Minimizing Dynamic Regret on Geodesic Metric Spaces

Zihao Hu ZIHAOHU@GATECH.EDU
Georgia Institute of Technology

Guanghui Wang GWANG369@GATECH.EDU
Georgia Institute of Technology

Jacob Abernethy ABERNETHYJ@GOOGLE.COM
Georgia Institute of Technology, Google Research

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

In this paper, we consider the sequential decision problem where the goal is to minimize the *general dynamic regret* on a complete Riemannian manifold. The task of offline optimization on such a domain, also known as a *geodesic metric space*, has recently received significant attention. The online setting has received significantly less attention, and it has remained an open question whether the body of results that hold in the Euclidean setting can be transplanted into the land of Riemannian manifolds where new challenges (e.g., *curvature*) come into play. In this paper, we show how to get optimistic regret bound on manifolds with non-positive curvature whenever improper learning is allowed and propose an array of adaptive no-regret algorithms. To the best of our knowledge, this is the first work that considers general dynamic regret and develops "optimistic" online learning algorithms which can be employed on geodesic metric spaces.

Keywords: Riemannian Manifolds, Optimistic Online Learning, Dynamic Regret

1. Introduction

Online convex optimization (OCO) in Euclidean space is a well-developed area with numerous applications. In each round, the learner takes an action from a decision set, while the *adversary* chooses a loss function. The long-term performance metric of the learner is (static) regret, which is defined as the difference between the learner's cumulative loss and the loss of playing the best-fixed decision in hindsight. As the name suggests, OCO requires both the losses and the decision set to be *convex*. From the theoretical perspective, convex functions and sets are well-behaved objects with many desirable properties that are generally required to obtain tight regret bounds.

Typical algorithms in OCO, such as *mirror descent*, determine how one should adjust parameter estimates in response to arriving data, typically by shifting parameters against the gradient of the loss. But in many cases of interest, the underlying parameter space is not only non-convex but non-Euclidean. The *hyperboloid*, for example, arising from the solution set of a degree-two polynomial, is a Riemannian manifold that has garnered interest as a tool in tree-embedding tasks (Lou et al., 2020). On such manifolds, we do have a generalized notion of convexity, known as *geodesic convexity* (Udriste, 2013). There are many popular problems of interest (Hosseini and Sra, 2015; Vishnoi, 2018; Sra et al., 2018) where the underlying objective function is geodesically convex (gsc-convex) under a suitable Riemannian metric. But there has thus far been significantly limited research on how to do *adaptive learning* in such spaces and to understand when regret bounds are obtainable.

Table 1: Summary of bounds. δ describes the discrepancy between the decision set and the comparator set. We define ζ in Def. 1 and let $B_T := \min\{V_T, F_T\}$.

Algorithm	Type	Dynamic regret
RADAR	Standard	$O(\sqrt{\zeta(1+P_T)T})$
	Lower bound	$\Omega(\sqrt{(1+P_T)T})$
RADAR$_{\text{v}}$	Gradient-variation	$O(\sqrt{\zeta(\frac{1+P_T}{\delta^2}+V_T)(P_T+1)})$
RADAR$_{\text{s}}$	Small-loss	$O(\sqrt{\zeta((1+P_T)\zeta+F_T)(P_T+1)})$
RADAR$_{\text{b}}$	Best-of-both-worlds	$O\left(\sqrt{\zeta(P_T(\zeta+\frac{1}{\delta^2})+B_T+1)(P_T+1)+B_T\ln T}\right)$

Let $\mathcal{N}$ be a gsc-convex subset of a geodesic metric space $\mathcal{M}$. In this paper, we consider the problem of minimizing the *general dynamic regret* on $\mathcal{N}$, defined as

$$\text{D-Regret}_T := \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}_t),$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_T \in \mathcal{N}$ is the sequence of actions taken by the learner, whose loss is evaluated relative to the sequence of "comparator" points $\mathbf{u}_1, \ldots, \mathbf{u}_T \in \mathcal{N}$. There has been recent work establishing that sublinear regret is possible as long as $\mathcal{N}$ and the f_t's are gsc-convex, for example using a Riemannian variant of Online Gradient Descent (Wang et al., 2021). But so far there are no such results that elicit better D-Regret using *adaptive* algorithms.

What do we mean by "adaptive" in this context? In the online learning literature there have emerged three key quantities of interest in the context of sequential decision making, the *comparator path length*, the *gradient variation*, and the *comparator loss*, defined respectively as:

$$\begin{aligned}
P_T &:= \textstyle\sum_{t=2}^{T} d(\mathbf{u}_t, \mathbf{u}_{t-1}), \\
V_T &:= \textstyle\sum_{t=2}^{T} \sup_{\mathbf{x}\in\mathcal{N}} \|\nabla f_{t-1}(\mathbf{x}) - \nabla f_t(\mathbf{x})\|^2, \\
F_T &:= \textstyle\sum_{t=1}^{T} f_t(\mathbf{u}_t).
\end{aligned} \tag{1}$$

Let us start by considering regret minimization with respect to path length. While it has been observed that $O(\sqrt{(1+P_T)T})$ is optimal in a minimax sense (Zhang et al., 2018), a great deal of research for the Euclidean setting (Srebro et al., 2010; Chiang et al., 2012; Rakhlin and Sridharan, 2013) has shown that significantly smaller regret is achievable when any one of the above quantities is small. These are not just cosmetic improvements either, many fundamental applications of online learning rely on these adaptive methods and bounds. We give a thorough overview in Section 3.

The goal of the present paper is to translate to the Riemannian setting an array of adaptive regret algorithms and prove corresponding bounds. We propose a family of algorithms which we call RADAR, for Riemannian adaptive dynamic regret. The three important variants of RADAR are RADAR$_{\text{v}}$, RADAR$_{\text{s}}$, and RADAR$_{\text{b}}$, we prove regret bounds on each, summarized in Table 1. We allow *improper learning* for the gradient-variation bound, which means the player can choose $\mathbf{x}_1, \ldots, \mathbf{x}_T$ from a slightly larger set $\mathcal{N}_{\delta G}$ (formally defined in Definition 2).

As a general matter, convex constraints on a Riemannian manifold introduce new difficulties in optimization that are not typically present in the Euclidean case, and there has been limited work on addressing these. To the best of our knowledge, there are only three papers considering how to incorporate constraints on manifolds, and these all make further assumptions on either the curvature or the diameter of the feasible set. Martínez-Rubio (2022) only applies to hyperbolic and spherical spaces. Criscitiello and Boumal (2022) works for complete Riemannian manifolds with sectional curvature in $[-K, K]$ but the diameter of the decision set is at most $O(\frac{1}{\sqrt{K}})$. Martínez-Rubio and Pokutta (2022) mainly works for locally symmetric Hadamard manifolds. Our paper is the first to consider the projective distortion in the *online* setting that applies to all Hadamard manifolds as long as improper learning is allowed without imposing further constraints on the diameter or the curvature.

Obtaining adaptive regret guarantees in the Riemannian setting is by no means a trivial task, as the new geometry introduces various additional technical challenges. Here is but one example: whereas the cost of a (Bregman) projection into a feasible region can be controlled using a generalized "Pythagorean" theorem in the Euclidean setting, this same issue becomes more difficult on a manifold as we encounter geometric distortion due to curvature. To better appreciate this, for a Hadamard manifold $\mathcal{M}$, assume the projection of $\mathbf{x} \in \mathcal{M}$ onto a convex subset $\mathcal{N} \subset \mathcal{M}$ is $\mathbf{z}$. While it is true that for any $\mathbf{y} \in \mathcal{N} \setminus \{\mathbf{z}\}$ the angle between geodesics $\overline{\mathbf{zx}}$ and $\overline{\mathbf{zy}}$ is obtuse, this property is only relevant at the tangent space of $\mathbf{z}$, yet we need to analyze gradients at the tangent space of $\mathbf{x}$. The use of *parallel transport* between $\mathbf{x}$ and $\mathbf{z}$ unavoidably incurs extra distortion and could potentially lead to $O(T)$ regret.

The last challenge comes from averaging on manifolds. For example, many adaptive OCO algorithms rely on the *meta-expert* framework, described by Van Erven and Koolen (2016), that runs several learning algorithms in parallel and combines them through appropriately-weighted averaging. There is not, unfortunately, a single way to take convex combinations in a geodesic metric space, and all such averaging schemes need to account for the curvature of the manifold and incorporate the associated costs. We finally find the *Fréchet mean* to be a desirable choice, but the analysis must proceed with care.

The key contributions of our work can be summarized as follows:

- We develop the optimistic mirror descent (OMD) algorithm on Hadamard manifolds[1] in the online improper learning setting. Interestingly, we also show Optimistic Hedge, a variant of OMD, works for gsc-convex losses. We believe these tools may have significant applications to research in online learning and Riemannian optimization.

