsto \frac{d}{\zeta_\kappa(d)}$ is an increasing function and that $\bar{d}_{\tau_k} \leq \bar{r}_{\tau_k} + d_0$. Furthermore, by Lemma D.14 we have for all $k \leq K$ with probability at least $1 - \delta - \mathbb{P}(\bar{\ell}_T > L)$,

$$\left|\sum_{t=\tau_{k-1}}^{\tau_k - 1} \frac{\bar{r}_t}{\sqrt{\zeta_\kappa(\bar{r}_t)}} \langle \Delta_t, \exp_{x_t}^{-1}(x_\star)\rangle_{x_t}\right| \leq \left|\sum_{t=0}^{\tau_k - 1} \frac{\bar{r}_t}{\sqrt{\zeta_\kappa(\bar{r}_t)}} \langle \Delta_t, \exp_{x_t}^{-1}(x_\star)\rangle_{x_t}\right| + \left|\sum_{t=0}^{\tau_{k-1} - 1} \frac{\bar{r}_t}{\sqrt{\zeta_\kappa(\bar{r}_t)}} \langle \Delta_t, \exp_{x_t}^{-1}(x_\star)\rangle_{x_t}\right| \tag{75}$$

$$\leq 16\frac{\bar{r}_{\tau_k - 1}}{\sqrt{\zeta_\kappa(\bar{r}_{\tau_k - 1})}}\bar{d}_{\tau_k - 1}\sqrt{\theta_{T,\delta}G_{T-1} + \theta_{T,\delta}^2 L^2}. \tag{76}$$

Now combining these two bounds, we have

$$\sum_{t=\tau_{k-1}}^{\tau_k - 1} f(x_t) - f(x_\star) \leq \frac{1}{\bar{r}_{\tau_{k-1}}/\sqrt{\zeta_\kappa(\bar{r}_{\tau_{k-1}})}} \sum_{t=\tau_{k-1}}^{\tau_k - 1} \frac{\bar{r}_t}{\sqrt{\zeta_\kappa(\bar{r}_t)}}[f(x_t) - f(x_\star)] \tag{77}$$

$$\leq \frac{1}{\bar{r}_{\tau_{k-1}}/\sqrt{\zeta_\kappa(\bar{r}_{\tau_{k-1}})}} \sum_{t=\tau_{k-1}}^{\tau_k - 1} \frac{\bar{r}_t}{\sqrt{\zeta_\kappa(\bar{r}_t)}} \langle -\operatorname{grad} f(x_t), \exp_{x_t}^{-1}(x_\star)\rangle_{x_t} \tag{78}$$

$$= \frac{1}{\bar{r}_{\tau_{k-1}}/\sqrt{\zeta_\kappa(\bar{r}_{\tau_{k-1}})}} \sum_{t=\tau_{k-1}}^{\tau_k - 1} \frac{\bar{r}_t}{\sqrt{\zeta_\kappa(\bar{r}_t)}} \left[\langle -g_t, \exp_{x_t}^{-1}(x_\star)\rangle_{x_t} + \langle \Delta_t, \exp_{x_t}^{-1}(x_\star)\rangle_{x_t}\right] \tag{79}$$

$$= O\left(\frac{\bar{r}_{\tau_k}/\sqrt{\zeta_\kappa(\bar{r}_{\tau_k})}}{\bar{r}_{\tau_{k-1}}/\sqrt{\zeta_\kappa(\bar{r}_{\tau_{k-1}})}}(d_0 + \bar{r}_{\tau_k})\sqrt{\zeta_\kappa(d_0 + \bar{r}_{\tau_k})}\sqrt{G_{T-1} + \theta_{T,\delta}G_{T-1} + \theta_{T,\delta}^2 L^2}\right) \tag{80}$$

$$= O\left((d_0 + \bar{r}_{\tau_k})\sqrt{\zeta_\kappa(d_0 + \bar{r}_{\tau_k})}\sqrt{G_{T-1} + \theta_{T,\delta}G_{T-1} + \theta_{T,\delta}^2 L^2}\right), \tag{81}$$

where final reduction holds since $d \mapsto \frac{d}{\zeta_\kappa(d)}$ is an increasing function, and for any t,

$$\bar{r}_{t+1} \leq \bar{r}_t + d(x_{t+1}, x_t) = \bar{r}_t\left(1 + \frac{\|g_t\|_{x_t}}{\sqrt{G_t}}\right) \leq 2\bar{r}_t. \tag{82}$$

Now summing over k from 1 to K we have

$$\sum_{t=0}^{T-1}[f(x_t) - f(x_\star)] = \sum_{k=1}^{K}\sum_{t=\tau_{k-1}}^{\tau_k - 1}[f(x_t) - f(x_\star)] \tag{83}$$

$$= O\left(\sum_{k=1}^{K}(d_0 + \bar{r}_{\tau_k})\sqrt{\zeta_\kappa(d_0 + \bar{r}_{\tau_k})}\sqrt{G_{T-1} + \theta_{T,\delta}G_{T-1} + \theta_{T,\delta}^2 L^2}\right) \tag{84}$$

$$\leq O\left(\sum_{k=1}^{K}(d_0 + \bar{r}_{\tau_k})\sqrt{\zeta_\kappa\left(d_0 + \sum_{k=1}^{K}\bar{r}_{\tau_k}\right)}\sqrt{G_{T-1} + \theta_{T,\delta}G_{T-1} + \theta_{T,\delta}^2 L^2}\right). \tag{85}$$

□

Where the final reduction holds since $d \mapsto \zeta_\kappa(d)$ is an increasing function. Now, recall that $K = O\left(\log_+ \frac{\bar{r}_T}{\epsilon}\right)$ and note that $\sum_{k=1}^{K} \bar{r}_{\tau_k} = O(\bar{r}_T)$ since $\frac{\bar{r}_{\tau_s}}{\bar{r}_{\tau_{K-1}}} \leq 2^{s-(K-1)}$ for all $s \leq K-1$. The proof is complete noting that via Lemma A.3 we deduce $f(\hat{x}_T) \leq \frac{1}{T}\sum_{t=0}^{T-1} f(x_t)$.

Now under smooth assumption, we present a result for bounding the stochastic term that depends on S.

Lemma D.6. *For all $\delta \in (0,1), T \in \mathbb{N}$ and $S > 0$, if Assumption 3.1, 3.2, and 3.4 hold, then the iterates of RDoG (Algorithm 1) satisfy*

$$\mathbb{P}\left(\exists t \leq T : \left|\sum_{s=0}^{t-1} \frac{\bar{r}_s}{\sqrt{\zeta_\kappa(\bar{r}_s)}} \langle \Delta_s, \exp_{x_s}^{-1}(x_\star)\rangle_{x_s}\right| \geq b_t\right) \leq \delta + \mathbb{P}(\bar{s}_T > S), \tag{86}$$

where $b_t = 8\frac{\bar{r}_{t-1}}{\sqrt{\zeta_\kappa(\bar{r}_{t-1})}}\bar{d}_{t-1}\sqrt{2S\theta_{t,\delta} + 8S\theta_{t,\delta}^2}\sqrt{\sum_{s=0}^{t-1}[f(x_s) - f(x_\star)]}$ and $\bar{s}_T \coloneqq \max_{k \leq T} s(x_k)$.

Proof. For $1 \leq s \leq T$ define the random variables

$$Y_s \coloneqq \frac{\bar{r}_{s-1}\bar{d}_{s-1}}{\sqrt{\zeta_\kappa(\bar{r}_{s-1})}}, \quad X_s \coloneqq \left\langle \Delta_{s-1}, \frac{\exp_{x_{s-1}}^{-1}(x_\star)}{\bar{d}_{s-1}} \right\rangle_{x_{s-1}}, \quad \hat{X}_s \coloneqq \left\langle -\operatorname{grad} f(x_{s-1}), \frac{\exp_{x_{s-1}}^{-1}(x_\star)}{\bar{d}_{s-1}} \right\rangle_{x_{s-1}}, \tag{87}$$

and consider the filtration $\mathcal{F}_s = \sigma(\mathcal{G}(x_0)\ldots,\mathcal{G}(x_s))$. Then we have that X_s is a martingale difference sequence adapted to $\mathcal{F}_s$ and $\hat{X}_s \in \mathcal{F}_{s-1}$. By construction or any $t \leq T$, we have

$$\sum_{s=1}^{t} Y_s X_s = \sum_{s=0}^{t-1} \bar{r}_s^2 \langle \Delta_s, \exp_{x_s}^{-1}(x_\star)\rangle_{x_s}. \tag{88}$$

Now we consider bounding, $\max\{|X_t|, |\hat{X}_t|\}$ by a constant c. Moreover, by the Cauchy-Schwartz inequality and Lemma A.5 we have with probability $\mathbb{P}(\bar{s}_T > S)$ we have

$$|X_s|^2 \leq \|\Delta_{s-1}\|_{x_{s-1}}^2 \cdot 1 \leq 8S(f(x_{s-1}) - f(x_\star)) \tag{89}$$

$$|\hat{X}_s|^2 \leq \|\operatorname{grad} f(x_{s-1})\|_{x_{s-1}}^2 \cdot 1 \leq 8S(f(x_{s-1}) - f(x_\star)). \tag{90}$$

Thus we have that,

$$|X_t| \leq \sqrt{\sum_{s=1}^{t} |X_s|^2} \leq \sqrt{8S}\sqrt{\sum_{s=0}^{t-1}[f(x_s) - f(x_\star)]} =: c \tag{91}$$

$$|\hat{X}_t| \leq \sqrt{\sum_{s=1}^{t} |\hat{X}_s|^2} \leq \sqrt{8S}\sqrt{\sum_{s=0}^{t-1}[f(x_s) - f(x_\star)]} =: c. \tag{92}$$

$$\tag{93}$$

Therefore, applying Lemma A.9 yields,

$$\mathbb{P}\left(\exists t \leq T : \left|\sum_{s=0}^{t-1} \frac{\bar{r}_s}{\zeta_\kappa(\bar{r}_s)} \langle \Delta_s, \exp_{x_s}^{-1}(x_\star)\rangle_{x_s}\right| \geq 8\frac{\bar{r}_{t-1}}{\sqrt{\zeta_\kappa(\bar{r}_{t-1})}}\bar{d}_{t-1}\sqrt{\theta_{t,\delta}\sum_{s=1}^{t}\left(X_s - \hat{X}_s\right)^2 + c^2\theta_{t,\delta}^2}\right) \leq \delta. \tag{94}$$

Now, finally noting

$$\sum_{s=1}^{t}(X_s - \hat{X}_s)^2 \leq \sum_{s=0}^{t-1}\|\operatorname{grad} f(x_s)\|_{x_s}^2 \leq 2S\sum_{s=0}^{t-1}(f(x_s) - f(x_\star)), \tag{95}$$

yields the result. □

Theorem D.7. *For all $\delta \in (0,1)$ and $S > 0$, if Assumption 3.1, 3.2, and 3.4 hold, then with probability at least $1 - \delta - \mathbb{P}(\bar{s}_T > S)$, for all $t \leq T$, the optimality gap of the uniformly averaged iterates $f(\hat{x}_T) - f(x_\star)$ of RDoG (Algorithm 1) satisfy:*

$$O\left(\frac{(d_0 \log_+ \frac{\bar{r}_T}{\epsilon} + \bar{r}_T)^2 \zeta_\kappa(d_0 + \bar{r}_T)\theta_{T,\delta}^2 S}{T}\right). \tag{96}$$

Proof. Similar to the non-smooth setting, define the times $\tau_s = \min\left\{\min\{k | \bar{r}_k \geq 2\bar{r}_{\tau_{k-1}}\}, T\right\}$, with $\tau_0 \coloneqq 0$. Moreover, let K be the first index such that $\tau_K = T$ and note that $K \leq 1 + \log_2 \frac{\bar{r}_T}{\epsilon}$ by construction. Now using the argument of Lemma D.13, we have that for $k \leq K$

$$\sum_{t=\tau_{k-1}}^{\tau_k - 1} \frac{\bar{r}_t}{\sqrt{\zeta_\kappa(\bar{r}_t)}} \langle -g_t, \exp_{x_t}^{-1}(x_\star)\rangle \leq \bar{r}_{\tau_k}\left(2\bar{d}_{\tau_k} + \frac{\bar{r}_{\tau_k}}{\zeta_\kappa(\bar{r}_{\tau_k})}\zeta_\kappa(\bar{d}_{\tau_k})\right)\sqrt{G_{\tau_k - 1}} \tag{97}$$

$$= O\left(\bar{r}_{\tau_k}(d_0 + \bar{r}_{\tau_k})\sqrt{G_{T-1}}\right) \tag{98}$$

$$\leq O\left(\bar{r}_{\tau_k}(d_0 + \bar{r}_{\tau_k})\sqrt{2S\sum_{t=0}^{T-1}[f(x_t) - f(x_\star)]}\right). \tag{99}$$

Where in the final inequality we have applied Lemma A.4. Now, by Lemma D.6 we have for all $k \leq K$ with probability at least $1 - \delta - \mathbb{P}(\bar{s}_T > L)$,

$$\left|\sum_{t=\tau_{k-1}}^{\tau_k - 1} \frac{\bar{r}_t}{\sqrt{\zeta_\kappa(\bar{r}_t)}} \langle \Delta_t, \exp_{x_t}^{-1}(x_\star)\rangle_{x_t}\right| \leq \left|\sum_{t=0}^{\tau_k - 1} \frac{\bar{r}_t}{\sqrt{\zeta_\kappa(\bar{r}_t)}} \langle \Delta_t, \exp_{x_t}^{-1}(x_\star)\rangle_{x_t}\right| + \left|\sum_{t=0}^{\tau_{k-1} - 1} \frac{\bar{r}_t}{\sqrt{\zeta_\kappa(\bar{r}_t)}} \langle \Delta_t, \exp_{x_t}^{-1}(x_\star)\rangle_{x_t}\right| \tag{100}$$

$$\leq 16\frac{\bar{r}_{\tau_k - 1}}{\sqrt{\zeta_\kappa(\bar{r}_{\tau_k - 1})}}\bar{d}_{\tau_k - 1}\sqrt{2S\theta_{T,\delta} + 8S\theta_{T,\delta}^2}\sqrt{\sum_{t=0}^{T-1}[f(x_t) - f(x_\star)]}. \tag{101}$$

Now combining these two bounds, we have

$$\sum_{t=\tau_{k-1}}^{\tau_k - 1} f(x_t) - f(x_\star) \leq \frac{1}{\bar{r}_{\tau_{k-1}}/\sqrt{\zeta_\kappa(\bar{r}_{\tau_{k-1}})}} \sum_{t=\tau_{k-1}}^{\tau_k - 1} \frac{\bar{r}_t}{\sqrt{\zeta_\kappa(\bar{r}_t)}}[f(x_t) - f(x_\star)] \tag{102}$$

$$\leq \frac{1}{\bar{r}_{\tau_{k-1}}/\sqrt{\zeta_\kappa(\bar{r}_{\tau_{k-1}})}} \sum_{t=\tau_{k-1}}^{\tau_k - 1} \frac{\bar{r}_t}{\sqrt{\zeta_\kappa(\bar{r}_t)}}\langle -\operatorname{grad} f(x_t), \exp_{x_t}^{-1}(x_\star)\rangle_{x_t} \tag{103}$$

$$= \frac{1}{\bar{r}_{\tau_{k-1}}/\sqrt{\zeta_\kappa(\bar{r}_{\tau_{k-1}})}} \sum_{t=\tau_{k-1}}^{\tau_k - 1} \frac{\bar{r}_t}{\sqrt{\zeta_\kappa(\bar{r}_t)}}\left[\langle -g_t, \exp_{x_t}^{-1}(x_\star)\rangle_{x_t} + \langle \Delta_t, \exp_{x_t}^{-1}(x_\star)\rangle_{x_t}\right] \tag{104}$$

$$= O\left(\frac{\bar{r}_{\tau_k}/\sqrt{\zeta_\kappa(\bar{r}_{\tau_k})}}{\bar{r}_{\tau_{k-1}}/\sqrt{\zeta_\kappa(\bar{r}_{\tau_{k-1}})}}(d_0 + \bar{r}_{\tau_k})\sqrt{\zeta_\kappa(d_0 + \bar{r}_{\tau_k})}\sqrt{S\theta_{T,\delta}^2}\sqrt{\sum_{t=0}^{T-1}[f(x_t) - f(x_\star)]}\right) \tag{105}$$

$$= O\left((d_0 + \bar{r}_{\tau_k})\sqrt{\zeta_\kappa(d_0 + \bar{r}_{\tau_k})}\sqrt{S\theta_{T,\delta}^2}\sqrt{\sum_{t=0}^{T-1}[f(x_t) - f(x_\star)]}\right). \tag{106}$$