- We combine our analysis on OMD with the *meta-expert framework* (Van Erven and Koolen, 2016) to get several adaptive regret bounds, as shown in Table 1.

- We develop a novel dynamic regret lower bound, which renders our $O(\sqrt{\zeta(1+P_T)T})$ bound to be tight up to the geometric constant ζ.

2. Preliminaries

In this section, we introduce background knowledge of OCO and Riemannian manifolds.

OCO in Euclidean space. We first formally describe OCO in Euclidean space. For each round $t = 1, \ldots, T$, the learner makes a decision $\mathbf{x}_t \in \mathcal{X}$ based on historical losses $f_1, \ldots, f_{t-1}$ where $\mathcal{X}$ is a convex decision set, and then the adversary reveals a convex loss function f_t. The goal of the

[1] We focus on Hadamard manifolds in the main paper and extend the guarantee to CAT(κ) spaces in Appendix B.2.

learner is to minimize the difference between the cumulative loss and that of the best-fixed decision in hindsight: $\text{Regret}_T = \sum_{t=1}^T f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^T f_t(\mathbf{x})$, which is usually referred to as the *static regret*, since the comparator is a fixed decision.

In the literature, there exist a large number of algorithms (Hazan et al., 2016) on minimizing the static regret. However, the underlying assumption of the static regret is the adversary's behavior does not change drastically, which can be unrealistic in real applications. To resolve this issue, dynamic regret stands out, which is defined as (Zinkevich, 2003)

$$\text{Regret}_T(\mathbf{u}_1, \ldots, \mathbf{u}_T) = \sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}_t),$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_T \in \mathcal{X}$ is a comparator sequence. Dynamic regret receives considerable attention recently (Besbes et al., 2015; Jadbabaie et al., 2015; Mokhtari et al., 2016; Zhang et al., 2017, 2018; Zhao et al., 2020; Wan et al., 2021; Zhao and Zhang, 2021; Baby and Wang, 2021) due to its flexibility. However, dynamic regret can be as large as $O(T)$ in general. Thus, regularizations need to be imposed on the comparator sequence to ensure no-regret online learning. A common assumption (Zinkevich, 2003) is the path-length (see Equation (1)) of the comparator sequence to be bounded. We refer to the corresponding dynamic regret as *general dynamic regret* because any assignments to $\mathbf{u}_1, \ldots, \mathbf{u}_T$ subject to the path-length constraint are feasible.

Riemannian manifolds. Here, we give a brief overview of Riemannian geometry, but this is a complex subject, and we refer the reader to, e.g., Petersen (2006) for a full treatment. A Riemannian manifold $(\mathcal{M}, g)$ is a smooth manifold $\mathcal{M}$ equipped with a Riemannian metric g. The tangent space $T_{\mathbf{x}}\mathcal{M} \cong \mathbb{R}^d$, generalizing the concept of the tangent plane, contains vectors tangent to any curve passing $\mathbf{x}$. The Riemannian metric g induces the inner product $\langle \mathbf{u}, \mathbf{v}\rangle_{\mathbf{x}}$ and the Riemannian norm $\|\mathbf{u}\|_{\mathbf{x}} = \sqrt{\langle \mathbf{u}, \mathbf{u}\rangle_{\mathbf{x}}}$ where $\mathbf{u}, \mathbf{v} \in T_{\mathbf{x}}\mathcal{M}$ (we omit the reference point $\mathbf{x}$ when it is clear from the context). We use $d(\mathbf{x}, \mathbf{y})$ to denote the Riemannian distance between $\mathbf{x}, \mathbf{y} \in \mathcal{M}$, which is the greatest lower bound of the length of all piecewise smooth curves joining $\mathbf{x}$ and $\mathbf{y}$.

A curve connecting $\mathbf{x}, \mathbf{y} \in \mathcal{M}$ is a geodesic if it is locally length-minimizing. For two points $\mathbf{x}, \mathbf{y} \in \mathcal{M}$, suppose there exists a geodesic $\gamma(t) : [0, 1] \to \mathcal{M}$ such that $\gamma(0) = \mathbf{x}, \gamma(1) = \mathbf{y}$ and $\gamma'(0) = \mathbf{v} \in T_{\mathbf{x}}\mathcal{M}$. The exponential map $\text{Exp}_{\mathbf{x}}(\cdot) : T_{\mathbf{x}}\mathcal{M} \to \mathcal{M}$ maps $\mathbf{v} \in T_{\mathbf{x}}\mathcal{M}$ to $\mathbf{y} \in \mathcal{M}$. Correspondingly, the inverse exponential map $\text{Exp}_{\mathbf{x}}^{-1}(\cdot) : \mathcal{M} \to T_{\mathbf{x}}\mathcal{M}$ maps $\mathbf{y} \in \mathcal{M}$ to $\mathbf{v} \in T_{\mathbf{x}}\mathcal{M}$. Since traveling along a geodesic is of constant velocity, we indeed have $d(\mathbf{x}, \mathbf{y}) = \|\text{Exp}_{\mathbf{x}}^{-1}\mathbf{y}\|_{\mathbf{x}}$. Also, it is useful to compare tangent vectors in different tangent spaces. Parallel transport $\Gamma_{\mathbf{x}}^{\mathbf{y}}\mathbf{u}$ translates $\mathbf{u}$ from $T_{\mathbf{x}}\mathcal{M}$ to $T_{\mathbf{y}}\mathcal{M}$ and preserves the inner product, i.e., $\langle \mathbf{u}, \mathbf{v}\rangle_{\mathbf{x}} = \langle \Gamma_{\mathbf{x}}^{\mathbf{y}}\mathbf{u}, \Gamma_{\mathbf{x}}^{\mathbf{y}}\mathbf{v}\rangle_{\mathbf{y}}$.

The curvature of Riemannian manifolds reflects the extent to which the manifold differs from a Euclidean surface. For optimization purposes, it usually suffices to consider the sectional curvature. Following Zhang and Sra (2016); Wang et al. (2021), in this paper we mainly consider Hadamard manifolds, which are complete and single-connected manifolds with non-positive sectional curvature. On such manifolds, every two points are connected by a unique and distance-minimizing geodesic by Hopf-Rinow Theorem (Petersen, 2006).

A subset $\mathcal{N}$ of $\mathcal{M}$ is gsc-convex if for any $\mathbf{x}, \mathbf{y} \in \mathcal{N}$, there exists a geodesic connecting $\mathbf{x}, \mathbf{y}$ and fully lies in $\mathcal{N}$. A function $f : \mathcal{N} \to \mathbb{R}$ is gsc-convex if $\mathcal{N}$ is gsc-convex and the composition $f(\gamma(t))$ satisfies $f(\gamma(t)) \le (1-t)f(\gamma(0)) + tf(\gamma(1))$ for any geodesic $\gamma(t) \subseteq \mathcal{N}$ and $t \in [0, 1]$. An alternative definition of geodesic convexity is

$$f(\mathbf{y}) \ge f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \text{Exp}_{\mathbf{x}}^{-1}\mathbf{y}\rangle, \qquad \forall\, \mathbf{x}, \mathbf{y} \in \mathcal{N}.$$

where the Riemannian gradient $\nabla f(\mathbf{x}) \in T_{\mathbf{x}}\mathcal{M}$ is the unique vector determined by $Df(\mathbf{x})[\mathbf{v}] = \langle \mathbf{v}, \nabla f(\mathbf{x}) \rangle$ and $Df(\mathbf{x})[\mathbf{v}]$ is the differential of f along $\mathbf{v} \in T_{\mathbf{x}}\mathcal{M}$.

Similarly, a L-geodesically-smooth (L-gsc-smooth) function f satisfies $\|\Gamma_{\mathbf{y}}^{\mathbf{x}} \nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\| \leq L \cdot d(\mathbf{x}, \mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{N}$, or

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathrm{Exp}_{\mathbf{x}}^{-1} \mathbf{y} \rangle + \tfrac{L}{2} d(\mathbf{x}, \mathbf{y})^2.$$

3. Related Work

In this part, we briefly review past work on OCO in Euclidean space, online optimization and optimism on Riemannian manifolds.

3.1. OCO in Euclidean Space

Static regret. We first consider work on static regret. In Euclidean space, it is well known that online gradient descent (OGD) guarantees $O(\sqrt{T})$ and $O(\log T)$ regret for convex and strongly convex losses (Hazan et al., 2016), which are also minimax optimal (Abernethy et al., 2008). However, the aforementioned bounds are not fully adaptive due to the dependence on T. Therefore, there is a tendency to replace T with problem-dependent quantities. Srebro et al. (2010) first notice that the smooth and non-negative losses satisfy the self-bounding property, thus establishing the small-loss bound $O(\sqrt{F_T^\star})$ where $F_T^\star = \sum_{t=1}^T f_t(\mathbf{x}^\star)$ is the cumulative loss of the best action in hindsight. Chiang et al. (2012) propose extra-gradient to get $O(\sqrt{V_T})$ gradient-variation regret bound for convex and smooth losses where $V_T = \sum_{t=2}^T \sup_{\mathbf{x} \in \mathcal{X}} \|\nabla f_{t-1}(\mathbf{x}) - \nabla f_t(\mathbf{x})\|_2^2$. Rakhlin and Sridharan (2013) generalize the work of Chiang et al. (2012) and propose optimistic mirror descent, which has become a standard tool in online learning since then.

Dynamic regret. Now we switch to the related work on dynamic regret. Zinkevich (2003) propose to use OGD to get a $O\left(\eta T + \frac{1+P_T}{\eta}\right)$ regret bound where η is the step size, but the result turns out to be $O((1+P_T)\sqrt{T})$ since the value of P_T is unknown to the learner. The seminal work of Zhang et al. (2018) use Hedge to combine the advice of experts with different step sizes, and show a $O\left(\sqrt{(1+P_T)T}\right)$ regret. A matching lower bound is also established therein. Zhao et al. (2020) utilize smoothness to get a gradient-variation bound, a small-loss bound, and a best-of-both-worlds bound in Euclidean space.

3.2. Online Learning and Optimism on Riemannian Manifolds

In the online setting, Bécigneul and Ganea (2019) consider adaptive stochastic optimization on Riemannian manifolds but their results only apply to the Cartesian product of one-manifolds. Maass et al. (2022) study the *restricted dynamic regret* on Hadamard manifolds under the gradient-free setting and provide $O(\sqrt{T} + P_T^\star)$ bound for gsc-strongly convex and gsc-smooth functions, where $P_T^\star$ is the path-length of the comparator formed by $\mathbf{u}_t = \mathrm{argmin}_{\mathbf{x} \in \mathcal{X}} f_t(\mathbf{x})$. On Hadamard manifolds, Wang et al. (2021) apply Riemannian OGD (R-OGD) to get $O(\sqrt{T})$ upper bound and $\Omega(\sqrt{T})$ randomized lower bound. Comparatively, we focus on general and adaptive dynamic regret on Hadamard manifolds. Our minimax lower bound is also novel.

There also exist algorithms considering optimism on Riemannian manifolds. Zhang et al. (2022) propose Riemannian Corrected Extra Gradient (RCEG) for unconstrained minimax optimization on manifolds. Karimi et al. (2022) consider a Robbins-Monro framework on Hadamard manifolds which subsumes Riemannian stochastic extra-gradient. By imposing the weakly asymptotically coercivity and using a decaying step size, the trajectory is guaranteed to be finite (Karimi et al., 2022). However, our paper is the first to consider the constrained case and the online setting. For the improper learning setting, we show a constant step size achieves the same guarantee as in Euclidean space.

4. Path Length Dynamic Regret Bound on Manifolds

In this section, we present the results related to the minimax path-length bound on manifolds. Before diving into the details, following previous work (Zinkevich, 2003; Wang et al., 2021), we introduce some standard assumptions and definitions.

Assumption 1 *$\mathcal{M}$ is a Hadamard manifold and its sectional curvature is lower bounded by $\kappa \leq 0$.*

Assumption 2 *The decision set $\mathcal{N}$ is a gsc-convex compact subset of $\mathcal{M}$ with diameter upper bounded by D, i.e., $\sup_{\mathbf{x},\mathbf{y}\in\mathcal{N}} d(\mathbf{x},\mathbf{y}) \leq D$. For optimistic online learning, we allow the player chooses decisions from $\mathcal{N}_{\delta M}$, which is defined in Definition 2 and the diameter becomes $(D+2\delta M)$.*

Assumption 3 *The norm of Riemannian gradients are bounded by G, i.e., $\sup_{\mathbf{x}\in\mathcal{N}} \|\nabla f_t(\mathbf{x})\| \leq G$. When improper learning is allowed, we assume $\sup_{\mathbf{x}\in\mathcal{N}_{\delta M}} \|\nabla f_t(\mathbf{x})\| \leq G$.*

Definition 1 *Under Assumptions 1, 2, we denote $\zeta := \sqrt{-\kappa}D\coth(\sqrt{-\kappa}D)$. When improper learning is allowed, $\zeta := \sqrt{-\kappa}(D+2\delta M)\coth(\sqrt{-\kappa}(D+2\delta M))$, where M is in Definition 2. Note that, on manifolds of zero sectional curvature ($\kappa = 0$), we have $\zeta = \lim_{x\to 0} x\cdot\coth x = 1$.*

The seminal work of Zinkevich (2003) shows that the classical OGD algorithm can minimize the general dynamic regret in Euclidean space. Motivated by this, we consider the Riemannian OGD (R-OGD) algorithm (Wang et al., 2021):

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{N}}\mathrm{Exp}_{\mathbf{x}_t}(-\eta\nabla f_t(\mathbf{x}_t)), \tag{2}$$

which is a natural extension of OGD to the manifold setting. We show R-OGD can also minimizes the general dynamic regret on manifolds. Due to page limitation, we postpone details to Appendix A.

Theorem 1 *Suppose Assumptions 1, 2 and 3 hold. Then the general dynamical regret of R-OGD defined in Equation (2) satisfies*

$$\textit{D-Regret}_T \leq \frac{D^2 + 2DP_T}{2\eta} + \frac{\eta\zeta G^2 T}{2}. \tag{3}$$

Theorem 1 implies that R-OGD yields $O(\frac{P_T+1}{\eta} + \eta T)$ general dynamic regret bound, which means the optimal step size is $\eta = O\left(\sqrt{\frac{1+P_T}{T}}\right)$. However, this configuration of η is invalid, as P_T is unknown to the learner. Although a sub-optimal choice for η, i.e., $\eta = O\left(\frac{1}{\sqrt{T}}\right)$, is accessible, the resulting algorithm suffers $O((1+P_T)\sqrt{T})$ regret.

The meta-expert framework (Van Erven and Koolen, 2016) consists of a meta algorithm and some expert algorithm instances. The constructions are modular such that we can use different meta algorithms and expert algorithms to achieve different regret guarantees. For optimizing dynamic regret, the seminal work of Zhang et al. (2018) propose Ader based on this framework. In every round t, each expert runs OGD with a different step size, and the meta algorithm applies Hedge to learn the best weights. The step sizes used by the experts are carefully designed so that there always exists an expert which is almost optimal. The regret of Ader is $O(\sqrt{(1+P_T)T})$, which is minimax-optimal in Euclidean space (Zhang et al., 2018).

However, it is unclear how to extend Ader to manifolds at first glance since we need to figure out the "correct" way to do averaging. In this paper, we successfully resolve this problem using the *Fréchet mean* and the *geodesic mean*. Our proposed algorithm, called RADAR, consists of N instances of the expert algorithm (Algorithm 2), each of which runs R-OGD with a different step size, and a meta algorithm (Algorithm 1), which enjoys a regret approximately the same as the best expert. We denote the set of all step sizes $\{\eta_i\}$ by $\mathcal{H}$. In the t-th round, the expert algorithms submit all $\mathbf{x}_{t,i}$'s ($i=1,\ldots,N$) to the meta algorithm. Then the meta algorithm either computes the Fréchet mean or the geodesic mean (see Algorithm 6 in Appendix E for details) as $\mathbf{x}_t$. After receiving f_t, the meta algorithm updates the weight of each expert $w_{t+1,i}$ via Hedge and sends $\nabla f_t(\mathbf{x}_{t,i})$ to the i-th expert, which computes $\mathbf{x}_{t+1,i}$ by R-OGD. The regret of the meta algorithm of RADAR can be bounded by Lemma 1.

```
Algorithm 1: RADAR: Meta Algorithm
Data: Learning rate $\beta$, set of step sizes $\mathcal{H}$,
      initial weights $w_{1,i} = \frac{N+1}{i(i+1)N}$
for $t = 1, \ldots, T$ do
    Receive $\mathbf{x}_{t,i}$ from experts with stepsize $\eta_i$
    $\mathbf{x}_t = \mathrm{argmin}_{\mathbf{x}\in\mathcal{N}} \sum_{i\in[N]} w_{t,i} d(\mathbf{x}, \mathbf{x}_{t,i})^2$
    Observe the loss function $f_t$
    Update $w_{t+1,i}$ by Hedge with $f_t(\mathbf{x}_{t,i})$
    Send gradient $\nabla f_t(\mathbf{x}_{t,i})$ to each expert
end
```

```
Algorithm 2: RADAR: Expert Algorithm
Data: A step size $\eta_i$
Let $\mathbf{x}^{\eta}_{1,i}$ be any point in $\mathcal{N}$
for $t = 1, \ldots, T$ do
    Submit $\mathbf{x}_{t,i}$ to the meta algorithm
    Receive gradient $\nabla f_t(\mathbf{x}_{t,i})$ from the
      meta algorithm
    Update:
    $\mathbf{x}_{t+1,i} = \Pi_{\mathcal{N}} \mathrm{Exp}_{\mathbf{x}_{t,i}}(-\eta_i \nabla f_t(\mathbf{x}_{t,i}))$
end
```

Lemma 1 *Under Assumptions 1, 2, 3, and setting $\beta = \sqrt{\frac{8}{G^2D^2T}}$, the regret of Algorithm 1 satisfies*

$$\sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{x}_{t,i}) \leq \sqrt{\frac{G^2D^2T}{8}} \left(1 + \ln \frac{1}{w_{1,i}}\right).$$

We show that, by configuring the step sizes in $\mathcal{H}$ carefully, RADAR ensures a $O(\sqrt{(1+P_T)T})$ bound on geodesic metric spaces.