Where final reduction holds since $d \mapsto \frac{d}{\zeta_\kappa(d)}$,and for any t,

$$\bar{r}_{t+1} \leq \bar{r}_t + d(x_{t+1}, x_t) = \bar{r}_t\left(1 + \frac{\|g_t\|_{x_t}}{\sqrt{G_t}}\right) \leq 2\bar{r}_t. \tag{107}$$

Now summing over k from 1 to K we have

$$\sum_{t=0}^{T-1}[f(x_t)-f(x_\star)] = \sum_{k=1}^{K}\sum_{t=\tau_{k-1}}^{\tau_k-1}[f(x_t)-f(x_\star)] \tag{108}$$

$$= O\left(\sum_{k=1}^{K}(d_0+\bar{r}_{\tau_k})\sqrt{\zeta_\kappa(d_0+\bar{r}_{\tau_k})}\sqrt{S\theta_{T,\delta}^2}\sqrt{\sum_{t=0}^{T-1}[f(x_t)-f(x_\star)]}\right) \tag{109}$$

$$\leq O\left(\sum_{k=1}^{K}(d_0+\bar{r}_{\tau_k})\sqrt{\zeta_\kappa\left(d_0+\sum_{k=1}^{K}\bar{r}_{\tau_k}\right)}\sqrt{S\theta_{T,\delta}^2}\sqrt{\sum_{t=0}^{T-1}[f(x_t)-f(x_\star)]}\right). \tag{110}$$

Where the final reduction holds since $d \mapsto \zeta_\kappa(d)$ is an increasing function. Now, recall that $K = O\left(\log_+ \frac{\bar{r}_T}{\epsilon}\right)$ and note that $\sum_{k=1}^{K}\bar{r}_{\tau_k} = O(\bar{r}_T)$ since $\frac{\bar{r}_{\tau_s}}{\bar{r}_{\tau_{K-1}}} \leq 2^{s-(K-1)}$ for all $s \leq K-1$. We then divide both sides through by $\sqrt{\sum_{t=0}^{T-1}[f(x_t)-f(x_\star)]}$, to yield

$$\sqrt{\sum_{t=0}^{T-1}[f(x_t)-f(x_\star)]} \leq O\left((d_0\bar{r}_T/\epsilon+\bar{r}_T)\sqrt{\zeta_\kappa\left(d_0+\bar{r}_T\right)}\sqrt{S\theta_{T,\delta}^2}\right), \tag{111}$$

squaring both sides, the proof is complete noting that via Lemma A.3 we deduce $f(\hat{x}_T) \leq \frac{1}{T}\sum_{t=0}^{T-1} f(x_t)$. □

D.4. Iterate Stability Bound

We introduce *Tamed Riemannian Distance over Gradients* (T-RDoG), a dampened version of RDoG (Algorithm 1) whose iterates are guaranteed to remain bounded with high probability. T-RDoG has the following step size scheme

$$\eta_t = \frac{\bar{r}_t}{\sqrt{\zeta_\kappa(\bar{r}_t)G_t'}}, \quad G_t' = 8^4\theta_{T,\delta}^2\log_+^2\left(\frac{(1+t)\bar{\ell}_t^2}{\bar{\ell}_0^2}\right)(G_{t-1}+16\bar{\ell}_t^2), \tag{112}$$

using $G_{-1} \coloneqq 0$ and recalling $\bar{\ell}_t \coloneqq \max_{s\leq t}\ell(x_s)$ for a function ℓ satisfying Assumption 3.3. To show iterate boundedness in the stochastic setting, we consider the stopping time

$$\mathcal{T}_{out} = \min\{t \geq 0 : \bar{r}_t > 3d_0\}, \tag{113}$$

so that the event $\{\bar{r}_T \leq 3d_0\}$ is the same as $\{\mathcal{T}_{out} > T\}$. We also define the following truncated step size sequence,

$$\tilde{\eta}_k \coloneqq \eta_k\mathbb{I}_{\{k<\mathcal{T}_{out}\}}. \tag{114}$$

Truncating as such allows us to handle the possibility that $\bar{r}_T$ exceeds $3d_0$. In particular, the following holds for $\{\tilde{\eta}_k\}$ but not for $\{\eta_k\}$.

Lemma D.8. *For all $t \leq T$, if Assumption 3.1, 3.2, and 3.3 hold, under the truncated T-RDoG step size sequence $\{\tilde{\eta}_t\}$, the iterates satisfy,*

$$\tilde{\eta}_t \in \sigma(\mathcal{G}(x_0),\ldots,\mathcal{G}(x_{t-1})), \tag{115}$$

$$|\tilde{\eta}_t\langle\gamma,\exp_{x_t}^{-1}(x_\star)\rangle_{x_t}| \leq \frac{6d_0^2}{8^2\sqrt{\zeta_\kappa(3d_0)}\theta_{T,\delta}} \text{ for } \gamma \in \{g_t, \operatorname{grad} f(x_t), \Delta_t\}, \tag{116}$$

$$\sum_{k=0}^{t}\tilde{\eta}_k^2\zeta_\kappa(d_k)\|g_k\|_{x_k}^2 \leq \frac{12d_0^2}{8^4\theta_{T,\delta}}, \text{ and} \tag{117}$$

$$\sum_{k=0}^{t}(\tilde{\eta}_k\langle g_k,\exp_{x_k}^{-1}(x_\star)\rangle_{x_k})^2 \leq \frac{3\cdot 4^3d_0^4}{8^4\theta_{T,\delta}}. \tag{118}$$

Proof. The first line holds directly by definition of the truncated T-RDoG iterates. For bound in the second line, note (recalling $\Delta_t = g_t - \operatorname{grad} f(x_t)$ we have $\|\Delta_t\|_{x_t} \le \|g_t\|_{x_t} + \|\operatorname{grad} f(x_t)\|_{x_t} \le 2\ell(x_t)$. Since $G'_t \ge 4^2 8^4 \ell^2(x_t)\theta^2_{T,\delta}$ for all t, the Cauchy-Schwartz inequality gives,

$$|\tilde{\eta}_t \langle \Delta_t, \exp_{x_t}^{-1}(x_\star)\rangle_{x_t}| \le \frac{\bar{r}_t}{\sqrt{\zeta_\kappa(\bar{r}_t)G'_t}}\|\Delta_t\|_{x_t} d_t \le \frac{1}{2\cdot 8^2\theta_{T,\delta}}\frac{\bar{r}_T}{\sqrt{\zeta_\kappa(\bar{r}_T)}} d_t \le \frac{6d_0^2}{8^2\sqrt{\zeta_\kappa(3d_0)}\theta_{T,\delta}}, \tag{119}$$

where we have used the fact that $d \mapsto \frac{d}{\sqrt{\zeta_\kappa(d)}}$ is an increasing function, and that $d_t \le d_0 + \bar{r}_t$ and $\bar{r}_t \le 3d_0$ (or else $\tilde{\eta}_t = 0$). Bounds for $\gamma \in \{g_t, \operatorname{grad} f(x_t)\}$ hold in a similar fashion.

Now for the third line, we have

$$\sum_{k=0}^{t} \tilde{\eta}_k^2 \zeta_\kappa(d_k)\|g_k\|^2_{x_k} \le \sum_{k=0}^{\mathcal{T}_{out}-1} \eta_k^2 \zeta_\kappa(d_k)\|g_k\|^2_{x_k} \tag{120}$$

$$= \sum_{k=0}^{\mathcal{T}_{out}-1} \frac{\bar{r}_k^2 \zeta_\kappa(d_k)(G_k - G_{k-1})}{\zeta_\kappa(\bar{r}_k)G'_k}. \tag{121}$$

Now $d \mapsto \zeta_\kappa(d)$ is increasing, thus $\zeta_\kappa(d_k) \le \zeta_\kappa(d_0 + \bar{r}_{\mathcal{T}_{out}-1})$ as $d_k \le \bar{d}_k \le \bar{d}_{\mathcal{T}_{out}-1} \le d_0 + \bar{r}_{\mathcal{T}_{out}-1}$. Additionally, since $d \mapsto \frac{d}{\zeta_\kappa(d)}$ is increasing, we have $\frac{\bar{r}_k^2}{\zeta_\kappa(\bar{r}_k)} = \bar{r}_k \frac{\bar{r}_k}{\zeta_\kappa(\bar{r}_k)} \le \bar{r}_{\mathcal{T}_{out}-1}\frac{\bar{r}_{\mathcal{T}_{out}-1}}{\zeta_\kappa(\bar{r}_{\mathcal{T}_{out}-1})} \le \bar{r}_{\mathcal{T}_{out}-1}\frac{(\bar{r}_{\mathcal{T}_{out}-1}+d_0)}{\zeta_\kappa(\bar{r}_{\mathcal{T}_{out}-1}+d_0)}$. Thus, we have

$$\sum_{k=0}^{\mathcal{T}_{out}-1} \frac{\bar{r}_k^2 \zeta_\kappa(d_k)(G_k - G_{k-1})}{\zeta_\kappa(\bar{r}_k)G'_k} \le \bar{r}_{\mathcal{T}_{out}-1}(\bar{r}_{\mathcal{T}_{out}-1} + d_0) \sum_{k=0}^{\mathcal{T}_{out}-1} \frac{G_k - G_{k-1}}{G'_k} \tag{122}$$

$$\overset{(i)}{\le} \frac{\bar{r}_{\mathcal{T}_{out}-1}(\bar{r}_{\mathcal{T}_{out}-1} + d_0)}{8^4\theta_{T,\delta}} \sum_{k=0}^{\mathcal{T}_{out}-1} \frac{G_k - G_{k-1}}{(G_k + \bar{\ell}_k^2)\log_+^2\left(\frac{G_k + \bar{\ell}_k^2}{\bar{\ell}_0^2}\right)} \tag{123}$$

$$\overset{(ii)}{\le} \frac{12 d_0^2}{8^4\theta_{T,\delta}}. \tag{124}$$

Where we have used in (i) that

$$G'_k \ge 8^4\theta_{T,\delta}(G_{k-1} + \|g_k\|^2_{x_k} + \ell_k^2)\log_+^2\left(\frac{\sum_{s=0}^k \bar{\ell}_s^2 + \bar{\ell}_k^2}{\bar{\ell}_0^2}\right) \ge 8^4\theta_{T,\delta}(G_k + \bar{\ell}_k^2)\log_+^2\left(\frac{G_k + \bar{\ell}_k^2}{\bar{\ell}_0^2}\right), \tag{125}$$

holding since $\|g_k\|_{x_k} \le \ell_k$. Additionally, in (ii) we have used Lemma A.8 with $a_k = G_k + \bar{\ell}_k^2$ and $\bar{r}_{\mathcal{T}_{out}-1} \le 3d_0$.

The final line holds from the previous noting that,

$$\sum_{k=0}^{t}(\tilde{\eta}_k\langle g_k, \exp_{x_k}^{-1}(x_\star)\rangle_{x_k})^2 \le \sum_{k=0}^{t}\tilde{\eta}_k^2\zeta_\kappa(d_k)\|g_k\|^2_{x_k} d_k^2 \le (4d_0)^2\sum_{k=0}^{t}\tilde{\eta}_k^2\zeta_\kappa(d_k)\|g_k\|^2_{x_k}, \tag{126}$$

where the first inequality follows from the Cauchy-Schwartz inequality and the second inequality holds from the fact that only terms with $k < \mathcal{T}_{out}$ contribute to the sum. □

Using the above lemma, we can establish the following concentration bound.

Lemma D.9. *If Assumption 3.1, 3.2, and 3.3 hold, under the truncated T-RDoG step size sequence $\{\tilde{\eta}_t\}$, the iterates satisfy,*

$$\mathbb{P}\left(\exists t \le T : \sum_{k=0}^{t-1}\tilde{\eta}_k\langle \Delta_k, \exp_{x_k}^{-1}(x_\star)\rangle_{x_k} > d_0^2\right) \le \delta. \tag{127}$$

Proof. Consider the filtration $\mathcal{F}_t = \sigma(\mathcal{G}(x_0)\ldots, \mathcal{G}(x_t))$ and define $X_t = \tilde{\eta}_t\langle \Delta_t, \exp_{x_t}^{-1}(x_\star)\rangle_{x_t}$ and $\hat{X}_t = -\tilde{\eta}_t\langle \operatorname{grad} f(x_t), \exp_{x_t}^{-1}(x_\star)\rangle_{x_t}$. Then we have that X_t is a martingale difference sequence adapted to $\mathcal{F}_t$ and $\hat{X}_t \in \mathcal{F}_{t-1}$.

Moreover, we have $\max\{|X_t|, |\hat{X}_t|\} \leq c$ almost surely for $c = \frac{24 d_0^2}{8^4 \theta_{T,\delta}}$. Substituting into Lemma A.9, we have

$$\mathbb{P}\left(\exists t \leq T : \left|\sum_{k=0}^{t-1} X_k\right| \geq 4\sqrt{\theta_{t,\delta} \sum_{k=0}^{t-1} (X_k - \hat{X}_k)^2 + c^2 \theta_{t,\delta}^2}\right) \leq \delta. \tag{128}$$

Noting that $X_t - \hat{X}_t = \tilde{\eta}_t \langle g_t, \exp_{x_t}^{-1}(x_\star)\rangle_{x_t}$ and substituting the definition of c and the bound gives for every $t < T$,

$$4\sqrt{\theta_{t,\delta} \sum_{k=0}^{t-1} (X_k - \hat{X}_k)^2 + c^2 \theta_{t,\delta}^2} \leq 4\sqrt{\theta_{t,\delta} \frac{3 \cdot 4^3 d_0^4}{8^4 \theta_{T,\delta}} + \left(\frac{6\theta_{t,\delta} d_0^2}{8^2 \sqrt{\zeta_\kappa(3d_0)}\theta_{T,\delta}}\right)^2} \leq d_0^2. \tag{129}$$

□

Finally, we show that the event defined in the previous lemma implies the desired distance bound.