Theorem 2 *Set $\mathcal{H} = \left\{\eta_i = 2^{i-1}\sqrt{\frac{D^2}{G^2\zeta T}} \middle| i \in [N]\right\}$ where $N = \lceil \frac{1}{2}\log_2(1+2T)\rceil + 1$ and $\beta = \sqrt{\frac{8}{G^2D^2T}}$. Under Assumptions 1, 2, 3, for any comparator sequence $\mathbf{u}_1, \ldots, \mathbf{u}_T \in \mathcal{N}$, the general dynamic regret of* RADAR *satisfies*

$$\textit{D-Regret}_T = O(\sqrt{\zeta(1+P_T)T}).$$

Remark 2 *Note that if $\mathcal{M}$ is Euclidean space, then $\zeta = 1$ and we get $O(\sqrt{(1+P_T)T})$ regret, which is the same as in Zhang et al. (2018).*

A disadvantage of Radar is $\Theta(\log T)$ gradient queries are required at each round. In Euclidean space, Zhang et al. (2018) use a linear surrogate loss to achieve the same bound by $O(1)$ gradient queries. But on manifolds, the existence of such functions implies the sectional curvature of the manifold is everywhere 0 (Kristály et al., 2016). It is interesting to investigate if $\Omega(\log T)$ gradient queries are necessary to achieve dynamic regret on manifolds. We would also like to point out that $O(\log T)$ is reasonable small and the work of Zhang et al. (2018) still needs $O(\log T)$ computational complexity per round.

Using the Busemann function as a bridge, we show the following dynamic regret lower bound, with proof deferred to Appendix A.4.

Theorem 3 *There exists a comparator sequence which satisfies $\sum_{t=2}^{T} d(\mathbf{u}_t, \mathbf{u}_{t-1}) \leq P_T$ and encounters $\Omega(\sqrt{(1+P_T)T})$ dynamic regret on Hadamard manifolds.*

Although the regret guarantee in Theorem 2 is optimal up to constants in terms of T and P_T by considering the corresponding lower bound, it still depends on T and thus cannot adapt to mild environments. In Euclidean space, the smoothness of losses induces adaptive regret bounds, including the gradient-variation bound (Chiang et al., 2012) and the small-loss bound (Srebro et al., 2010). It is then natural to ask if similar bounds can be established on manifolds by assuming gsc-smoothness. We provide an affirmative answer to this question and show how to get problem-dependent bounds under the Radar framework.

5. Gradient-variation Bound on Manifolds

In this section, we show how to obtain the gradient-variation bound on manifolds under the Radar framework with alternative expert and meta algorithms.

Expert Algorithm. For minimax optimization on Riemannian manifolds, Zhang et al. (2022) propose Riemannian Corrected Extra Gradient (RCEG), which performs the following iterates:

$$\begin{aligned} \mathbf{x}_t &= \mathrm{Exp}_{\mathbf{y}_t}(-\eta \nabla f_{t-1}(\mathbf{y}_t)) \\ \mathbf{y}_{t+1} &= \mathrm{Exp}_{\mathbf{x}_t}\left(-\eta \nabla f_t(\mathbf{x}_t) + \mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{y}_t)\right). \end{aligned}$$

However, this algorithm does not work in the constrained case, which has been left as an open problem (Zhang et al., 2022). The online improper learning setting (Hazan et al., 2018; Baby and Wang, 2021) allows the decision set to be different from (usually larger than) the set of strategies we want to compete against. Under such a setting, we find the geometric distortion due to projection can be bounded in an elegant way, and generalize RCEG to incorporate an optimism term $M_t \in T_{\mathbf{y}_t}\mathcal{M}$.

Definition 2 *We use M_t to denote the optimism at round t and assume there exists M such that $\|M_t\| \leq M$ for all t. We define $\mathcal{N}_c = \{\mathbf{x} | d(x, \mathcal{N}) \leq c\}$ where $d(\mathbf{x}, \mathcal{N}) := \inf_{\mathbf{y} \in \mathcal{N}} d(\mathbf{x}, \mathbf{y})$. In the improper setting, we allow the player to choose decisions from $\mathcal{N}_{\delta M}$.*

Theorem 4 *Suppose all losses f_t are L-gsc-smooth on $\mathcal{M}$. Under Assumptions 1, 2, 3, the iterates*

$$\begin{aligned} \mathbf{x}'_t &= \mathrm{Exp}_{\mathbf{y}_t}(-\eta M_t) \\ \mathbf{x}_t &= \Pi_{\mathcal{N}_{\delta M}} \mathbf{x}'_t \\ \mathbf{y}_{t+1} &= \Pi_{\mathcal{N}} \mathrm{Exp}_{\mathbf{x}'_t}\left(-\eta \nabla f_t(\mathbf{x}'_t) + \mathrm{Exp}_{\mathbf{x}'_t}^{-1}(\mathbf{y}_t)\right). \end{aligned} \tag{4}$$

satisfies

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}_t) \leq \eta\zeta \sum_{t=1}^{T} \|\nabla f_t(\mathbf{y}_t) - M_t\|^2 + \frac{D^2 + 2DP_T}{2\eta}.$$

for any $\mathbf{u}_1, \dots, \mathbf{u}_T \in \mathcal{N}$ and $\eta \leq \frac{\delta M}{G+(G^2+2\zeta\delta^2M^2L^2)^{\frac{1}{2}}}$. Specifically, we achieve $\eta\zeta V_T + \frac{D^2+2DP_T}{2\eta}$ regret by choosing $M_t = \nabla f_{t-1}(\mathbf{y}_t)$. In this case, $M = G$ and we need $\eta \leq \frac{\delta}{1+(1+2\zeta\delta^2L^2)^{\frac{1}{2}}}$.

Proof sketch. We use the special case $M_t = \nabla f_{t-1}(\mathbf{y}_t)$ to illustrate the main idea of the proof. We first decompose $f_t(\mathbf{x}_t) - f_t(\mathbf{u}_t)$ into two terms,

$$\begin{aligned} f_t(\mathbf{x}_t) - f_t(\mathbf{u}_t) &= (f_t(\mathbf{x}_t) - f_t(\mathbf{x}'_t)) + (f_t(\mathbf{x}'_t) - f_t(\mathbf{u}_t)) \\ &\leq G \cdot d(\mathbf{x}_t, \mathbf{x}'_t) + (f_t(\mathbf{x}'_t) - f_t(\mathbf{u}_t)) \leq \underbrace{G \cdot d(\mathbf{x}'_t, \mathbf{y}_t)}_{\text{troublesome term 1}} + \underbrace{(f_t(\mathbf{x}'_t) - f_t(\mathbf{u}_t))}_{\text{unconstrained RCEG}} \end{aligned}$$

where the first inequality is because the gradient Lipschitzness condition, and the second one follows from the non-expansiveness of the projection. For the unconstrained RCEG term, we have the following decomposition,

$$\begin{aligned} f_t(\mathbf{x}'_t) - f_t(\mathbf{u}_t) \leq& \underbrace{\frac{1}{2\eta}(2\eta^2\zeta L^2 - 1) d(\mathbf{x}'_t, \mathbf{y}_t)^2}_{\text{troublesome term 2}} \\ &+ \underbrace{\eta\zeta\|\nabla f_t(\mathbf{y}_t) - \nabla f_{t-1}(\mathbf{y}_t)\|^2}_{\eta\zeta V_T} + \underbrace{\frac{1}{2\eta}\left(d(\mathbf{y}_t, \mathbf{u}_t)^2 - d(\mathbf{y}_{t+1}, \mathbf{u}_t)^2\right)}_{\frac{D^2+2DP_T}{2\eta}} \end{aligned}$$

where the second and the third term corresponds to the gradient variation term and the dynamic regret term, respectively.

In the improper learning setting, we can show $d(\mathbf{x}'_t, \mathbf{y}_t) \geq \delta G$. Combining both troublesome terms, it suffices to find η which satisfies

$$2\eta G + 2\eta^2\zeta L^2\lambda - \lambda \leq 0, \ \forall\lambda := d(\mathbf{x}'_t, \mathbf{y}_t) \geq \delta G.$$

Remark 3 *We generalize Theorem 4 from Hadamard manifolds to CAT(κ) spaces in Appendix B.2. Note that although we allow the player to make improper decisions, V_T is still defined on $\mathcal{N}$ instead of $\mathcal{N}_{\delta G}$. For the static setting, $P_T = 0$ and the resulting regret bound is $O(\sqrt{V_T} + \frac{1}{\delta})$. Also, in this setting, we can use an adaptive step-size*

$$\eta_t = \min\left\{\frac{1}{\sqrt{1 + \sum_{s=2}^{t} \|\nabla f_t(\mathbf{y}_t) - \nabla f_{t-1}(\mathbf{y}_t)\|^2}}, \frac{\delta}{1 + (1 + 2\zeta\delta^2L^2)^{\frac{1}{2}}}\right\}$$

to eliminate the dependence on V_T.

Meta algorithm. Intuitively, we can run OMD with different step sizes and apply a meta algorithm to estimate the optimal step size. Previous studies in learning with multiple step sizes usually adopt Hedge to aggregate the experts' advice. However, the regret of Hedge is $O(\sqrt{T \ln N})$ and thus is undesirable for our purpose. Inspired by optimistic online learning (Rakhlin and Sridharan, 2013; Syrgkanis et al., 2015), Zhao et al. (2020) adopt Optimistic Hedge as the meta algorithm to get $O(\sqrt{(V_T + P_T)P_T})$ gradient-variation bound. After careful analysis, we show Optimistic Hedge works for gsc-convex losses regardless of the geometric distortion and get the desired gradient-variation bound.

```
Algorithm 3: RADAR_v: Expert Algorithm
Data: A step size $\eta_i$
Let $\mathbf{x}_{1,i}^{\eta}$ be any point in $\mathcal{N}$
for $t = 1, \dots, T$ do
    Submit $\mathbf{x}_{t,i}$ to the meta algorithm
    Receive gradient $\nabla f_t(\cdot)$ from the meta algorithm
    Each expert runs Equation (4) with $M_t = \nabla f_{t-1}(\mathbf{y}_t)$, $M = G$ and step size $\eta_i$
end
```

```
Algorithm 4: RADAR_v: Meta Algorithm
Data: A learning rate $\beta$, a set of step sizes $\mathcal{H}$, initial weights $w_{1,i} = w_{0,i} = \frac{1}{N}$
for $t = 1, \dots, T$ do
    Receive all $\mathbf{x}_{t,i}$'s from experts with step size $\eta_i$
    $\bar{\mathbf{x}}_t = \mathrm{argmin}_{\mathbf{x} \in \mathcal{N}_{\delta G}} \sum_{i \in [N]} w_{t-1,i} d(\mathbf{x}, \mathbf{x}_{t,i})^2$
    Update $w_{t,i} \propto \exp\left(-\beta \left(\sum_{s=1}^{t-1} \ell_{s,i} + m_{t,i}\right)\right)$ by Equation (5)
    $\mathbf{x}_t = \mathrm{argmin}_{\mathbf{x} \in \mathcal{N}_{\delta G}} \sum_{i \in [N]} w_{t,i} d(\mathbf{x}, \mathbf{x}_{t,i})^2$
    Observe $f_t(\cdot)$ and send $\nabla f_t(\cdot)$ to experts
end
```

We denote $\boldsymbol{\ell}_t, \mathbf{m}_t \in \mathbb{R}^N$ as the surrogate loss and the optimism at round t. The update rule of Optimistic Hedge is:

$$w_{t,i} \propto \exp\left(-\beta \left(\textstyle\sum_{s=1}^{t-1} \ell_{s,i} + m_{t,i}\right)\right),$$

which achieves adaptive regret due to the optimism. The following technical lemma (Syrgkanis et al., 2015) is critical for our analysis of Optimistic Hedge, and the proof is in Appendix B.3 for completeness.

Lemma 4 *For any $i \in [N]$, Optimistic Hedge satisfies*

$$\textstyle\sum_{t=1}^{T} \langle \mathbf{w}_t, \boldsymbol{\ell}_t \rangle - \ell_{t,i} \leq \frac{2 + \ln N}{\beta} + \beta \sum_{t=1}^{T} \|\boldsymbol{\ell}_t - \mathbf{m}_t\|_\infty^2 - \frac{1}{4\beta} \sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_1^2$$

Following the insightful work of Zhao et al. (2020), we also adopt the Optimistic Hedge algorithm as the meta algorithm, but there are some key differences in the design of the surrogate loss and optimism. To respect the Riemannian metric, we propose the following:

$$\begin{aligned} \ell_{t,i} &= \left\langle \nabla f_t(\mathbf{x}_t), \mathrm{Exp}_{\mathbf{x}_t}^{-1} \mathbf{x}_{t,i} \right\rangle \\ m_{t,i} &= \left\langle \nabla f_{t-1}(\bar{\mathbf{x}}_t), \mathrm{Exp}_{\bar{\mathbf{x}}_t}^{-1} \mathbf{x}_{t,i} \right\rangle \end{aligned} \tag{5}$$

where $\mathbf{x}_t$ and $\bar{\mathbf{x}}_t$ are Fréchet averages of $\mathbf{x}_{t,i}$ w.r.t. linear combination coefficients $\mathbf{w}_t$ and $\mathbf{w}_{t-1}$ respectively. Under the Fréchet mean, we can show

$$f_t(\mathbf{x}_t) - f_t(\mathbf{x}_{t,i}) \leq \langle \mathbf{w}_t, \boldsymbol{\ell}_t \rangle - \ell_{t,i},$$

which ensures Lemma 4 can be applied to bound the meta-regret and the geodesic mean does not meet this requirement. We also emphasize that the design of the surrogate loss and optimism is highly non-trivial. As we will see in the proof of Theorem 5, the combination of the surrogate loss and the gradient-vanishing property of the Fréchet mean ensures Lemma 4 can be invoked to upper bound the regret of the meta algorithm. However, $\mathbf{m}_t$ cannot rely on $\mathbf{x}_t$ thus, we need to design optimism based on the tangent space of $\bar{\mathbf{x}}_t$, which incurs extra cost. Luckily, under Equation (5), we find a reasonable upper bound of this geometric distortion by showing

$$\begin{aligned}\|\boldsymbol{\ell}_t - \mathbf{m}_t\|_\infty^2 &\leq O(1) \cdot \sup_{\mathbf{x} \in \mathcal{N}_{\delta G}} \|\nabla f_t(\mathbf{x}) - \nabla f_{t-1}(\mathbf{x})\|^2 + O(1) \cdot d(\mathbf{x}_t, \bar{\mathbf{x}}_t)^2 \\ &\leq O(1) \cdot \sup_{\mathbf{x} \in \mathcal{N}_{\delta G}} \|\nabla f_t(\mathbf{x}) - \nabla f_{t-1}(\mathbf{x})\|^2 + \tilde{O}(1) \cdot \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_1^2.\end{aligned}$$

Thus we can apply the negative term in Lemma 4 to eliminate undesired terms in $\|\boldsymbol{\ell}_t - \mathbf{m}_t\|_\infty^2$.

Algorithms 3 and 4 describe the expert algorithm and meta algorithm of Radar$_\text{v}$. We show the meta-regret and total regret of Radar$_\text{v}$ in Theorems 5 and 6, respectively. Detailed proof in this section is deferred to Appendix B.