Lemma D.10. *Suppose Assumption 3.1, 3.2, and 3.3 hold. If $\sum_{k=0}^{t-1} \tilde{\eta}_k \langle \Delta_k, \exp_{x_k}^{-1}(x_\star)\rangle_{x_k} \leq d_0^2$ for all $t \leq T$ then $\mathcal{T}_{out} > T$ i.e., $\bar{r}_t \leq 3d_0$.*

Proof. To condense notation, let $B_t := \max_{t' \leq t} \sum_{k=0}^{t'-1} \tilde{\eta}_k \langle \Delta_k, \exp_{x_k}^{-1}(x_\star)\rangle_{x_k}$, so the claim becomes $B_t \leq d_0^2$ implies $\mathcal{T}_{out} > t$ for all $t \leq T$. We prove the claim by induction on t. The basis of the induction is that $\mathcal{T}_{out} > 0$ always hold since $\bar{r}_0 = \epsilon \leq 3d_0$ by assumption. For the induction step, we assume that B_{t-1} implies $\mathcal{T}_{out} \geq t$ and show that $B_t \leq d_0^2$ implies $\mathcal{T}_{out} > t$. To that end, we use $\langle \operatorname{grad} f(x_t), \exp_{x_t}^{-1}(x_\star)\rangle_{x_t} \geq f(x_t) - f(x_\star) \geq 0$ to rearrange Lemma A.2 as

$$d_{k+1}^2 - d_k^2 \leq \eta_k^2 \zeta_\kappa(d_k) \|g_k\|_{x_k}^2 + 2\eta_k \langle \Delta_k, \exp_{x_k}^{-1}(x_\star)\rangle_{x_k} \tag{130}$$

for all k. Summing from $0 \leq k \leq t-1$, we have

$$\begin{aligned} d_t^2 - d_0^2 &\leq \sum_{k=0}^{t-1} \eta_k^2 \zeta_\kappa(d_k) \|g_k\|_{x_k}^2 + 2\sum_{k=0}^{t-1} \eta_k \langle \Delta_k, \exp_{x_k}^{-1}(x_\star)\rangle_{x_k} &(131)\\ &= \sum_{k=0}^{t-1} \tilde{\eta}_k^2 \zeta_\kappa(d_k) \|g_k\|_{x_k}^2 + 2\sum_{k=0}^{t-1} \tilde{\eta}_k \langle \Delta_k, \exp_{x_k}^{-1}(x_\star)\rangle_{x_k}. &(132)\end{aligned}$$

where the equality holds since $\mathcal{T}_{out} > t-1$ and therefore $\eta_k = \tilde{\eta}_k$ for all $0 \leq k \leq t-1$. Now, by previous lemma we have $\sum_{k=0}^{t-1} \tilde{\eta}_k^2 \zeta_\kappa(d_k) \|g_k\|_{x_k}^2 \leq \frac{12 d_0^2}{8^4 \theta_{T,\delta}} \leq d_0^2$. Moreover, by assumption we have $\sum_{k=0}^{t-1} \tilde{\eta}_k \langle \Delta_k, \exp_{x_k}^{-1}(x_\star)\rangle_{x_k} \leq B_t \leq d_0^2$, from which we conclude, $d_t^2 \leq 4d_0^2$ and hence $r_t \leq d_0 + d_t \leq 3d_0$. Finally, since $\bar{r}_t = \max\{\bar{r}_{t-1}, r_t\}$ and $\bar{r}_{t-1} \leq 3d_0$ by the induction assumption, we have that $\bar{r}_t \leq 3d_0$. □

Theorem D.11. *Suppose $\epsilon \leq 3d_0$ and Assumption 3.1, 3.2, and 3.3 hold. Then for any $\delta \in (0,1)$ and $t \in \mathbb{N}$, under the T-RDoG step size sequence $\{\eta_t\}$, the iterates satisfy $\mathbb{P}(\bar{r}_t > 3d_0) \leq \delta$.*

Proof. A consequence of combining the previous two lemmas. □

Corollary D.12. *Suppose that Assumption 3.1, 3.2, and 3.3 hold. For any $\delta \in (0, 1/2)$, $t \in \mathbb{N}$, consider T iterations of T-RDoG, with an initial step size of $\epsilon \leq 3d_0$. Then for $\tau \in \arg\max_{t \leq T} \sum_{s=0}^{t-1} \frac{\bar{r}_s/\sqrt{\zeta_\kappa(\bar{r}_s)}}{\bar{r}_t/\sqrt{\zeta_\kappa(\bar{r}_t)}}$ we have, with probability at least $1 - 2\delta$, that*

$$f(\tilde{x}_\tau) - f(x_\star) = O\left(c_{\delta,\epsilon,T} \frac{d_0 \sqrt{\zeta_\kappa(d_0)(G_{\tau-1}^2 + L_\star^2)}}{T}\right) = O\left(c_{\delta,\epsilon,T} \frac{d_0 \sqrt{\zeta_\kappa(d_0)} L_\star}{\sqrt{T}}\right). \tag{133}$$

where $L_\star := \max_{x \in \mathcal{M}: d(x,x_0) \leq 3d(x_\star,x_0)} \ell(x)$ and $c_{\delta,\epsilon,T} = \log_+(T \frac{d_0 L_\star}{f(x_0) - f(x_\star)}) \log_+(\frac{d_0}{\epsilon}) \log(\frac{\log_+(T)}{\delta})$.

Proof. Here we adapt Theorem D.3. Using that $\bar{r}_t \leq 3d_0$, we have $\zeta_\kappa(d_0 + \bar{r}_t) \leq \zeta_\kappa(4d_0) \leq \frac{\sqrt{\kappa}4d_0}{\tanh(\sqrt{\kappa}4d_0)} \leq \frac{\sqrt{\kappa}4d_0}{\tanh(\sqrt{\kappa}d_0)} = O(\zeta_\kappa(d_0))$ for $\kappa > 0$, otherwise $\zeta_\kappa(d_0 + \bar{r}_t) = 1 = \zeta_\kappa(d_0) = O(\zeta_\kappa(d_0))$ in the case $\kappa = 0$. Now, by Assumption 3.3 we have $\ell_0 \geq \|\operatorname{grad} f(x_0)\|_{x_0} \geq (f(x_0) - f(x_\star))/d_0$, while $\bar{r}_T \leq 3d_0$ gives $\bar{\ell}_T \leq L_\star$. Therefore, $\log_+\left(1 + \frac{T\bar{\ell}_T^2}{\ell_0^2}\right) = O\left(\log_+\left(T\frac{d_0 L_\star}{f(x_0) - f(x_\star)}\right)\right)$. □

D.5. Omitting Geometric Curvature Term Analysis

We analyze omitting the geometric curvature term from the denominator RDoG (Algorithm 1). Thus we consider step sizes of the form

$$\eta_t = \frac{\bar{r}_t}{\sqrt{\sum_{s=0}^t \|g_s\|_{x_s}^2}}. \tag{134}$$

We term this algorithm *Curvature Omitted Riemannian Distance over Gradients* (CO-RDoG).

We consider bounding the error of the weighted average sequence,

$$\tilde{x}_{t+1} = \exp_{\tilde{x}_t}\left(\frac{\bar{r}_t}{\sum_{s=0}^t \bar{r}_s} \exp_{\tilde{x}_t}^{-1}(x_t)\right), \quad \tilde{x}_1 = x_0.$$

For a geodesically convex function $f\colon \mathcal{M} \to \mathbb{R}$, we have by Jensens inequality (Lemma A.3) that $\tilde{x}_t$ satisfies,

$$f(\tilde{x}_t) - f(x_\star) \leq \frac{1}{\sum_{s=0}^{t-1} \bar{r}_s} \sum_{s=0}^{t-1} \bar{r}_s \langle -\operatorname{grad} f(x_s), \exp_{x_s}^{-1}(x_\star)\rangle_{x_s}. \tag{135}$$

Recalling g_s is the stochastic oracle evaluation, $\mathcal{G}(x_s)$, the numerator decomposes into two components:

$$\underbrace{\sum_{s=0}^{t-1} \bar{r}_s \langle -g_s, \exp_{x_s}^{-1}(x_\star)\rangle_{x_s}}_{\text{weighted regret}} + \underbrace{\sum_{s=0}^{t-1} \bar{r}_s \langle \Delta_s, \exp_{x_s}^{-1}(x_\star)\rangle_{x_s}}_{\text{noise}}, \tag{136}$$

with $\Delta_s \coloneqq g_s - \operatorname{grad} f(x_s)$.

We give deterministic bounds for the weighted regret (Lemma D.13) and high probability bounds for the noise term (Lemma D.14).

Lemma D.13. *Under Assumption 3.1 and 3.2, the iterates of CO-RDoG, satisfy*

$$\sum_{s=0}^{t-1} \bar{r}_s \langle -g_s, \exp_{x_s}^{-1}(x_\star)\rangle_{x_s} \leq \bar{r}_t \left(2\bar{d}_t + \bar{r}_t \zeta_\kappa(\bar{d}_t)\right) \sqrt{G_{t-1}}. \tag{137}$$

Proof. Applying Lemma A.2, we can bound the weighted average as

$$\sum_{s=0}^{t-1} \bar{r}_s \langle -g_s, \exp_{x_s}^{-1}(x_\star)\rangle_{x_s} \leq \underbrace{\frac{1}{2}\sum_{s=0}^{t-1} \frac{\bar{r}_s}{\eta_s}\left(d_s^2 - d_{s+1}^2\right)}_{(A)} + \underbrace{\frac{1}{2}\sum_{s=0}^{t-1} \bar{r}_s \eta_s \zeta_\kappa(d_s)\|g_s\|_{x_s}^2}_{(B)}. \tag{138}$$

We bound the terms (A) and (B) in turn, beginning with the former:

$$(A) = \sum_{s=0}^{t-1} \sqrt{G_s}\left(d_s^2 - d_{s+1}^2\right) = d_0^2\sqrt{G_0} - d_t^2\sqrt{G_{t-1}} + \sum_{s=1}^{t-1} d_s^2\left(\sqrt{G_s} - \sqrt{G_{s-1}}\right) \tag{139}$$

$$\overset{(i)}{\leq} \bar{d}_t^2\sqrt{G_0} - d_t^2\sqrt{G_{t-1}} + \bar{d}_t^2 \sum_{s=1}^{t-1}\left(\sqrt{G_s} - \sqrt{G_{s-1}}\right) = \sqrt{G}_{t-1}(\bar{d}_t^2 - d_t^2) \overset{(ii)}{\leq} 4\bar{r}_t\bar{d}_t\sqrt{G_{t-1}}. \tag{140}$$

Inequality (i) uses $d_s \leq \bar{d}_t$ and that G_t is nondecreasing. Inequality (ii) use that for $k \in \arg\max_{s \leq t} d_s$, we have $\bar{d}_t^2 - d_t^2 = d_k^2 - d_t^2 = (d_k - d_t)(d_k + d_t) \leq d(x_k, x_t)(d_k + d_t) \leq (\bar{r}_k + \bar{r}_t)(d_k + d_t) \leq 4\bar{r}_t \bar{d}_t$.

Bounding the second term (B), using $d \mapsto \zeta_\kappa(d)$ is an increasing function, we have:

$$(B) = \sum_{s=0}^{t-1} \frac{\bar{r}_s^2 \zeta_\kappa(d_s) \|g_s\|_{x_s}^2}{\sqrt{G_s}} \leq \sum_{s=0}^{t-1} \frac{\bar{r}_s^2 \zeta_\kappa(\bar{d}_s) \|g_s\|_{x_s}^2}{\sqrt{G_s}} \leq \bar{r}_t^2 \zeta_\kappa(\bar{d}_t) \sum_{s=0}^{t-1} \frac{\|g_s\|_{x_s}^2}{\sqrt{G_s}} \leq 2\bar{r}_t^2 \zeta_\kappa(\bar{d}_t) \sqrt{G_{t-1}}. \tag{141}$$

Thus, combining (A) and (B) together, gives the result. □

Lemma D.14. *For all $\delta \in (0,1), T \in \mathbb{N}$ and $L > 0$, if Assumption 3.1, 3.2, and 3.3 hold, the iterates of CO-RDoG satisfy*

$$\mathbb{P}\left(\exists t \leq T : \left| \sum_{s=0}^{t-1} \bar{r}_s \langle \Delta_s, \exp_{x_s}^{-1}(x_\star) \rangle_{x_s} \right| \geq b_t \right) \leq \delta + \mathbb{P}(\bar{\ell}_T > L), \tag{142}$$

where $b_t = 8\bar{r}_{t-1}\bar{d}_{t-1}\sqrt{\theta_{t,\delta} G_{t-1} + \theta_{t,\delta}^2 L^2}$ and $\bar{\ell}_T \coloneqq \max_{s \leq T} \ell(x_s)$.

Proof. For $1 \leq s \leq T$ define the random variables

$$Y_s \coloneqq \bar{r}_{s-1}\bar{d}_{s-1}, \quad X_s \coloneqq \left\langle \Delta_{s-1}, \frac{\exp_{x_{s-1}}^{-1}(x_\star)}{\bar{d}_{s-1}} \right\rangle_{x_{s-1}}, \quad \hat{X}_s \coloneqq \left\langle -\operatorname{grad} f(x_{s-1}), \frac{\exp_{x_{s-1}}^{-1}(x_\star)}{\bar{d}_{s-1}} \right\rangle_{x_{s-1}}. \tag{143}$$

By the Cauchy-Schwartz inequality and Assumption 3.3 we have each $|X_s| \leq \ell(x)$, and each $|\hat{X}_s| \leq \ell(x)$ with probability 1. Moreover, for any $t \leq T$, we have

$$\sum_{s=1}^{t} Y_s X_s = \sum_{s=0}^{t-1} \bar{r}_s \langle \Delta_s, \exp_{x_s}^{-1}(x_\star) \rangle_{x_s}. \tag{144}$$

Therefore, applying Lemma A.9 yields the result as required. □

Combining the above results, we obtain the following.

Theorem D.15. *For all $\delta \in (0,1)$ and $L > 0$, if Assumption 3.1, 3.2, and 3.3 hold, then with probability at least $1 - \delta - \mathbb{P}(\bar{\ell}_T > L)$, for all $t \leq T$, the optimality gap on the weighted iterates $f(\tilde{x}_t) - f(x_\star)$ of CO-RDoG satisfy*

$$O\left(\frac{(d_0 + \bar{r}_t)\zeta_\kappa(d_0 + \bar{r}_t)\sqrt{G_{t-1} + \theta_{t,\delta} G_{t-1} + \theta_{t,\delta}^2 L^2}}{\sum_{s=0}^{t-1} \bar{r}_s / \bar{r}_t} \right) \tag{145}$$

with probability at least $1 - \delta - \mathbb{P}(\bar{\ell}_T > L)$.

Proof. Combining Lemma D.13 and Lemma D.14 and utilizing $\bar{d}_t \leq d_0 + \bar{r}_t$ and that $d \mapsto \zeta_\kappa(d)$ is an increasing function yields the result as required. □

Thus in comparison to standard RDoG, we pay an additional cost of $O\left(\sqrt{\zeta_\kappa(d_0 + \bar{r}_t)}\right)$ for omitting the geometric curvature term with CO-RDoG.

We then have a useful result when the manifold is bounded but its exact diameter is unknown.

Corollary D.16. *Under Assumption 3.1, 3.2, and 3.3, for any $D \geq d_0$ let $L_D \coloneqq \max_{x \in \mathcal{M} : d(x, x_0) \leq D} \ell(x)$. Then, for all $\delta \in (0,1)$ and for $\tau \in \arg\max_{t \leq T} \sum_{s=0}^{t-1} \frac{\bar{r}_s / \sqrt{\zeta_\kappa(\bar{r}_s)}}{\bar{r}_t / \sqrt{\zeta_\kappa(\bar{r}_t)}}$, with probability at least $1 - \delta - \mathbb{P}(\bar{\ell}_T > L)$, iterates of CO-RDoG satisfy the optimality gap bound*

$$f(\tilde{x}_\tau) - f(x_\star) = O\left(\frac{D\zeta_\kappa(D)\sqrt{G_{\tau-1}\theta_{\tau,\delta} + L_D^2 \theta_{\tau,\delta}^2}}{T} \log_+(D/\epsilon) \right). \tag{146}$$

Proof. Apply Lemma A.6 to the denominator term of Theorem D.15. □

Thus in comparison to standard RDoG, we pay an additional cost of $O\left(\sqrt{\zeta_\kappa(D)}\right)$ for omitting the curvature term with CO-RDoG.