Theorem 5 *Assume all losses are L-gsc-smooth on $\mathcal{M}$. Then under Assumptions 1, 2, 3, the regret of Algorithm 4 satisfies:*

$$\begin{aligned}&\sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{x}_{t,i}) \\ \leq &\frac{2 + \ln N}{\beta} + 3D^2\beta(V_T + G^2) + \sum_{t=2}^T \left(3\beta(D^4L^2 + D^2G^2\zeta^2) - \frac{1}{4\beta}\right) \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_1^2.\end{aligned}$$

Theorem 6 *Let $\beta = \min\left\{\sqrt{\frac{2+\ln N}{3D^2V_T}}, \frac{1}{\sqrt{12(D^4L^2+D^2G^2\zeta^2)}}\right\}$, $\mathcal{H} = \left\{\eta_i = 2^{i-1}\sqrt{\frac{D^2}{8\zeta G^2 T}} \middle| i \in [N]\right\}$, where $N = \left\lceil \frac{1}{2} \log \frac{8\zeta\delta^2G^2T}{(1+(1+2\zeta\delta^2L^2)^{\frac{1}{2}})^2} \right\rceil + 1$. Assume all losses are L-gsc-smooth on $\mathcal{M}$ and allow improper learning. Under Assumptions 1, 2 and 3, the regret of* Radar$_\text{v}$ *satisfies*

$$\textit{D-Regret}_T = O\left(\sqrt{\zeta(V_T + (1 + P_T)/\delta^2)(1 + P_T)}\right).$$

In Theorem 6, β relies on V_T, and this dependence can be eliminated by showing a variant of Lemma 4 with an adaptive learning rate β_t.

6. Small-loss Bound on Manifolds

For dynamic regret, the small-loss bound replaces the dependence on T by $F_T = \sum_{t=1}^T f_t(\mathbf{u}_t)$, which adapts to the function values of the comparator sequence. In Euclidean space, Srebro et al.

(2010) show this adaptive regret by combining OGD with the self-bounding property of smooth and non-negative functions, which reads $\|\nabla f(\mathbf{x})\|_2^2 \leq 4L \cdot f(\mathbf{x})$ where L is the smoothness constant. We show a similar conclusion on manifolds and defer proof details in this part to Appendix C.

Lemma 5 *Suppose $f : \mathcal{M} \to \mathbb{R}$ is both L-gsc-smooth and non-negative on its domain where $\mathcal{M}$ is a Hadamard manifold, then we have $\|\nabla f(\mathbf{x})\|^2 \leq 2L \cdot f(\mathbf{x})$.*

To facilitate the discussion, we denote $\bar{F}_T = \sum_{t=1}^T f_t(\mathbf{x}_t)$ and $\bar{F}_{T,i} = \sum_{t=1}^T f_t(\mathbf{x}_{t,i})$. We use R-OGD as the expert algorithm (Algorithm 2) and Hedge with surrogate loss

$$\ell_{t,i} = \left\langle \nabla f_t(\mathbf{x}_t), \text{Exp}_{\mathbf{x}_t}^{-1} \mathbf{x}_{t,i} \right\rangle$$

as the meta algorithm (Algorithm 1). The following Lemma considers the regret of a single expert and shows that R-OGD achieves a small-loss dynamic regret on geodesic metric spaces.

Lemma 6 *Suppose all losses are L-gsc-smooth and non-negative on $\mathcal{M}$. Under Assumptions 1, 2, by choosing any step size $\eta \leq \frac{1}{2\zeta L}$, R-OGD achieves $O\left(\frac{P_T}{\eta} + \eta F_T\right)$ regret.*

Again, we can not directly set $\eta = O\left(\frac{1+P_T}{F_T}\right)$ because P_T is unknown, which is precisely why we need the meta algorithm. The meta-regret of Hedge is as follows.

Lemma 7 *Suppose all losses are L-gsc-smooth and non-negative on $\mathcal{M}$. Under Assumptions 1, 2, by setting the learning rate of Hedge as $\beta = \sqrt{\frac{(2+\ln N)}{D^2 \bar{F}_T}}$, the regret of the meta algorithm satisfies*

$$\textstyle\sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{x}_{t,i}) \leq 8D^2 L(2 + \ln N) + \sqrt{8D^2 L(2 + \ln N) F_{T,i}}.$$

Now as we have the guarantee for both the expert algorithm and the meta algorithm, a direct combination yields the following general dynamic regret guarantee.

Theorem 7 *Suppose all losses are L-gsc-smooth and non-negative on $\mathcal{M}$. Under Assumptions 1, 2. Setting $\mathcal{H} = \left\{\eta_i = 2^{i-1}\sqrt{\frac{D}{2\zeta LGT}} \middle| i \in [N]\right\}$ where $N = \lceil \frac{1}{2} \log \frac{GT}{2LD\zeta} \rceil + 1$ and $\beta = \sqrt{\frac{(2+\ln N)}{D^2 \bar{F}_T}}$. Then for any comparator $\mathbf{u}_1, \ldots, \mathbf{u}_T \in \mathcal{N}$, we have*

$$\textit{D-Regret}_T = O(\sqrt{\zeta(\zeta(P_T + 1) + F_T)(P_T + 1)}).$$

Remark 8 *If we take $\mathcal{M}$ as Euclidean space, the regret guarantee shown in Theorem 7 becomes $O(\sqrt{(P_T + F_T + 1)(P_T + 1)})$, which matches the result of Zhao et al. (2020).*

7. Best-of-both-worlds Bound on Manifolds

Now we already achieve the gradient-variation bound and the small-loss bound on manifolds. To highlight the differences between them, we provide an example in Appendix D.1 to show under certain scenarios, one bound can be much tighter than the other. The next natural question is, is that possible to get a best-of-both-worlds bound on manifolds?

We initialize $N := N^v + N^s$ experts as shown in Theorems 6 and 7 where N^v and N^s are the numbers of experts running OMD and R-OGD, respectively. For each expert $i \in [N]$, the surrogate loss and the optimism are

$$\begin{aligned} \ell_{t,i} &= \left\langle \nabla f_t(\mathbf{x}_t), \mathrm{Exp}^{-1}_{\mathbf{x}_t}\mathbf{x}_{t,i} \right\rangle \\ m_{t,i} &= \gamma_t \left\langle \nabla f_{t-1}(\bar{\mathbf{x}}_t), \mathrm{Exp}^{-1}_{\bar{\mathbf{x}}_t}\mathbf{x}_{t,i} \right\rangle . \end{aligned} \tag{6}$$

γ_t controls the optimism used in the meta algorithm. When $\gamma_t = 1$, the optimism for the gradient-variation bound is recovered, and $\gamma_t = 0$ corresponds to the optimism for the small-loss bound.

Following Zhao et al. (2020), we use Hedge for two experts to get a best-of-both-worlds bound. The analysis therein relies on the strong convexity of $\|\nabla f_t(\mathbf{x}_t) - \mathbf{m}\|_2^2$ in $\mathbf{m}$, which is generally not the case on manifolds. So an alternative scheme needs to be proposed. We denote

$$\begin{aligned} m^v_{t,i} &= \left\langle \nabla f_{t-1}(\bar{\mathbf{x}}_t), \mathrm{Exp}^{-1}_{\bar{\mathbf{x}}_t}\mathbf{x}_{t,i} \right\rangle \\ m^s_{t,i} &= 0, \end{aligned} \tag{7}$$

while $\mathbf{m}_t^v$ and $\mathbf{m}_t^s$ be the corresponding vectors respectively. Then $\mathbf{m}_t = \gamma_t \mathbf{m}_t^v + (1-\gamma_t)\mathbf{m}_t^s$, which is exactly the combination rule of Hedge. The function $\|\boldsymbol{\ell}_t - \mathbf{m}\|_\infty^2$ is convex with respect to $\mathbf{m}$ but not strongly convex so we instead use $d_t(\mathbf{m}) := \|\boldsymbol{\ell}_t - \mathbf{m}\|_2^2$ for Hedge, and the learning rate is updated as

$$\gamma_t = \frac{\exp\left(-\tau\left(\sum_{r=1}^{t-1} d_r(\mathbf{m}_r^v)\right)\right)}{\exp\left(-\tau\left(\sum_{r=1}^{t-1} d_r(\mathbf{m}_r^v)\right)\right) + \exp\left(-\tau\left(\sum_{r=1}^{t-1} d_r(\mathbf{m}_r^s)\right)\right)} \tag{8}$$

Algorithm 5 summarizes the meta algorithm as well as the expert algorithm for RADAR$_\text{b}$.

Algorithm 5: RADAR$_\text{b}$: Algorithm

```
Data: Learning rates $\beta$ for Optimistic Hedge and $\gamma_t$ for Hedging the two experts, $\mathcal{H} = \{\eta_i\}$
      consists of $N = N^v + N^s$ step sizes, $\tau = \frac{1}{8NG^2D^2}$
for $t = 1, \dots, T$ do
    Run Algorithms 3 and 2 on the first $N^v$ experts and the later $N^s$ experts, resp.
    $\bar{\mathbf{x}}_t = \mathrm{argmin}_{\mathbf{x} \in \mathcal{N}_{\delta G}} \sum_{i \in [N]} w_{t-1,i} d(\mathbf{x}, \mathbf{x}_{t,i})^2$
    Update $\gamma_t$ as in Equation (8)
    Update $w_{t,i} \propto \exp\left(-\beta\left(\sum_{s=1}^{t-1} \ell_{s,i} + m_{t,i}\right)\right)$ by Equation (6)
    $\mathbf{x}_t = \mathrm{argmin}_{\mathbf{x} \in \mathcal{N}_{\delta G}} \sum_{i \in [N]} w_{t,i} d(\mathbf{x}, \mathbf{x}_{t,i})^2$
    Observe $f_t$ and send $\nabla f_t(\cdot)$ to each expert
end
```

In Theorem 8 we show the guarantee of the meta algorithm of RADAR$_\text{b}$ and postpone proof details of this section to Appendix D.

Theorem 8 *Setting learning rates* $\tau = \frac{1}{8NG^2D^2}$ *and*

$$\beta = \min\left\{ \sqrt{\frac{2 + \ln N}{N(D^2 \min\{3(V_T + G^2), \bar{F}_T\} + 8G^2D^2 \ln 2)}}, \frac{1}{\sqrt{12(D^4L^2 + D^2G^2\zeta^2)}} \right\}.$$

Suppose all losses are L-gsc-smooth and non-negative on $\mathcal{M}$. Under Assumptions 1, 2, 3, the regret of the meta algorithm satisfies

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_{t,i}) = O\left(\sqrt{\ln T \min\{V_T, \bar{F}_T\}}\right)$$

where $N = N^v + N^s$ and $\bar{F}_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t)$.

Finally, we show the regret of $\textsc{Radar}_{\text{b}}$ is bounded by the smaller of two problem-dependent bounds as follows.

Theorem 9 *Suppose all losses are L-gsc-smooth and non-negative on $\mathcal{M}$ and allow improper learning. Under Assumptions 1, 2, 3, if we set the set of candidate step sizes as*

$$\mathcal{H} = \mathcal{H}^v \cup \mathcal{H}^s, \tag{9}$$

where $\mathcal{H}^v$ and $\mathcal{H}^s$ are sets of step sizes in Theorem 6 with $N = N^v$ and Theorem 7 with $N = N^s$ respectively. Then Algorithm 5 satisfies

$$\textit{D-Regret}_T = O\left(\sqrt{\zeta(P_T(\zeta + 1/\delta^2) + B_T + 1)(1 + P_T) + \ln T \cdot B_T}\right)$$

where $B_T := \min\{V_T, F_T\}$.

Remark 9 *Comparing to the result in Zhao et al. (2020), we find the result of Theorem 9 has an additional $\sqrt{\ln T}$ factor, which comes from our construction of hedging two experts. It will be interesting to remove this dependence.*

8. Conclusion

In this paper, we consider adaptive online learning on Riemannian manifolds. Equipped with the idea of learning with multiple step sizes and optimistic mirror descent, we derive a series of no-regret algorithms that adapt to quantities reflecting the intrinsic difficulty of the online learning problem in different aspects. In the future, it is interesting to investigate how to achieve optimistic online learning in the proper learning setting. Moving forward, one could further examine whether $\Omega(\log T)$ gradient queries in each round are truly necessary. A curvature-dependent lower bound like the one in Criscitiello and Boumal (2022) for Riemannian online optimization also remains open.

Acknowledgments

We would like to thank three anonymous referees for constructive comments and suggestions. We gratefully thank the AI4OPT Institute for funding, as part of NSF Award 2112533. We also acknowledge the NSF for their support through Award IIS-1910077. GW would like to acknowledge an ARC-ACO fellowship provided by Georgia Tech.

REFERENCES

Jacob Abernethy, Peter L Bartlett, Alexander Rakhlin, and Ambuj Tewari. Optimal strategies and minimax lower bounds for online convex games. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 415–423, 2008.

Kwangjun Ahn and Suvrit Sra. From nesterov's estimate sequence to riemannian acceleration. In *Proceedings of the 33rd Annual Conference on Learning Theory*, pages 84–118. PMLR, 2020.

Foivos Alimisis, Antonio Orvieto, Gary Bécigneul, and Aurelien Lucchi. A continuous-time perspective for modeling acceleration in riemannian optimization. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pages 1297–1307. PMLR, 2020.

Dheeraj Baby and Yu-Xiang Wang. Optimal dynamic regret in exp-concave online learning. In *Proceedings of 34th Conference on Learning Theory*, pages 359–409. PMLR, 2021.

Miroslav Bacák. Computing medians and means in hadamard spaces. *SIAM journal on optimization*, 24(3):1542–1566, 2014a.

Miroslav Bacák. *Convex analysis and optimization in Hadamard spaces*. de Gruyter, 2014b.

Werner Ballmann. *Lectures on spaces of nonpositive curvature*, volume 25. Birkhäuser, 2012.

Gary Bécigneul and Octavian-Eugen Ganea. Riemannian adaptive optimization methods. In *7th International Conference on Learning Representations*, 2019.

GC Bento, JX Neto, and IDL Melo. Elements of convex geometry in hadamard manifolds with application to equilibrium problems. *arXiv preprint arXiv:2107.02223*, 2021.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244, 2015.

Rajendra Bhatia. Positive definite matrices. In *Positive Definite Matrices*. Princeton university press, 2009.

Martin R Bridson and André Haefliger. *Metric spaces of non-positive curvature*, volume 319. Springer Science & Business Media, 2013.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *Proceedings of the 25th Annual Conference on Learning Theory*, pages 6.1–6.20, 2012.

Christopher Criscitiello and Nicolas Boumal. Negative curvature obstructs acceleration for strongly geodesically convex optimization, even with exact first-order oracles. In *Proceedings of the 35th Annual Conference on Learning Theory*, pages 496–542. PMLR, 2022.

John C Duchi, Alekh Agarwal, and Martin J Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57(3): 592–606, 2012.

Elad Hazan, Wei Hu, Yuanzhi Li, and Zhiyuan Li. Online improper learning with an approximation oracle. *Advances in Neural Information Processing Systems*, 31, 2018.

Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Reshad Hosseini and Suvrit Sra. Matrix manifold optimization for gaussian mixtures. *In Advances in Neural Information Processing Systems 29*, 28:910–918, 2015.

Ali Jadbabaie, Alexander Rakhlin, Shahin Shahrampour, and Karthik Sridharan. Online optimization: Competing with dynamic comparators. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pages 398–406. PMLR, 2015.

Mohammad Karimi, Ya-Ping Hsieh, Panayotis Mertikopoulos, and Andreas Krause. Riemannian stochastic approximation algorithms. *arXiv preprint arXiv:2206.06795*, 2022.

Alexandru Kristály, Chong Li, Genaro López-Acedo, and Adriana Nicolae. What do 'convexities' imply on hadamard manifolds? *Journal of Optimization Theory and Applications*, 170:1068–1074, 2016.

Aaron Lou, Isay Katsman, Qingxuan Jiang, Serge Belongie, Ser-Nam Lim, and Christopher De Sa. Differentiating through the fréchet mean. In *Proceeddings of the 37th International Conference on Machine Learning*, pages 6393–6403. PMLR, 2020.

Haipeng Luo and Robert E Schapire. Achieving all with no parameters: Adanormalhedge. In *Proceedings of the 28th Annual Conference on Learning Theory*, pages 1286–1304. PMLR, 2015.

Alejandro I Maass, Chris Manzie, Dragan Nesic, Jonathan H Manton, and Iman Shames. Tracking and regret bounds for online zeroth-order euclidean and riemannian optimization. *SIAM Journal on Optimization*, 32(2):445–469, 2022.

David Martínez-Rubio. Global riemannian acceleration in hyperbolic and spherical spaces. In *Proceedings of the 33rd International Conference on Algorithmic Learning Theory*, pages 768–826. PMLR, 2022.

David Martínez-Rubio and Sebastian Pokutta. Accelerated riemannian optimization: Handling constraints with a prox to bound geometric penalties. *arXiv preprint arXiv:2211.14645*, 2022.

Aryan Mokhtari, Shahin Shahrampour, Ali Jadbabaie, and Alejandro Ribeiro. Online optimization in dynamic environments: Improved regret rates for strongly convex problems. In *2016 IEEE 55th Conference on Decision and Control*, pages 7195–7201. IEEE, 2016.

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

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, pages 993–1019. PMLR, 2013.

Takashi Sakai. *Riemannian geometry*, volume 149. American Mathematical Soc., 1996.

Suvrit Sra, Nisheeth K Vishnoi, and Ozan Yildiz. On geodesically convex formulations for the brascamp-lieb constant. In *Proceedings of the 21st International Conference on Approximation Algorithms for Combinatorial Optimization Problems*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in neural information processing systems 23*, 2010.

Karl-Theodor Sturm. Probability measures on metric spaces of nonpositive curvature. *Heat Kernels and Analysis on Manifolds, Graphs, and Metric Spaces: Lecture Notes from a Quarter Program on Heat Kernels, Random Walks, and Analysis on Manifolds and Graphs*, 338:357, 2003.