E. RDoWG Theoretical Analysis

E.1. Overview

In this section, we analyze RDoWG (Algorithm 2). Thus we consider RSGD with step sizes given by,

$$\eta_t = \frac{\bar{r}_t^2}{\zeta_\kappa(\bar{r}_t)\sqrt{v_t}}, \quad v_t = v_{t-1} + \frac{\bar{r}_t^2}{\zeta_\kappa(\bar{r}_t)}\|g_t\|_{x_t}^2, \quad v_{-1} = 0. \tag{147}$$

We consider the bounding the error of the weighted average sequence,

$$\tilde{x}_{t+1} = \exp_{\tilde{x}_t}\left(\frac{\bar{r}_t^2/\zeta_\kappa(\bar{r}_t)}{\sum_{s=0}^t \bar{r}_s^2/\zeta_\kappa(\bar{r}_s)} \exp_{\tilde{x}_t}^{-1}(x_t)\right), \quad \tilde{x}_1 = x_0.$$

For a geodesically convex function $f\colon \mathcal{M} \to \mathbb{R}$, we have by Lemma A.3 that $\tilde{x}_t$ satisfies,

$$f(\tilde{x}_t) - f(x_\star) \le \frac{1}{\sum_{s=0}^{t-1}(\bar{r}_s^2/\zeta_\kappa(\bar{r}_s))} \sum_{s=0}^{t-1}(\bar{r}_s^2/\zeta_\kappa(\bar{r}_s))\langle -\operatorname{grad} f(x_s), \exp_{x_s}^{-1}(x_\star)\rangle_{x_s}. \tag{148}$$

Recalling that g_s represents the stochastic oracle evaluation at x_s, denoted as $\mathcal{G}(x_s)$, we can decompose the numerator into two components:

$$\underbrace{\sum_{s=0}^{t-1}(\bar{r}_s^2/\zeta_\kappa(\bar{r}_s))\langle -g_s, \exp_{x_s}^{-1}(x_\star)\rangle_{x_s}}_{\text{weighted regret}} + \underbrace{\sum_{s=0}^{t-1}(\bar{r}_s^2/\zeta_\kappa(\bar{r}_s))\langle \Delta_s, \exp_{x_s}^{-1}(x_\star)\rangle_{x_s}}_{\text{noise}}, \tag{149}$$

with $\Delta_s \coloneqq g_s - \operatorname{grad} f(x_s)$.

E.2. Supporting Analysis

Our first result gives deterministic bounds for the weighted regret (Lemma E.2).

Lemma E.1. *Under Assumption 3.1 and 3.2, we have that the iterates of RDoWG (Algorithm 2) satisfy*

$$\sum_{s=0}^{t-1}(\bar{r}_s^2/\zeta_\kappa(\bar{r}_s))\langle -g_s, \exp_{x_s}^{-1}(x_\star)\rangle_{x_s} \le \bar{r}_t\left(2\bar{d}_t + \frac{\bar{r}_t}{\zeta_\kappa(\bar{r}_t)}\zeta_\kappa(\bar{d}_t)\right)\sqrt{v_{t-1}}. \tag{150}$$

Proof. Follow same argument as Lemma D.1 but with weights $\frac{\bar{r}_s^2}{\zeta_\kappa(\bar{r}_s)}$ replacing $\frac{\bar{r}_s}{\sqrt{\zeta_\kappa(\bar{r}_s)}}$ and weighted gradient sum v_s replacing the standard gradient sum G_s. □

E.3. Non-Smooth Analysis

We give deterministic bounds for the weighted regret (Lemma E.2) and high probability bounds for the noise term (Lemma E.3) for the non-smooth setting.

Lemma E.2. *Under Assumption 3.1 and 3.2, we have that the iterates of RDoWG (Algorithm 2) satisfy*

$$\sum_{s=0}^{t-1}(\bar{r}_s^2/\zeta_\kappa(\bar{r}_s))\langle -g_s, \exp_{x_s}^{-1}(x_\star)\rangle_{x_s} \le \frac{\bar{r}_t^2}{\sqrt{\zeta_\kappa(\bar{r}_t)}}\left(2\bar{d}_t + \frac{\bar{r}_t}{\zeta_\kappa(\bar{r}_t)}\zeta_\kappa(\bar{d}_t)\right)\sqrt{G_{t-1}}. \tag{151}$$

Proof. Using the bound of E.2 and that $\sqrt{v_{t-1}} \leq \frac{\bar{r}_{t-1}}{\sqrt{\zeta_\kappa(\bar{r}_{t-1})}}\sqrt{G_{t-1}} \leq \frac{\bar{r}_t}{\sqrt{\zeta_\kappa(\bar{r}_t)}}\sqrt{G_{t-1}}$ gives the result. □

Lemma E.3. *For all $\delta \in (0,1), T \in \mathbb{N}$ and $L > 0$, if Assumption 3.1, 3.2, and 3.3 hold, the iterates of RDoWG (Algorithm 2) satisfy*

$$\mathbb{P}\left(\exists t \leq T : \left|\sum_{s=0}^{t-1} \frac{\bar{r}_s^2}{\zeta_\kappa(\bar{r}_s)} \langle \Delta_s, \exp_{x_s}^{-1}(x_\star)\rangle_{x_s}\right| \geq b_t\right) \leq \delta + \mathbb{P}(\bar{\ell}_T > L), \tag{152}$$

where $b_t = 8\frac{\bar{r}_{t-1}^2}{\zeta_\kappa(\bar{r}_{t-1})}\bar{d}_{t-1}\sqrt{\theta_{t,\delta}G_{t-1} + \theta_{t,\delta}^2 L^2}$ and $\bar{\ell}_T := \max_{s\leq T} \ell(x_s)$.

Proof. Following the argument of Lemma D.2. □

Combining the above results, we obtain the following.

Theorem E.4. *For all $\delta \in (0,1)$ and $L > 0$, if Assumption 3.1, 3.2, and 3.3 hold, then with probability at least $1 - \delta - \mathbb{P}(\bar{\ell}_T > L)$, for all $t \leq T$, the optimality gap on the weighted iterates $f(\tilde{x}_t) - f(x_\star)$ of RDoWG (Algorithm 2) satisfy*

$$O\left(\frac{(d_0 + \bar{r}_t)\sqrt{\zeta_\kappa(d_0 + \bar{r}_t)}\sqrt{G_{t-1} + \theta_{t,\delta}G_{t-1} + \theta_{t,\delta}^2 L^2}}{\sum_{s=0}^{t-1} \frac{\bar{r}_s^2/\zeta_\kappa(\bar{r}_s)}{\bar{r}_t^2/\zeta_\kappa(\bar{r}_t)}}\right). \tag{153}$$

Proof. Using Lemma E.2 and Lemma E.3 we have

$$f(\tilde{x}_t) - f(x_\star) \leq \frac{\left(2\bar{d}_t\sqrt{\zeta_\kappa(\bar{r}_t)} + \frac{\bar{r}_t}{\sqrt{\zeta_\kappa(\bar{r}_t)}}\zeta_\kappa(\bar{d}_t)\right)\sqrt{G_{t-1}} + 8\bar{d}_t\sqrt{\theta_{t,\delta}G_{t-1} + \theta_{t,\delta}^2 L^2}}{\sum_{s=0}^{t-1} \frac{\bar{r}_s^2/\zeta_\kappa(\bar{r}_s)}{\bar{r}_t^2/\zeta_\kappa(\bar{r}_t)}}. \tag{154}$$

Now using the fact $\bar{d}_t \leq d_0 + \bar{r}_t$ and that $d \mapsto \zeta_\kappa(d)$ and $d \mapsto \frac{d}{\sqrt{\zeta_\kappa(d)}}$ are increasing functions gives the result. □

Corollary E.5. *Suppose Assumption 3.1, 3.2, and 3.3 hold, and for any $D \geq d_0$ let $L_D := \max_{x\in\mathcal{M}:d(x,x_0)\leq D} \ell(x)$. Then, for all $\delta \in (0,1)$ and for $\tau \in \arg\max_{t\leq T} \sum_{s=0}^{t-1} \frac{\bar{r}_s/\sqrt{\zeta_\kappa(\bar{r}_s)}}{\bar{r}_t/\sqrt{\zeta_\kappa(\bar{r}_t)}}$, with probability at least $1 - \delta - \mathbb{P}(\bar{\ell}_T > L)$, iterates of RDoWG (Algorithm 2) satisfy the optimality gap bound*

$$f(\tilde{x}_\tau) - f(x_\star) = O\left(\frac{D\sqrt{\zeta_\kappa(D)}\sqrt{G_{\tau-1}\theta_{\tau,\delta} + L_D^2\theta_{\tau,\delta}^2}}{T}\log_+\left(\frac{D/\sqrt{\zeta_\kappa(D)}}{\epsilon/\sqrt{\zeta_\kappa(\epsilon)}}\right)\right). \tag{155}$$

Proof. Apply Lemma A.6 to the denominator term of Theorem E.4. □

E.4. Smooth Analysis

Lemma E.6. *Suppose f is S-smooth and assume Assumption 3.1 and 3.2 hold. Then we have that the iterates of RDoWG (Algorithm 2) satisfy*

$$\sum_{s=0}^{t-1}(\bar{r}_s^2/\zeta_\kappa(\bar{r}_s))\langle -g_s, \exp_{x_s}^{-1}(x_\star)\rangle_{x_s} \leq \bar{r}_t\left(2\bar{d}_t + \frac{\bar{r}_t}{\zeta_\kappa(\bar{r}_s)}\zeta_\kappa(\bar{d}_t)\right)\sqrt{2S\sum_{s=0}^{t-1}\frac{\bar{r}_s^2}{\zeta_\kappa(\bar{r}_s)}(f(x_s) - f(x_\star))}. \tag{156}$$

Proof. By smoothness we can use Lemma A.4 to deduce $\|\operatorname{grad} f(x)\|_x^2 \leq 2S(f(x) - f(x_\star))$ for all $x \in \mathcal{M}$. Therefore

$$v_t = \sum_{s=0}^{t-1}\frac{\bar{r}_s^2}{\zeta_\kappa(\bar{r}_s)}\|\operatorname{grad} f(x_s)\|_{x_s}^2 \leq 2S\sum_{s=0}^{t-1}\frac{\bar{r}_s^2}{\zeta_\kappa(\bar{r}_s)}(f(x_s) - f(x_\star)). \tag{157}$$

Taking square roots and substituting this into Lemma E.2 gives the result. □

Lemma E.7. *Suppose Assumption 3.1, 3.2, and 3.4 hold. Then for all $\delta \in (0,1), T \in \mathbb{N}$ and $S > 0$, Then we have that the iterates of RDoWG (Algorithm 2) satisfy*

$$\mathbb{P}\left(\exists t \leq T : \left|\sum_{s=0}^{t-1} \frac{\bar{r}_s^2}{\zeta_\kappa(\bar{r}_s)} \langle \Delta_s, \exp_{x_s}^{-1}(x_\star)\rangle_{x_s}\right| \geq b_t\right) \leq \delta + \mathbb{P}(\bar{s}_T > S), \tag{158}$$

where $b_t = 8\frac{\bar{r}_{t-1}}{\sqrt{\zeta_\kappa(\bar{r}_{t-1})}}\bar{d}_{t-1}\sqrt{2S\theta_{t,\delta} + 8S\theta_{t,\delta}^2}\sqrt{\sum_{s=0}^{t-1}\frac{\bar{r}_s^2}{\zeta_\kappa(\bar{r}_s)}[f(x_s) - f(x_\star)]}$ *and* $\bar{s}_T \coloneqq \max_{t\leq T} s(x_t)$.

Proof. Define for $1 \leq s \leq T$ the following random variables as

$$Y_s \coloneqq \frac{\bar{r}_{s-1}}{\sqrt{\zeta_\kappa(\bar{r}_{s-1})}}\bar{d}_{s-1}, \quad X_s \coloneqq \frac{\bar{r}_{s-1}}{\sqrt{\zeta_\kappa(\bar{r}_{s-1})}}\left\langle \Delta_{s-1}, \frac{\exp_{x_{s-1}}^{-1}(x_\star)}{\bar{d}_{s-1}}\right\rangle_{x_{s-1}}, \tag{159}$$

$$\hat{X}_s \coloneqq \frac{\bar{r}_{s-1}}{\sqrt{\zeta_\kappa(\bar{r}_{s-1})}}\left\langle -\operatorname{grad} f(x_{s-1}), \frac{\exp_{x_{s-1}}^{-1}(x_\star)}{\bar{d}_{s-1}}\right\rangle_{x_{s-1}}, \tag{160}$$

and follow similar argument to Lemma D.6. □

Combining the above results, we obtain the following.

Theorem E.8. *Suppose Assumption 3.1, 3.2, and 3.4 hold. Then for all $\delta \in (0,1)$ and $S > 0$, with probability at least $1 - \delta - \mathbb{P}(\bar{s}_T > S)$, for all $t \leq T$, the optimality gap on the weighted iterates $f(\tilde{x}_t) - f(x_\star)$ of RDoWG (Algorithm 2) satisfy*

$$O\left(\frac{(d_0 + \bar{r}_t)^2\zeta_\kappa(d_0 + \bar{r}_t)(S\theta_{t,\delta}^2)}{\sum_{s=0}^{t-1}\frac{\bar{r}_s^2/\zeta_\kappa(\bar{r}_s)}{\bar{r}_t^2/\zeta_\kappa(\bar{r}_t)}}\right). \tag{161}$$

Proof. Using Lemma E.20 and Lemma E.21 above, we have with the relevant probabilistic conditions,

$$\sum_{s=0}^{t-1}\frac{\bar{r}_s^2}{\zeta_\kappa(\bar{r}_s)}[f(x_s) - f(x_\star)] \leq \left(\sqrt{2S}\bar{r}_t\left(2\bar{d}_t + \frac{\bar{r}_t}{\zeta_\kappa(\bar{r}_t)}\zeta_\kappa(\bar{d}_t)\right) + 8\frac{\bar{r}_t}{\sqrt{\zeta_\kappa(\bar{r}_t)}}\bar{d}_t\sqrt{2S\theta_{t,\delta} + 8S\theta_{t,\delta}^2}\right) \times \sqrt{\sum_{s=0}^{t-1}\frac{\bar{r}_s^2}{\zeta_\kappa(\bar{r}_s)}[f(x_s) - f(x_\star)]}.$$

Now if $f(x_s) - f(x_\star) = 0$ for some iterate, then the statement is trivial. Otherwise diving by sides by the square root term, we have

$$\sqrt{\sum_{s=0}^{t-1}\frac{\bar{r}_s^2}{\zeta_\kappa(\bar{r}_s)}[f(x_s) - f(x_\star)]} \leq \left(\sqrt{2S}\bar{r}_t\left(2\bar{d}_t + \frac{\bar{r}_t}{\zeta_\kappa(\bar{r}_t)}\zeta_\kappa(\bar{d}_t)\right) + 8\frac{\bar{r}_t}{\sqrt{\zeta_\kappa(\bar{r}_t)}}\bar{d}_t\sqrt{2S\theta_{t,\delta} + 8S\theta_{t,\delta}^2}\right). \tag{162}$$

We square both sides and divide through by $\frac{\bar{r}_s^2}{\zeta_\kappa(\bar{r}_s)}$. Finally using the fact, $\bar{d}_t \leq d_0 + \bar{r}_t$, in the above bound gives the result since $d \mapsto \zeta_\kappa(d)$ and $d \mapsto \frac{d}{\sqrt{\zeta_\kappa(d)}}$ are increasing functions. □

We then have a useful result when the manifold is bounded but its exact diameter is unknown.

Corollary E.9. *Under Assumption 3.1, 3.2, and 3.4 hold, for any $D \geq d_0$ let $S_D \coloneqq \max_{x\in\mathcal{M}:d(x,x_0)\leq D} s(x)$. Then, for all $\delta \in (0,1)$ and for $\tau \in \arg\max_{t\leq T}\sum_{s=0}^{t-1}\frac{\bar{r}_s^2/\zeta_\kappa(\bar{r}_s)}{\bar{r}_t^2/\zeta_\kappa(\bar{r}_t)}$, with probability at least $1 - \delta - \mathbb{P}(\bar{s}_T > S)$, iterates of Algorithm 1 satisfy the optimally gap on the weighted iterates $f(\tilde{x}_\tau) - f(x_\star)$ of RDoWG (Algorithm 2) satisfy*

$$O\left(\frac{D^2\zeta_\kappa(D)S_D\theta_{\tau,\delta}^2}{T}\log_+\left(\frac{D/\sqrt{\zeta_\kappa(D)}}{\epsilon/\sqrt{\zeta_\kappa(\epsilon)}}\right)\right). \tag{163}$$

Proof. Apply Lemma A.6 to the denominator term of Corollary E.23. □

E.5. Iterate Stability Bound

We introduce *Tamed Riemannian Distance over Weighted Gradients* (T-RDoWG), a dampened version of RDoWG (Algorithm 2) whose iterates are guaranteed to remain bounded with high probability. T-RDoWG has the following step size scheme

$$v_t = v_{t-1} + \frac{\bar{r}_t^2}{\zeta_\kappa(\bar{r}_t)}\|g_t\|_{x_t}^2, \quad v_{-1} = 0, \tag{164}$$

$$\eta_t = \frac{\bar{r}_t^2}{\zeta_\kappa(\bar{r}_t)\sqrt{v_t'}}, \quad v_t' = 8^4\theta_{T,\delta}^2 \log_+^2\left(\frac{(1+t)\bar{r}_t^2\bar{\ell}_t^2/\zeta_\kappa(\bar{r}_t)}{\bar{r}_0^2\bar{\ell}_0^2/\zeta_\kappa(\bar{r}_0)}\right)(v_{t-1} + 16\frac{\bar{r}_t^2}{\zeta_\kappa(\bar{r}_t)}\bar{\ell}_t^2). \tag{165}$$

To show iterate boundedness in the stochastic setting, we consider the stopping time

$$\mathcal{T}_{out} = \min\{t \geq 0 : \bar{r}_t > 3d_0\}, \tag{166}$$

so that the event $\{\bar{r}_T \leq 3d_0\}$ is the same as $\{\mathcal{T}_{out} > T\}$. We also define the following truncated step size sequence,

$$\tilde{\eta}_k \coloneqq \eta_k \mathbb{I}_{\{k < \mathcal{T}_{out}\}}. \tag{167}$$

Truncating as such allows us to handle the possibility that $\bar{r}_T$ exceeds $3d_0$. In particular, the following holds for $\{\tilde{\eta}_k\}$ but not for $\{\eta_k\}$.

Lemma E.10. *For all $t \leq T$, if Assumption 3.1, 3.2, and 3.3 hold, under the truncated T-RDoWG step size sequence $\{\tilde{\eta}_t\}$, the iterates satisfy*

$$\tilde{\eta}_t \in \sigma(\mathcal{G}(x_0), \ldots, \mathcal{G}(x_{t-1})), \tag{168}$$

$$|\tilde{\eta}_t\langle\gamma, \exp_{x_t}^{-1}(x_\star)\rangle_{x_t}| \leq \frac{6d_0^2}{8^2\sqrt{\zeta_\kappa(3d_0)}\theta_{T,\delta}} \text{ for } \gamma \in \{g_t, \operatorname{grad} f(x_t), \Delta_t\}, \tag{169}$$

$$\sum_{k=0}^{t} \tilde{\eta}_k^2\zeta_\kappa(d_k)\|g_k\|_{x_k}^2 \leq \frac{12d_0^2}{8^4\theta_{T,\delta}}, \text{ and} \tag{170}$$

$$\sum_{k=0}^{t}(\tilde{\eta}_k\langle g_k, \exp_{x_k}^{-1}(x_\star)\rangle_{x_k})^2 \leq \frac{3\cdot 4^3 d_0^4}{8^4\theta_{T,\delta}}. \tag{171}$$

Proof. The first line holds directly by definition of the truncated T-RDoWG iterates. For bound in the second line, note (recalling $\Delta_t = g_t - \operatorname{grad} f(x_t)$ we have $\|\Delta_t\|_{x_t} \leq \|g_t\|_{x_t} + \|\operatorname{grad} f(x_t)\|_{x_t} \leq 2\ell(x_t)$. Since $v_t' \geq 4^2 8^4 \frac{\bar{r}_t^2}{\zeta_\kappa(\bar{r}_t)}\ell^2(x_t)\theta_{T,\delta}^2$ for all t, the Cauchy-Schwartz inequality gives,

$$|\tilde{\eta}_t\langle\Delta_t, \exp_{x_t}^{-1}(x_\star)\rangle_{x_t}| \leq \frac{\bar{r}_t^2}{\zeta_\kappa(\bar{r}_t)\sqrt{v_t'}}\|\Delta_t\|_{x_t}d_t \leq \frac{1}{2\cdot 8^2\theta_{T,\delta}}\frac{\bar{r}_T}{\sqrt{\zeta_\kappa(\bar{r}_T)}}d_t \leq \frac{6d_0^2}{8^2\sqrt{\zeta_\kappa(3d_0)}\theta_{T,\delta}}, \tag{172}$$

where we have used the fact that $d \mapsto \frac{d}{\sqrt{\zeta_\kappa(d)}}$ is an increasing function, and that $d_t \leq d_0 + \bar{r}_t$ and $\bar{r}_t \leq 3d_0$ (or else $\tilde{\eta}_t = 0$). Bounds for $\gamma \in \{g_t, \operatorname{grad} f(x_t)\}$ hold in a similar fashion.

Now for the third line, we have

$$\sum_{k=0}^{t} \tilde{\eta}_k^2\zeta_\kappa(d_k)\|g_k\|_{x_k}^2 \leq \sum_{k=0}^{\mathcal{T}_{out}-1} \eta_k^2\zeta_\kappa(d_k)\|g_k\|_{x_k}^2 \tag{173}$$

$$= \sum_{k=0}^{\mathcal{T}_{out}-1} \frac{\bar{r}_k^4\zeta_\kappa(d_k)\|g_k\|_{x_k}^2}{\zeta_\kappa(\bar{r}_k)^2 v_k'} \tag{174}$$

$$= \sum_{k=0}^{\mathcal{T}_{out}-1} \frac{\bar{r}_k^2\zeta_\kappa(d_k)(v_k - v_{k-1})}{\zeta_\kappa(\bar{r}_k)v_k'}. \tag{175}$$

Now $d \mapsto \zeta_\kappa(d)$ is increasing, thus $\zeta_\kappa(d_k) \leq \zeta_\kappa(d_0 + \bar{r}_{\mathcal{T}_{out}-1})$ as $d_k \leq \bar{d}_k \leq \bar{d}_{\mathcal{T}_{out}-1} \leq d_0 + \bar{r}_{\mathcal{T}_{out}-1}$. Additionally, since $d \mapsto \frac{d}{\zeta_\kappa(d)}$ is increasing, we have $\frac{\bar{r}_k^2}{\zeta_\kappa(\bar{r}_k)} = \bar{r}_k \frac{\bar{r}_k}{\zeta_\kappa(\bar{r}_k)} \leq \bar{r}_{\mathcal{T}_{out}-1} \frac{\bar{r}_{\mathcal{T}_{out}-1}}{\zeta_\kappa(\bar{r}_{\mathcal{T}_{out}-1})} \leq \bar{r}_{\mathcal{T}_{out}-1} \frac{(\bar{r}_{\mathcal{T}_{out}-1}+d_0)}{\zeta_\kappa(\bar{r}_{\mathcal{T}_{out}-1}+d_0)}$. Thus, we have

$$\sum_{k=0}^{\mathcal{T}_{out}-1} \frac{\bar{r}_k^2 \zeta_\kappa(d_k)(G_k - G_{k-1})}{\zeta_\kappa(\bar{r}_k) G_k'} \leq \bar{r}_{\mathcal{T}_{out}-1}(\bar{r}_{\mathcal{T}_{out}-1} + d_0) \sum_{k=0}^{\mathcal{T}_{out}-1} \frac{v_k - v_{k-1}}{v_k'} \tag{176}$$

$$\overset{(i)}{\leq} \frac{\bar{r}_{\mathcal{T}_{out}-1}(\bar{r}_{\mathcal{T}_{out}-1} + d_0)}{8^4 \theta_{T,\delta}} \sum_{k=0}^{\mathcal{T}_{out}-1} \frac{v_k - v_{k-1}}{(v_k + \frac{\bar{r}_k^2}{\zeta_\kappa(\bar{r}_k)} \bar{\ell}_k^2) \log_+^2 \left(\frac{v_k + \frac{\bar{r}_k^2}{\zeta_\kappa(\bar{r}_k)} \bar{\ell}_k^2}{\frac{\bar{r}_0^2}{\zeta_\kappa(\bar{r}_0)} \bar{\ell}_0^2} \right)} \tag{177}$$

$$\overset{(ii)}{\leq} \frac{12 d_0^2}{8^4 \theta_{T,\delta}}. \tag{178}$$

Where we have used in (i) that

$$v_k' \geq 8^4 \theta_{T,\delta} \left(v_{k-1} + \frac{\bar{r}_k^2}{\zeta_\kappa(\bar{r}_k)} \|g_k\|_{x_k}^2 + \frac{\bar{r}_k^2}{\zeta_\kappa(\bar{r}_k)} \ell_k^2 \right) \log_+^2 \left(\frac{\sum_{s=0}^k \frac{\bar{r}_s^2}{\zeta_\kappa(\bar{r}_s)} \bar{\ell}_s^2 + \frac{\bar{r}_k^2}{\zeta_\kappa(\bar{r}_k)} \bar{\ell}_k^2}{\frac{\bar{r}_0^2}{\zeta_\kappa(\bar{r}_0)} \bar{\ell}_0^2} \right) \tag{179}$$

$$\geq 8^4 \theta_{T,\delta} \left(v_k + \frac{\bar{r}_k^2}{\zeta_\kappa(\bar{r}_k)} \bar{\ell}_k^2 \right) \log_+^2 \left(\frac{v_k + \frac{\bar{r}_k^2}{\zeta_\kappa(\bar{r}_k)} \bar{\ell}_k^2}{\frac{\bar{r}_0^2}{\zeta_\kappa(\bar{r}_0)} \bar{\ell}_0^2} \right), \tag{180}$$

holding since $\|g_k\|_{x_k} \leq \ell_k$. Additionally, in (ii) we have used Lemma A.8 with $a_k = v_k + \frac{\bar{r}_k^2}{\zeta_\kappa(\bar{r}_k)} \bar{\ell}_k^2$ and $\bar{r}_{\mathcal{T}_{out}-1} \leq 3d_0$.

The final line holds from the previous noting that,

$$\sum_{k=0}^{t} (\tilde{\eta}_k \langle g_k, \exp_{x_k}^{-1}(x_\star) \rangle_{x_k})^2 \leq \sum_{k=0}^{t} \tilde{\eta}_k^2 \zeta_\kappa(d_k) \|g_k\|_{x_k}^2 d_k^2 \leq (4d_0)^2 \sum_{k=0}^{t} \tilde{\eta}_k^2 \zeta_\kappa(d_k) \|g_k\|_{x_k}^2, \tag{181}$$

where the first inequality follows from the Cauchy-Schwartz inequality and the second inequality holds from the fact that only terms with $k < \mathcal{T}_{out}$ contribute to the sum. □

Using the above lemma, we can establish the following concentration bound.

Lemma E.11. *If Assumption 3.1, 3.2, and 3.3 hold, under the truncated T-RDoWG step size sequence $\{\tilde{\eta}_t\}$, the iterates satisfy,*

$$\mathbb{P}\left(\exists t \leq T : \sum_{k=0}^{t-1} \tilde{\eta}_k \langle \Delta_k, \exp_{x_k}^{-1}(x_\star) \rangle_{x_k} > d_0^2 \right) \leq \delta. \tag{182}$$

Proof. Consider the filtration $\mathcal{F}_t = \sigma(\mathcal{G}(x_0) \ldots, \mathcal{G}(x_t))$ and define $X_t = \tilde{\eta}_t \langle \Delta_t, \exp_{x_t}^{-1}(x_\star) \rangle_{x_t}$ and $\hat{X}_t = -\tilde{\eta}_t \langle \operatorname{grad} f(x_t), \exp_{x_t}^{-1}(x_\star) \rangle_{x_t}$. Then we have that X_t is a martingale difference sequence adapted to $\mathcal{F}_t$ and $\hat{X}_t \in \mathcal{F}_{t-1}$. Moreover, we have $\max\{|X_t|, |\hat{X}_t|\} \leq c$ almost surely for $c = \frac{24 d_0^2}{8^4 \theta_{T,\delta}}$. Substituting into Lemma A.9, we have

$$\mathbb{P}\left(\exists t \leq T : \left| \sum_{k=0}^{t-1} X_k \right| \geq 4 \sqrt{\theta_{t,\delta} \sum_{k=0}^{t-1} (X_k - \hat{X}_k)^2 + c^2 \theta_{t,\delta}^2} \right) \leq \delta. \tag{183}$$

Noting that $X_t - \hat{X}_t = \tilde{\eta}_t \langle g_t, \exp_{x_t}^{-1}(x_\star) \rangle_{x_t}$ and substituting the definition of c and the bound gives for every $t < T$,

$$4 \sqrt{\theta_{t,\delta} \sum_{k=0}^{t-1} (X_k - \hat{X}_k)^2 + c^2 \theta_{t,\delta}^2} \leq 4 \sqrt{\theta_{t,\delta} \frac{3 \cdot 4^3 d_0^4}{8^4 \theta_{T,\delta}} + \left(\frac{6 \theta_{t,\delta} d_0^2}{8^2 \sqrt{\zeta_\kappa(3d_0)} \theta_{T,\delta}} \right)^2} \leq d_0^2. \tag{184}$$

□

Lemma E.12. *Suppose Assumption 3.1, 3.2, and 3.3 hold. If* $\sum_{k=0}^{t-1} \tilde{\eta}_k \langle \Delta_k, \exp_{x_k}^{-1}(x_\star)\rangle_{x_k} \leq d_0^2$ *for all* $t \leq T$ *then* $\mathcal{T}_{out} > T$ *i.e.,* $\bar{r}_t \leq 3d_0$.

Proof. To condense notation, let $B_t := \max_{t' \leq t} \sum_{k=0}^{t'-1} \tilde{\eta}_k \langle \Delta_k, \exp_{x_k}^{-1}(x_\star)\rangle_{x_k}$, so the claim becomes $B_t \leq d_0^2$ implies $\mathcal{T}_{out} > t$ for all $t \leq T$. We prove the claim by induction on t. The basis of the induction is that $\mathcal{T}_{out} > 0$ always hold since $\bar{r}_0 = \epsilon \leq 3d_0$ by assumption. For the induction step, we assume that B_t implies $\mathcal{T}_{out} \geq t$ and show that $B_t \leq d_0^2$ implies $\mathcal{T}_{out} > t$. To that end, we use $\langle \operatorname{grad} f(x_t), \exp_{x_t}^{-1}(x_\star)\rangle_{x_t} \geq f(x_t) - f(x_\star) \geq 0$ to rearrange Lemma A.2 as

$$d_{k+1}^2 - d_k^2 \leq \eta_k^2 \zeta_\kappa(d_k) \|g_k\|_{x_k}^2 + 2\eta_k \langle \Delta_k, \exp_{x_k}^{-1}(x_\star)\rangle_{x_k} \tag{185}$$

for all k. Summing from $0 \leq k \leq t-1$, we have

$$d_t^2 - d_0^2 \leq \sum_{k=0}^{t-1} \eta_k^2 \zeta_\kappa(d_k) \|g_k\|_{x_k}^2 + 2\sum_{k=0}^{t-1} \eta_k \langle \Delta_k, \exp_{x_k}^{-1}(x_\star)\rangle_{x_k} \tag{186}$$

$$= \sum_{k=0}^{t-1} \tilde{\eta}_k^2 \zeta_\kappa(d_k) \|g_k\|_{x_k}^2 + 2\sum_{k=0}^{t-1} \tilde{\eta}_k \langle \Delta_k, \exp_{x_k}^{-1}(x_\star)\rangle_{x_k}. \tag{187}$$

where the equality holds since $\mathcal{T}_{out} \geq t$ and therefore $\eta_k = \tilde{\eta}_k$ for all $0 \leq k \leq t-1$. Now, by previous lemma we have $\sum_{k=0}^{t-1} \tilde{\eta}_k^2 \zeta_\kappa(d_k) \|g_k\|_{x_k}^2 \leq \frac{12 d_0^2}{8^4 \theta_{T,\delta}} \leq d_0^2$. Moreover, by assumption we have $\sum_{k=0}^{t-1} \tilde{\eta}_k \langle \Delta_k, \exp_{x_k}^{-1}(x_\star)\rangle_{x_k} \leq B_t \leq d_0^2$, from which we conclude, $d_t^2 \leq 4d_0^2$ and hence $r_t \leq d_0 + d_t \leq 3d_0$. Finally, since $\bar{r}_t = \max\{\bar{r}_{t-1}, r_t\}$ and $\bar{r}_{t-1} \leq 3d_0$ by the induction assumption, we have that $\bar{r}_t \leq 3d_0$. □

Theorem E.13. *Suppose Assumption 3.1, 3.2, and 3.3 hold, and* $\epsilon \leq 3d_0$. *Then for any* $\delta \in (0,1)$ *and* $t \in \mathbb{N}$, *under the T-RDoWG step size sequence* $\{\eta_k\}$, *the iterates satisfy* $\mathbb{P}(\bar{r}_t > 3d_0) \leq \delta$.

Proof. A consequence of combining the previous two lemmas. □

Corollary E.14. *Suppose that Assumption 3.1, 3.2, and 3.3 hold. For any* $\delta \in (0, 1/2)$, $t \in \mathbb{N}$, *consider T iterations of* $\{\eta_k\}$, *with initial step size of* $\epsilon \leq 3d_0$. *Then for* $\tau \in \arg\max_{t \leq T} \sum_{s=0}^{t-1} \frac{\bar{r}_s^2/\zeta_\kappa(\bar{r}_s)}{\bar{r}_t^2/\zeta_\kappa(\bar{r}_t)}$ *we have, with probability at least* $1 - 2\delta$, *that*

$$f(\tilde{x}_\tau) - f(x_\star) = O\left(c_{\delta,\epsilon,T} \frac{d_0 \sqrt{\zeta_\kappa(d_0)(G_\tau + L_\star^2)}}{T}\right) = O\left(c_{\delta,\epsilon,T} \frac{d_0 \sqrt{\zeta_\kappa(d_0)} L_\star}{\sqrt{T}}\right). \tag{188}$$

where $L_\star := \max_{x \in \mathcal{M}: d(x,x_0) \leq 3d(x_\star,x_0)} \ell(x)$ *and* $c_{\delta,\epsilon,T} = \log_+(T \frac{d_0 L_\star}{f(x_0) - f(x_\star)}) \log_+(\frac{d_0}{\epsilon}) \log(\frac{\log_+(T)}{\delta})$.

Proof. Here we adapt theorem Theorem E.4. Using that $\bar{r}_t \leq 3d_0$, we have $\zeta_\kappa(d_0 + \bar{r}_t) \leq \zeta_\kappa(4d_0) \leq \frac{\sqrt{\kappa} 4d_0}{\tanh(\sqrt{\kappa} 4d_0)} \leq \frac{\sqrt{\kappa} 4d_0}{\tanh(\sqrt{\kappa} d_0)} = O(\zeta_\kappa(d_0))$ for $\kappa > 0$, otherwise $\zeta_\kappa(d_0 + \bar{r}_t) = 1 = \zeta_\kappa(d_0) = O(\zeta_\kappa(d_0))$ in the case $\kappa = 0$. Now, by Assumption 3.3 we have $\ell_0 \geq \|\operatorname{grad} f(x_0)\|_{x_0} \geq (f(x_0) - f(x_\star))/d_0$, while $\bar{r}_T \leq 3d_0$ gives $\bar{\ell}_T \leq L_\star$. Therefore, $\log_+\left(1 + \frac{T\bar{\ell}_T^2}{\ell_0^2}\right) = O\left(\log_+\left(T \frac{d_0 L_\star}{f(x_0) - f(x_\star)}\right)\right)$. □

E.6. Omitting Geometric Curvature Term Analysis

We analyze omitting the geometric curvature term from the denominator RDoWG (Algorithm 2). Thus we consider step sizes of the form

$$\eta_t = \frac{\bar{r}_t^2}{\sqrt{v_t}}, \quad v_t = v_{t-1} + \bar{r}_t^2 \|g_s\|_{x_s}^2, \quad v_{-1} = 0. \tag{189}$$

We term this algorithm *Curvature Omitted Riemannian Distance over Weighted Gradients* (CO-RDoWG).

We consider the bound the error of the weighted average sequence,

$$\tilde{x}_{t+1} = \exp_{\tilde{x}_t}\left(\frac{\bar{r}_t^2}{\sum_{s=0}^t \bar{r}_s^2} \exp_{\tilde{x}_t}^{-1}(x_t)\right), \quad \tilde{x}_1 = x_0.$$

For a geodesically convex function $f\colon \mathcal{M} \to \mathbb{R}$, we have by Lemma A.3 that $\tilde{x}_t$ satisfies,

$$f(\tilde{x}_t) - f(x_\star) \leq \frac{1}{\sum_{s=0}^{t-1} \bar{r}_s^2} \sum_{s=0}^{t-1} \bar{r}_s^2 \langle -\operatorname{grad} f(x_s), \exp_{x_s}^{-1}(x_\star)\rangle_{x_s}. \tag{190}$$

Recalling g_s is the stochastic oracle evaluation, $\mathcal{G}(x_s)$, the numerator decomposes into two components:

$$\underbrace{\sum_{s=0}^{t-1} \bar{r}_s^2 \langle -g_s, \exp_{x_s}^{-1}(x_\star)\rangle_{x_s}}_{\text{weighted regret}} + \underbrace{\sum_{s=0}^{t-1} \bar{r}_s^2 \langle \Delta_s, \exp_{x_s}^{-1}(x_\star)\rangle_{x_s}}_{\text{noise}}, \tag{191}$$

with $\Delta_s \coloneqq g_s - \operatorname{grad} f(x_s)$.

Supporting Analysis

We give a deterministic bound for the weighted regret (Lemma E.15).

Lemma E.15. *Under Assumption 3.1 and 3.2, the iterates of CO-RDoWG, satisfy*

$$\sum_{s=0}^{t-1} \bar{r}_s^2 \langle -g_s, \exp_{x_s}^{-1}(x_\star)\rangle_{x_s} \leq \bar{r}_t \left(2\bar{d}_t + \bar{r}_t \zeta_\kappa(\bar{d}_t)\right) \sqrt{v_{t-1}}. \tag{192}$$

Proof. Follow the same argument as Lemma D.13 but with weights $\bar{r}_s^2$ replacing $\bar{r}_s$ and weighted gradient sum v_s replacing the standard gradient sum G_s. □

Non-Smooth Analysis

We give deterministic bounds for the weighted regret (Lemma E.16) and high probability bounds for the noise term (Lemma E.17) in the non-smooth setting.

Lemma E.16. *Under Assumption 3.1 and 3.2, the iterates of CO-RDoWG, satisfy*

$$\sum_{s=0}^{t-1} \bar{r}_s^2 \langle -g_s, \exp_{x_s}^{-1}(x_\star)\rangle_{x_s} \leq \bar{r}_t^2 \left(2\bar{d}_t + \bar{r}_t \zeta_\kappa(\bar{d}_t)\right) \sqrt{G_{t-1}}. \tag{193}$$

Proof. Using the bound of E.15 and that $\sqrt{v_{t-1}} \leq \bar{r}_{t-1}\sqrt{G_{t-1}} \leq \bar{r}_t \sqrt{G_{t-1}}$ gives the result. □

Lemma E.17. *For all $\delta \in (0,1), T \in \mathbb{N}$ and $L > 0$, if Assumption 3.1, 3.2, and 3.3 hold, the iterates of CO-RDoWG satisfy*

$$\mathbb{P}\left(\exists t \leq T : \left|\sum_{s=0}^{t-1} \bar{r}_s^2 \langle \Delta_s, \exp_{x_s}^{-1}(x_\star)\rangle_{x_s}\right| \geq b_t\right) \leq \delta + \mathbb{P}(\bar{\ell}_T > L), \tag{194}$$

where $b_t = 8\bar{r}_{t-1}^2 \bar{d}_{t-1} \sqrt{\theta_{t,\delta} G_{t-1} + \theta_{t,\delta}^2 L^2}$ and $\bar{\ell}_T \coloneqq \max_{s \leq T} \ell(x_s)$.

Proof. Follow argument of Lemma D.14. □

Combining the above results, we obtain the following.

Theorem E.18. *For all $\delta \in (0,1)$ and $L > 0$, if Assumption 3.1, 3.2, and 3.3 hold, then with probability at least $1 - \delta - \mathbb{P}(\bar{\ell}_T > L)$, for all $t \leq T$, the optimality gap on the weighted iterates $f(\tilde{x}_t) - f(x_\star)$ of CO-RDoWG satisfy*

$$O\left(\frac{(d_0 + \bar{r}_t)\zeta_\kappa(d_0 + \bar{r}_t)\sqrt{G_{t-1} + \theta_{t,\delta} G_{t-1} + \theta_{t,\delta}^2 L^2}}{\sum_{s=0}^{t-1} \bar{r}_s^2 / \bar{r}_t^2}\right). \tag{195}$$

Proof. Combine Lemma E.16 and Lemma E.17 and use the fact $\bar{d}_t \leq d_0 + \bar{r}_t$. □

Thus in comparison to standard RDoWG, we pay an additional cost of $O\left(\sqrt{\zeta_\kappa(d_0 + \bar{r}_t)}\right)$ for omitting the geometric curvature term with CO-RDoWG.

We then have a useful result when the manifold is bounded but its exact diameter is unknown.

Corollary E.19. *Under Assumption 3.1, 3.2, and 3.3, for any $D \geq d_0$ let $L_D \coloneqq \max_{x \in \mathcal{M}: d(x,x_0) \leq D} \ell(x)$. Then, for all $\delta \in (0,1)$ and for $\tau \in \arg\max_{t \leq T} \sum_{s=0}^{t-1} \frac{\bar{r}_s^2}{\bar{r}_t^2}$, with probability at least $1 - \delta - \mathbb{P}(\bar{\ell}_T > L)$, iterates of CO-DoWG satisfy the optimality gap bound*

$$f(\tilde{x}_\tau) - f(x_\star) = O\left(\frac{D\zeta_\kappa(D)\sqrt{G_{\tau-1}\theta_{\tau,\delta} + L_D^2\theta_{\tau,\delta}^2}}{T}\log_+(D/\epsilon)\right). \tag{196}$$

Proof. Apply Lemma A.6 to the denominator term of Theorem E.18. □

Thus in comparison to standard RDoWG, we pay an additional cost of $O\left(\sqrt{\zeta_\kappa(D)}\right)$ for omitting the curvature term with CO-RDoWG.

SMOOTH ANALYSIS

Lemma E.20. *Suppose f is S-smooth and assume Assumption 3.1 and 3.2 hold. Then we have that the iterates of CO-RDoWG satisfy*

$$\sum_{s=0}^{t-1} \bar{r}_s^2 \langle -g_s, \exp_{x_s}^{-1}(x_\star)\rangle_{x_s} \leq \bar{r}_t\left(2\bar{d}_t + \bar{r}_t\zeta_\kappa(\bar{d}_t)\right)\sqrt{2S\sum_{s=0}^{t-1}\bar{r}_s^2(f(x_s) - f(x_\star))}. \tag{197}$$

Proof. By smoothness we can use Lemma A.4 to deduce $\|\operatorname{grad} f(x)\|_x^2 \leq 2S(f(x) - f(x_\star))$ for all $x \in \mathcal{M}$. Therefore

$$v_t = \sum_{s=0}^{t-1}\bar{r}_s^2\|\operatorname{grad} f(x_s)\|_{x_s}^2 \leq 2S\sum_{s=0}^{t-1}\bar{r}_s^2(f(x_s) - f(x_\star)). \tag{198}$$

Taking square roots and substituting this into Lemma A.4 gives the result. □

Lemma E.21. *Suppose Assumption 3.1, 3.2, and 3.4 hold. Then for all $\delta \in (0,1), T \in \mathbb{N}$ and $S > 0$, Then we have that the iterates of CO-RDoWG satisfy*

$$\mathbb{P}\left(\exists t \leq T : \left|\sum_{s=0}^{t-1}\bar{r}_s^2\langle \Delta_s, \exp_{x_s}^{-1}(x_\star)\rangle_{x_s}\right| \geq b_t\right) \leq \delta + \mathbb{P}(\bar{s}_T > S), \tag{199}$$

where $b_t = 8\bar{r}_{t-1}\bar{d}_{t-1}\sqrt{2S\theta_{t,\delta} + 8S\theta_{t,\delta}^2}\sqrt{\sum_{s=0}^{t-1}\bar{r}_s^2\left[f(x_s) - f(x_\star)\right]}$ and $\bar{s}_T \coloneqq \max_{t \leq T} s(x_t)$.

Proof. Follow argument of Lemma E.7. □

Combining the above results, we obtain the following.

Theorem E.22. *For all $\delta \in (0,1)$ and $S > 0$, if Assumption 3.1, 3.2, and 3.4 hold, then with probability at least $1 - \delta - \mathbb{P}(\bar{s}_T > S)$, for all $t \leq T$, CO-RDoWG satisfies the optimality gap $f(\tilde{x}_t) - f(x_\star)$ of*

$$O\left(\frac{\left((d_0 + \bar{r}_t)\zeta_\kappa(d_0 + \bar{r}_t)\right)^2(S\theta_{t,\delta}^2)}{\sum_{s=0}^{t-1}\frac{\bar{r}_s^2}{\bar{r}_t^2}}\right). \tag{200}$$

Proof. Using Lemma E.20 and Lemma E.21 above, we have with the relevant probabilistic conditions,

$$\sum_{s=0}^{t-1} \bar{r}_s^2[f(x_s) - f(x_\star)] \leq \left(\sqrt{2S}\bar{r}_t \left(2\bar{d}_t + \bar{r}_t\zeta_\kappa(\bar{d}_t)\right) + 8\bar{r}_t\bar{d}_t\sqrt{2S\theta_{t,\delta} + 8S\theta_{t,\delta}^2}\right)\sqrt{\sum_{s=0}^{t-1} \bar{r}_s^2\left[f(x_s) - f(x_\star)\right]}. \tag{201}$$

Now if $f(x_s) - f(x_\star) = 0$ for some iterate, then the statement is trivial. Otherwise diving by sides by the square root term, we have

$$\sqrt{\sum_{s=0}^{t-1} \bar{r}_s^2\left[f(x_s) - f(x_\star)\right]} \leq \left(\sqrt{2S}\bar{r}_t \left(2\bar{d}_t + \bar{r}_t\zeta_\kappa(\bar{d}_t)\right) + 8\bar{r}_t\bar{d}_t\sqrt{2S\theta_{t,\delta} + 8S\theta_{t,\delta}^2}\right). \tag{202}$$

We square both sides and divide through by $\sum_{s=0}^{t-1} \bar{r}_s^2$,

$$\frac{1}{\sum_{s=0}^{t-1} \bar{r}_s^2}\sum_{s=0}^{t-1} \bar{r}_s^2\left[f(x_s) - f(x_\star)\right] \leq O\left(\frac{\left(2\bar{d}_t + \bar{r}_t\zeta_\kappa(\bar{d}_t)\right)^2\left(S\theta_{t,\delta}^2\right)}{\sum_{s=0}^{t-1}\frac{\bar{r}_s^2}{\bar{r}_t^2}}\right). \tag{203}$$

Now using the fact, $\bar{d}_t \leq d_0 + \bar{r}_t$, in the above bound gives the result. □

Thus in comparison to standard RDoWG, we pay an additional cost of $O\left(\sqrt{\zeta_\kappa(d_0 + \bar{r}_t)}\right)$ for omitting the geometric curvature term with CO-RDoWG.

We then have a useful result when the manifold is bounded but its exact diameter is unknown.

Corollary E.23. *Under Assumption 3.1, 3.2, and 3.4, for any $D \geq d_0$ let $S_D \coloneqq \max_{x\in\mathcal{M}:d(x,x_0)\leq D} s(x)$. Then, for all $\delta \in (0,1)$ and for $\tau \in \arg\max_{t\leq T}\sum_{s=0}^{t-1}\frac{\bar{r}_s^2}{\bar{r}_t^2}$, with probability at least $1 - \delta - \mathbb{P}(\bar{s}_T > S)$, the iterates of CO-RDoWG satisfies the optimally gap $f(\tilde{x}_\tau) - f(x_\star)$ of*

$$O\left(\frac{D^2\zeta_\kappa(D)^2 S_D\theta_{\tau,\delta}^2}{T}\log_+(D/\epsilon)\right). \tag{204}$$

Proof. Apply Lemma A.6 to the denominator term of Theorem E.22. □

Thus in comparison to standard RDoWG, we pay an additional cost of $O\left(\sqrt{\zeta_\kappa(D)}\right)$ for omitting the curvature term with CO-RDoWG.

F. NRDoG Overview

Here we present a learning-rate-free schedule for NRSGD: *Normalized Riemannian Distance over Gradients* (NRDoG).

Algorithm 4 NRDoG

Input: initial point x_0, initial estimate $\epsilon > 0$, $G_{-1} = 0$.
for $t = 0$ **to** $T - 1$ **do**
  $g_t = \mathcal{G}(x_t)$
  $\bar{r}_t = \max\left(\epsilon, \max_{s\leq t} d(x_s, x_0)\right)$
  $\eta_t = \frac{\bar{r}_t}{\sqrt{(t+1)\zeta_\kappa(\bar{r}_t)}}$
  $x_{t+1} = \exp_{x_t}\left(-\eta_t\frac{g_t}{\|g_t\|_{x_t}}\right)$
end for

G. Geometry of Specific Riemannian Manifolds

G.1. Sphere Manifold

The sphere manifold $\mathbb{S}^{d-1} \coloneqq \{x \in \mathbb{R}^d : \|x\| = 1\}$ is an embedded submanifold of $\mathbb{R}^d$ with tangent space $\mathcal{T}_x\mathbb{S}^{d-1} = \{v \in \mathbb{R}^d : x^T v = 0\}$. The Riemannian metric is given by the Euclidean inner product $\langle \cdot, \cdot' \rangle_x = \langle \cdot, \cdot' \rangle$. The exponential map is given by $\exp_x(v) = \cos(\|v\|)x + \sin(\|v\|)\frac{v}{\|v\|}$ with inverse exponential map as $\exp_x^{-1}(y) = \arccos(x^T y)\frac{\operatorname{Proj}_x(y-x)}{\|\operatorname{Proj}_x(y-x)\|}$ where $\operatorname{Proj}_x(v) = v - (x^T v)x$ is the orthogonal projection of any $v \in \mathbb{R}^d$ to the tangent space $\mathcal{T}_x\mathbb{S}^{d-1}$. Following the Pymanopt implementation (Townsend et al., 2016), parallel transport is approximated with the projection operation, i.e., $\Gamma_x^y v \approx \operatorname{Proj}_x(v)$.

G.2. Grassmann Manifold

The Grassmann manifold of dimension $d \times r$, denoted as $\mathbb{G}(d, r)$ is the set of all r dimensional subspaces in $\mathbb{R}^d$ $(d \geq r)$. Each point on the Grassmann manifold can be identified as a column of orthonormal matrices $x \in \mathbb{R}^{d \times r}$, $x^T x = \mathrm{I}$ and two points x, y are equivalent if $x = yo$ for some $r \times r$ orthogonal matrix o. For our implementation of the exponential map, inverse exponential map, and parallel transport, we directly translate the Pymanopt code (Townsend et al., 2016) from NumPy to JAX.

G.3. Poincaré Manifold

The Poincaré manifold of dimension d is given by the open d-dimensional unit ball $\mathbb{B}_d \coloneqq \{x \in \mathbb{R}^d : \|x\| < 1\}$ equipped with Riemannian metric $\langle \cdot, \cdot' \rangle_x = 4/(1 - \|x\|^2)^2 \langle \cdot, \cdot' \rangle$. The *Möbius addition* of x and y in $\mathbb{B}^d$ is defined as (Ungar, 2008)

$$x \oplus y \coloneqq \frac{(1 + 2\langle x, y \rangle + \|y\|^2)x + (1 - \|x\|^2)y}{1 + 2\langle x, y \rangle + \|x\|^2\|y\|^2}.$$

Defining the *conformal factor* as $\lambda_x \coloneqq 2/(1 - \|x\|^2)$, the exponential map is given by $\exp_x(v) = x \oplus \left(\tanh\left(\lambda_x \frac{\|x\|}{2}\right)\right)\frac{v}{\|v\|}$ and the inverse exponential map is given by $\exp_x^{-1}(y) = \frac{2}{\lambda_x}\tanh^{-1}(\|-x \oplus y\|)\frac{-x \oplus y}{\|-x \oplus y\|}$. Parallel transport can also be given in closed form (see Ungar, 2008, for further details).

H. Additional Numerical Results

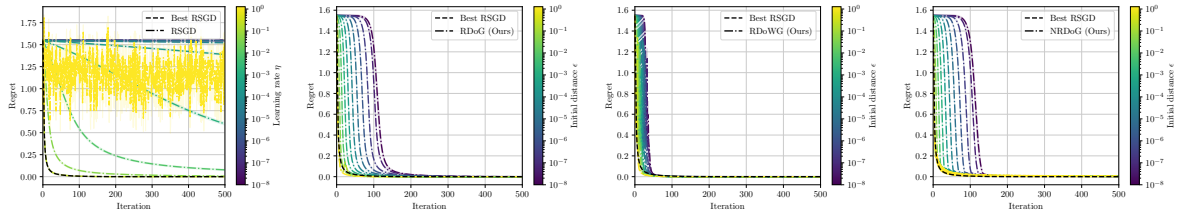
H.1. Rayleigh Quotient Maximization on the Sphere

In this section, we provide additional results for the Rayleigh quotient maximization discussed in Section 5.1 with a consistent setup across $d = 1000$ dimensions. The initial figures in Figure 5 emphasize the learning-rate-free adaptability and insensitivity to the choice of the initial distance estimate, $\epsilon \in [10^{-8}, 10^0]$, for RDoG, RDoWG, and NRDoG, particularly after a few hundred iterations. In contrast, we observe a notable impact on the performance of RSGD due to the choice of the learning rate, $\eta \in [10^{-8}, 10^0]$. This sensitivity in regret also influences solution quality, as illustrated in Figure 6.

We proceed to evaluate the algorithms for various numbers of iterations $T \in \{100, 500, 1000, 2000\}$, showcasing regret in Figure 7 and geodesic distance to a numerically computed optimum in Figure 8 for different learning rates, $\eta \in [10^{-8}, 10^6]$, for RSGD and RADAM. Additionally, we explore different initial distance estimates, $\epsilon \in [10^{-8}, 10^{-1}]$, for RDoG, RDoWG, and NRDoG. Notably, we observe that for $T = 100$ iterations, the initial distance estimate does impact the algorithms, but after $T = 500$ iterations, the effect becomes insensitive over several orders of magnitude, mirroring Figure 7 and Figure 8. Conversely, RADAM and RSGD exhibit a requirement for careful tuning.

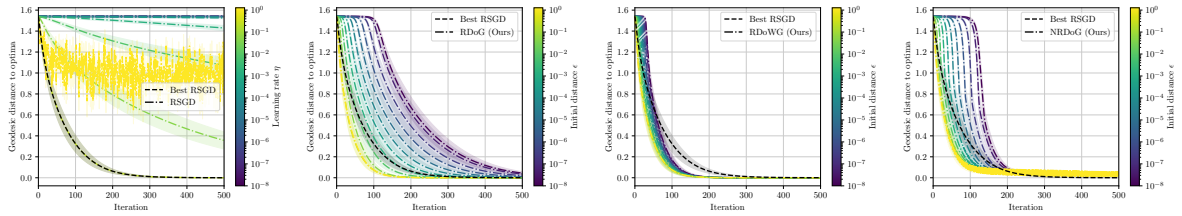
H.2. PCA on the Grassmann Manifold

In this section, we present additional results for the PCA on the Grassmann manifold discussed in Section 5.2, maintaining a consistent experimental setup. In Table 1, we observe that while RSGD exhibits sensitivity to the learning rate $\eta \in [10^{-8}, 10^0]$, RDoG, RDoWG, and NRDoG quickly adapt and achieve performance comparable to the best learning rate for RSGD within 500 iterations, irrespective of the chosen initial distance $\epsilon \in [10^{-8}, 10^0]$. This adaptability is further evident in Table 2, where we consider halting the algorithms for $T \in \{100, 500, 1000, 2000\}$ iterations and comparing the geodesic distance of the output of the optimizer with the numerical solution. Discrepancies are noticeable for $T = 100$, but these discrepancies diminish for $T = 500$ iterations and beyond.



(a) RSGD. (b) RDoG. (c) RDoWG. (d) NRDoG.

Figure 5. **Supplementary results for Rayleigh quotient maximization on the sphere (Section 5.1).** The plots depict regret as a function of the iteration, considering various learning rates. Results are averaged over ten random replications. The optimal RSGD is chosen based on minimizing the regret after 5000 iterations. Note that (a) and (b) are equivalent to Figure 2 (b) and (c) respectively.



(a) RSGD. (b) RDoG. (c) RDoWG. (d) NRDoG.

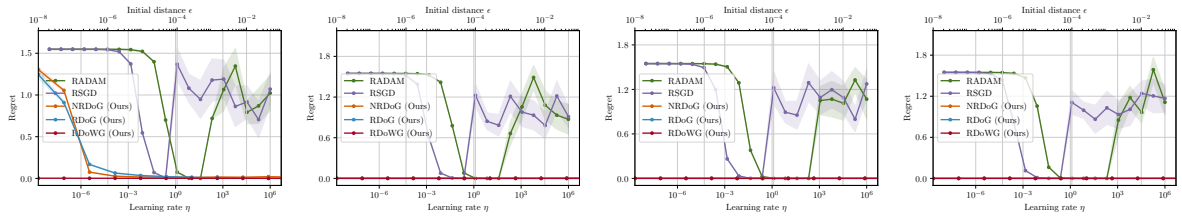
Figure 6. **Supplementary results for Rayleigh quotient maximization on the sphere (Section 5.1).** The plots display the geodesic distance from an optimum as a function of the iteration, considering various learning rates. Results are averaged over ten replicates with different initial points. The optimal RSGD is selected based on minimizing the geodesic distance from the optimum after 5000 iterations.

H.3. Embedding Graphs in the Poincaré Embeddings

In this section, we provide supplementary results concerning Poincaré embeddings, as detailed in Section 5.3.

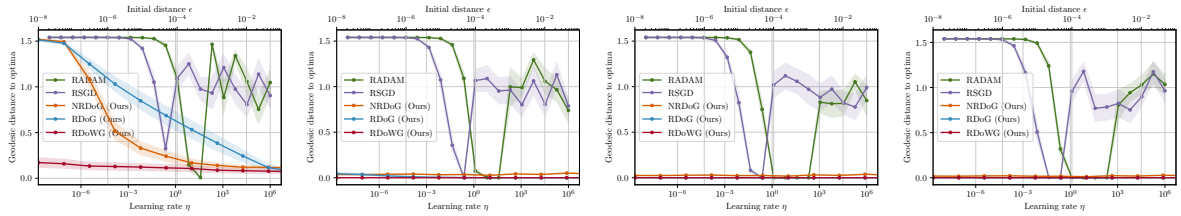
Table 3 maintains a consistent experimental framework with the main paper, focusing on five-dimensional embeddings. The top-left section of the table corresponds to Figure 4(a) presented in the main paper, serving as a reference for comparison. In the bottom-left segment, we explore the algorithmic performance of RADAM and RSGD without implementing the burn-in heuristic, which results in inferior performance. Notably, our optimizers demonstrate robustness, eliminating the need for such heuristics. On the left-hand column, we investigate the impact of omitting the curvature term from the learning rates. For RDoG and RDoWG, the curvature omission corresponds to CO-RDoG (Appendix D.5) and CO-RDoWG (Appendix E.6). This omission leads to a performance decrease for NRDoG and RDoWG, while RDoG remains unaffected in performance.

Table 4 adheres to a consistent experimental framework for two-dimensional embeddings outlined in the main paper. In the right-hand column, we discern meaningful groupings across various categories without resorting to burn-in heuristics for RDoG, RDoWG, and NRDoG. Conversely, in the left-hand column, we emphasize the pivotal role of geometric curvature in governing step sizes; its absence results in inferior groupings. This discrepancy is reflected in the mean average precision



(a) T = 100. (b) T = 500. (c) T = 1000. (d) T = 2000.

Figure 7. **Supplementary Results for Rayleigh Quotient Maximization (Section 5.1).** Each plot illustrates the regret after the algorithm is halted for the specified number of iterations. Results are averaged over ten replicates with different initial points.



(a) T = 100. (b) T = 500. (c) T = 1000. (d) T = 2000.

Figure 8. **Supplementary results for Rayleigh quotient maximization on the sphere (Section 5.1).** Each plot illustrates the geodesic distance to a numerically computed optimum after the algorithm is halted for the specified number of iterations. Results are averaged over ten replicates with different initial points.

metric.

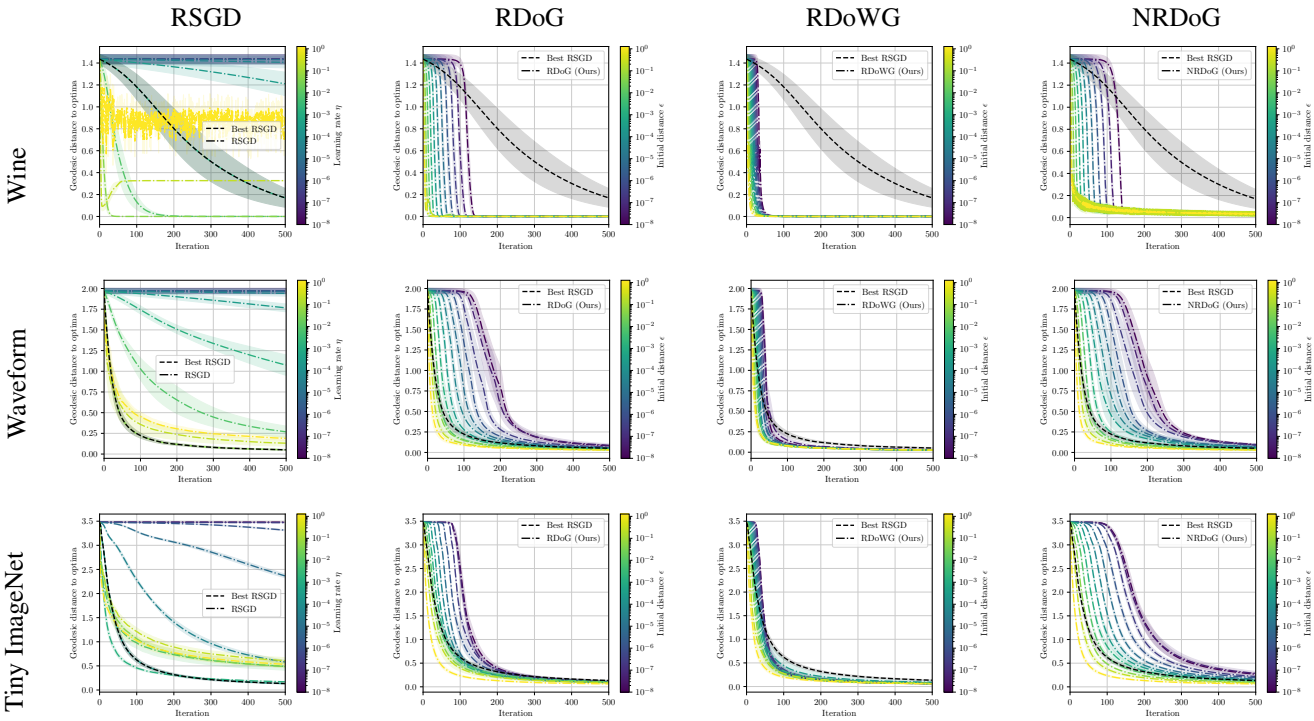


Table 1. **Supplementary results for PCA on the Grassmann manifold (Section 5.2).** The plots display the geodesic distance from a numerically computed optimum as a function of the iteration, considering various learning rates. Results are averaged over five replicates with different initial points. The optimal RSGD is selected based on minimizing the geodesic distance from the optimum after 2000 iterations.

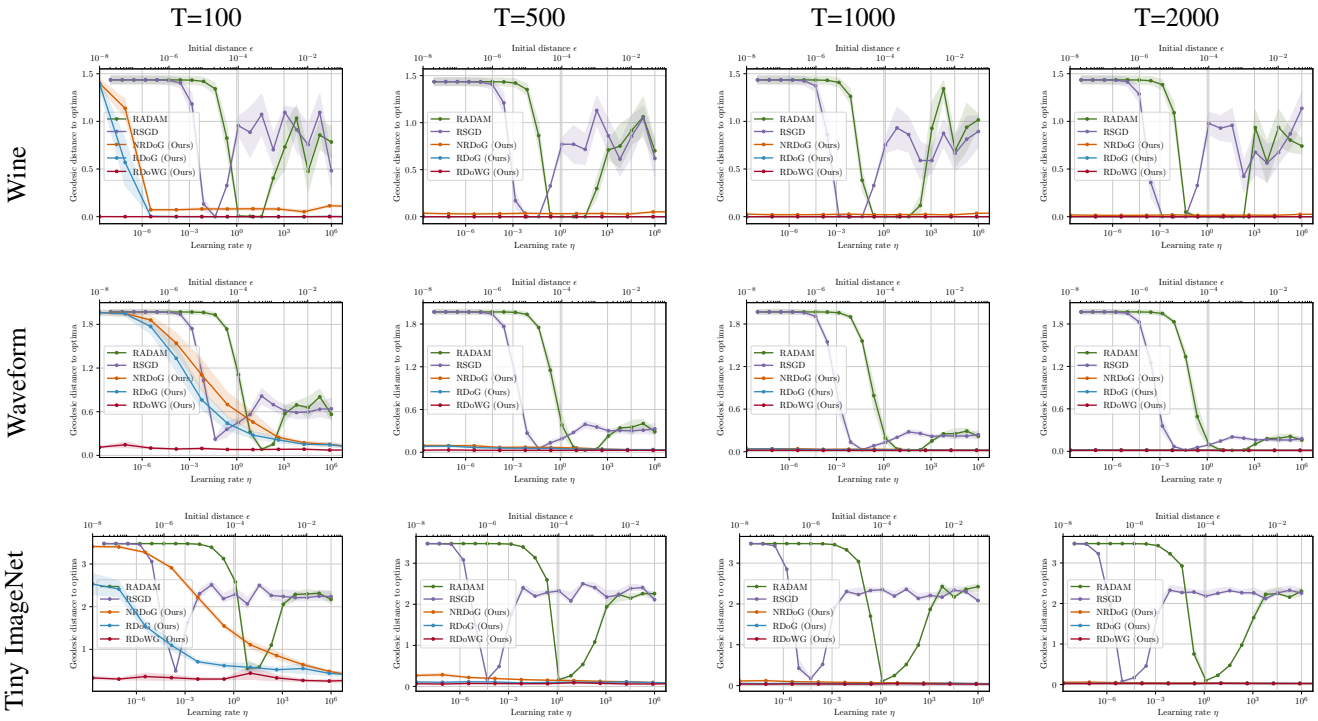


Table 2. **Supplementary results for PCA on the Grassmann manifold (Section 5.2).** Results for different datasets and methods. Each plot illustrates the geodesic distance to a numerically computed optimum after the algorithm is halted for the specified number of iterations. Results are averaged over ten replicates with different initial points.

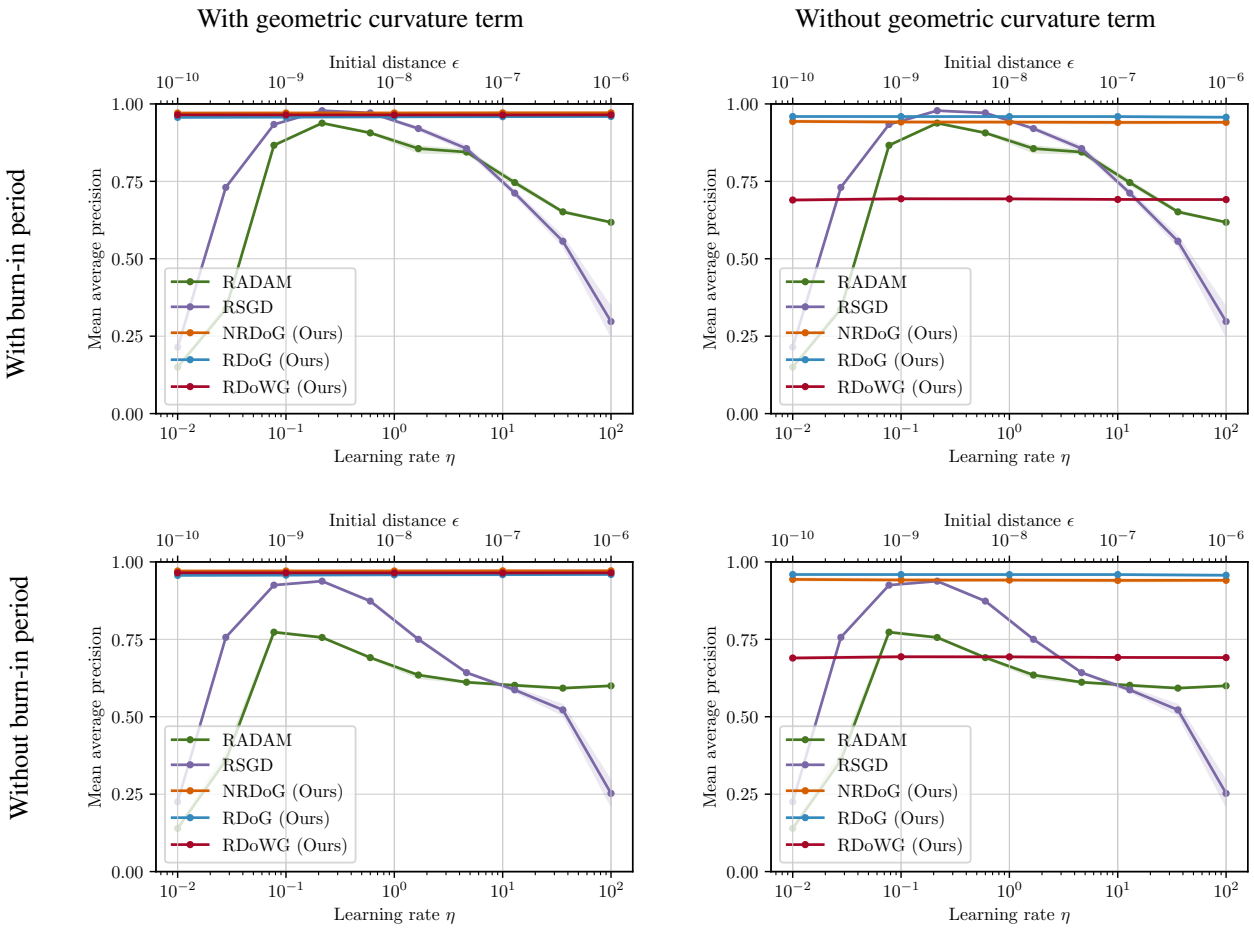


Table 3. **Supplementary results for the five-dimensional Poincaré word embeddings (Section 5.3).** We compute the mean average precision of the embeddings against the ground truth after 1000 training epochs. The reported results represent the average over five replications, with the dimension of the embeddings set to five. In the columns, "with geometric curvature term" corresponds to learning schedulers for RDoG, RDoWG, and NRDoG that retain the geometric curvature term in the denominator, while "without geometric curvature term" denotes the omission of this term. On the rows, "with burn-in period" indicates running RADAM and RSGD with a burn-in heuristic. In this case, the algorithms are executed with learning rates divided by ten for the initial ten epochs before regular training. "without burn-in period" signifies the absence of this heuristic.

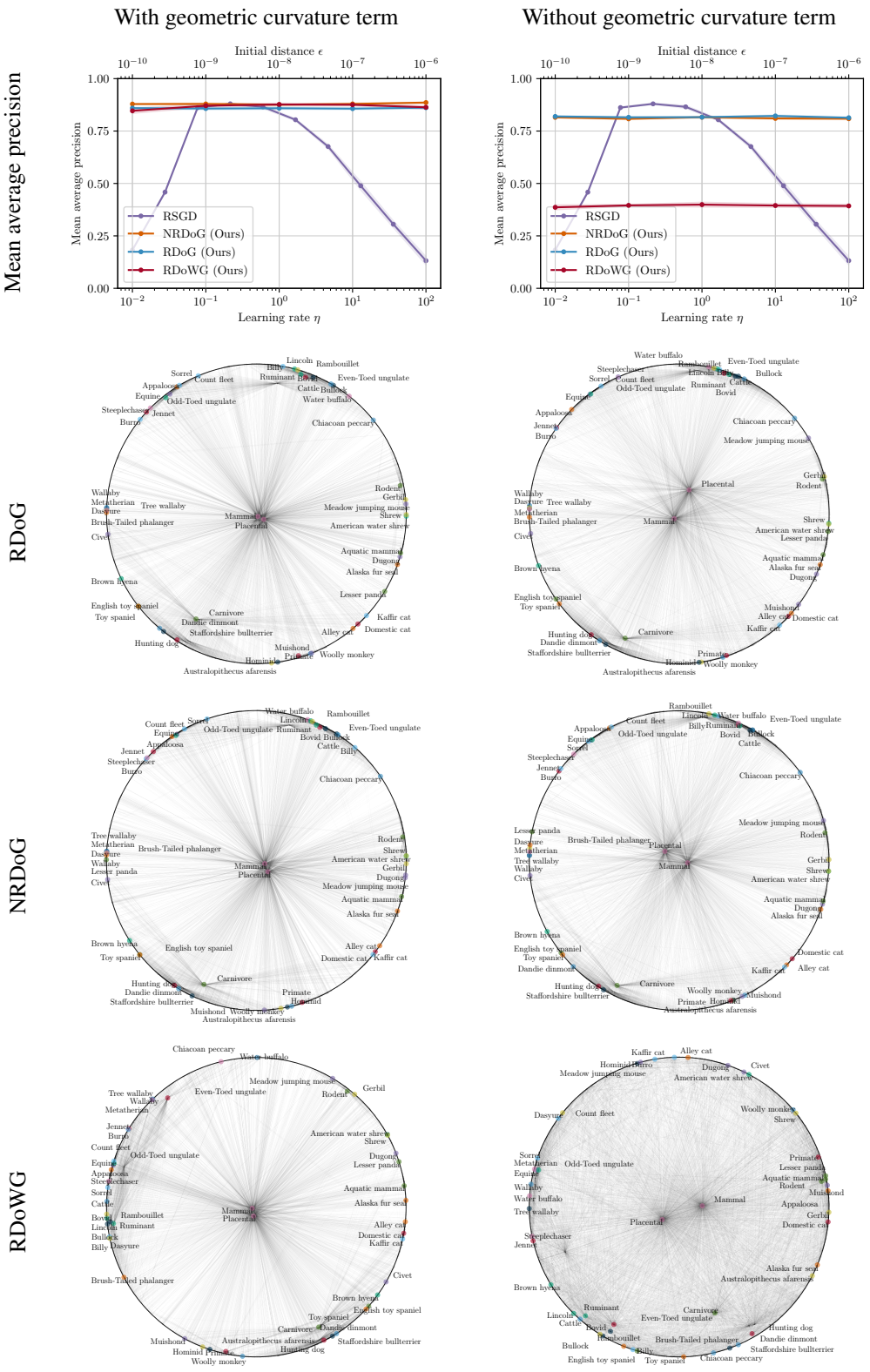


Table 4. **Supplementary results for the two-dimensional Poincaré word embeddings (Section 5.3).** We compute the mean average precision of the embeddings against the ground truth after 2000 training epochs. The reported results represent the average over five replications, with the dimension of the embeddings set to two. Plots of embeddings obtained under each optimizer are visualized and annotated for the first 50 nouns of the mammal's subtree. In the columns, "with geometric curvature term" corresponds to learning schedulers for RDoG, RDoWG, and NRDoG that retain the geometric curvature term in the denominator, while "without geometric curvature term" denotes the omission of this term.