Yue Sun, Nicolas Flammarion, and Maryam Fazel. Escaping from saddle points on riemannian manifolds. In *In Advances in Neural Information Processing Systems 32*, pages 7276–7286, 2019.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems 28*, pages 2989–2997, 2015.

Constantin Udriste. *Convex functions and optimization methods on Riemannian manifolds*, volume 297. Springer Science & Business Media, 2013.

Tim Van Erven and Wouter M Koolen. Metagrad: Multiple learning rates in online learning. In *In Advances in Neural Information Processing Systems 29*, pages 3666–3674, 2016.

Nisheeth K Vishnoi. Geodesic convex optimization: Differentiation on manifolds, geodesics, and convexity. *arXiv preprint arXiv:1806.06373*, 2018.

Yuanyu Wan, Bo Xue, and Lijun Zhang. Projection-free online learning in dynamic environments. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pages 10067–10075, 2021.

Xi Wang, Zhipeng Tu, Yiguang Hong, Yingyi Wu, and Guodong Shi. No-regret online learning over riemannian manifolds. In *Advances in Neural Information Processing Systems 34*, pages 28323–28335, 2021.

Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *The 29th Annual Conference on Learning Theory*, pages 1617–1638. PMLR, 2016.

Lijun Zhang, Tianbao Yang, Jinfeng Yi, Rong Jin, and Zhi-Hua Zhou. Improved dynamic regret for non-degenerate functions. In *Advance in Neural Information Processing Systems 30*, pages 732–741, 2017.

Lijun Zhang, Shiyin Lu, and Zhi-Hua Zhou. Adaptive online learning in dynamic environments. *In Advances in Neural Information Processing Systems*, 31, 2018.

Peiyuan Zhang, Jingzhao Zhang, and Suvrit Sra. Minimax in geodesic metric spaces: Sion's theorem and algorithms. *arXiv preprint arXiv:2202.06950*, 2022.

Peng Zhao and Lijun Zhang. Improved analysis for dynamic regret of strongly convex and smooth functions. In *Proceedings of the 3rd Conference on Learning for Dynamics and Control*, pages 48–59, 2021.

Peng Zhao, Yu-Jie Zhang, Lijun Zhang, and Zhi-Hua Zhou. Dynamic regret of convex and smooth functions. In *In Advances in Neural Information Processing Systems 33*, pages 12510–12520, 2020.

Li-wen Zhou and Nan-jing Huang. A revision on geodesic pseudo-convex combination and knaster–kuratowski–mazurkiewicz theorem on hadamard manifolds. *Journal of Optimization Theory and Applications*, 182(3):1186–1198, 2019.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, pages 928–936, 2003.

Appendix A. Omitted Proof for Section 4

A.1. Proof of Theorem 1

We denote $\mathbf{x}'_{t+1} = \mathrm{Exp}_{\mathbf{x}_t}(-\eta\nabla f_t(\mathbf{x}_t))$ and start from the geodesic convexity:

$$\begin{aligned}
f_t(\mathbf{x}_t) - f_t(\mathbf{u}_t) &\overset{(1)}{\leq} \langle \nabla f_t(\mathbf{x}_t), -\mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{u}_t)\rangle \\
&= \frac{1}{\eta}\langle \mathrm{Exp}_{\mathbf{x}_t}^{-1}\mathbf{x}'_{t+1}, \mathrm{Exp}_{\mathbf{x}_t}^{-1}\mathbf{u}_t\rangle \\
&\overset{(2)}{\leq} \frac{1}{2\eta}\left(\|\mathrm{Exp}_{\mathbf{x}_t}^{-1}\mathbf{u}_t\|^2 - \|\mathrm{Exp}_{\mathbf{x}'_{t+1}}^{-1}\mathbf{u}_t\|^2 + \zeta\|\mathrm{Exp}_{\mathbf{x}_t}^{-1}\mathbf{x}'_{t+1}\|^2\right) \\
&\overset{(3)}{\leq} \frac{1}{2\eta}\left(\|\mathrm{Exp}_{\mathbf{x}_t}^{-1}\mathbf{u}_t\|^2 - \|\mathrm{Exp}_{\mathbf{x}_{t+1}}^{-1}\mathbf{u}_t\|^2\right) + \frac{\eta\zeta G^2}{2} \\
&= \frac{1}{2\eta}\left(\|\mathrm{Exp}_{\mathbf{x}_t}^{-1}\mathbf{u}_t\|^2 - \|\mathrm{Exp}_{\mathbf{x}_{t+1}}^{-1}\mathbf{u}_{t+1}\|^2 + \|\mathrm{Exp}_{\mathbf{x}_{t+1}}^{-1}\mathbf{u}_{t+1}\|^2 - \|\mathrm{Exp}_{\mathbf{x}_{t+1}}^{-1}\mathbf{u}_t\|^2\right) + \frac{\eta\zeta G^2}{2} \\
&\overset{(4)}{\leq} \frac{1}{2\eta}\left(\|\mathrm{Exp}_{\mathbf{x}_t}^{-1}\mathbf{u}_t\|^2 - \|\mathrm{Exp}_{\mathbf{x}_{t+1}}^{-1}\mathbf{u}_{t+1}\|^2 + 2D\|\mathrm{Exp}_{\mathbf{u}_t}^{-1}\mathbf{u}_{t+1}\|\right) + \frac{\eta\zeta G^2}{2},
\end{aligned} \tag{10}$$

where for the second inequality we use Lemma 21, while the third is due to Lemma 23 and Assumption 3. For the last inequality, we invoke triangle inequality and Assumption 2.

WLOG, we can assume $\mathbf{u}_{T+1} = \mathbf{u}_T$ and sum from $t = 1$ to T:

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{u}_t) \leq \frac{D^2}{2\eta} + \frac{DP_T}{\eta} + \frac{\eta\zeta G^2 T}{2}. \tag{11}$$

A.2. Proof of Lemma 1

This is a generalization of (Cesa-Bianchi and Lugosi, 2006, Theorem 2.2) to the Riemannian manifold. Let $L_{t,i} = \sum_{s=1}^{t} f_s(\mathbf{x}_{s,i})$ and $W_t = \sum_{i=1}^{N} w_{1,i}e^{-\beta L_{t,i}}$. We have the following lower bound for $\ln W_T$,

$$\ln(W_T) = \ln\left(\sum_{i\in[N]} w_{1,i}e^{-\beta L_{t,i}}\right) \geq -\beta \min_{i\in[N]}\left(L_{T,i} + \frac{1}{\beta}\ln\frac{1}{w_{1,i}}\right).$$

For the next step, we try to get an upper bound on $\ln W_T$. When $t \geq 2$, we have

$$\ln\left(\frac{W_t}{W_{t-1}}\right) = \ln\frac{\sum_{i\in[N]} w_{1,i}e^{-\beta L_{t-1,i}}e^{-\beta f_t(\mathbf{x}_{t,i})}}{\sum_{i\in[N]} w_{1,i}e^{-\beta L_{t-1,i}}} = \ln\left(\sum_{i\in[N]} w_{t,i}e^{-\beta f_t(\mathbf{x}_{t,i})}\right),$$

where the updating rule of Hedge

$$w_{t,i} = \frac{w_{1,i}e^{-\beta L_{t-1,i}}}{\sum_{j\in[N]} w_{1,j}e^{-\beta L_{t-1,j}}}$$

is applied. Therefore

$$
\begin{aligned}
\ln W_T = \ln W_1 + \sum_{t=2}^{T} \ln\left(\frac{W_t}{W_{t-1}}\right) &= \sum_{t=1}^{T} \ln\left(\sum_{i\in[N]} w_{t,i} e^{-\beta f_t(\mathbf{x}_{t,i})}\right) \\
\leq \sum_{t=1}^{T}\left(-\beta \sum_{i\in[N]} w_{t,i} f_t(\mathbf{x}_{t,i}) + \frac{\beta^2 G^2 D^2}{8}\right) &\leq \sum_{t=1}^{T}\left(-\beta f_t(\mathbf{x}_t) + \frac{\beta^2 G^2 D^2}{8}\right),
\end{aligned}
\tag{12}
$$

where the first inequality follows from Hoeffding's inequality and $f_t(\mathbf{x}^\star) \leq f_t(\mathbf{x}) \leq f_t(\mathbf{x}^\star) + G \cdot D$ holds for any $\mathbf{x} \in \mathcal{N}$ and $\mathbf{x}^\star = \operatorname{argmin}_{\mathbf{x}\in\mathcal{N}} f_t(\mathbf{x})$, and the second inequality is due to both the Fréchet mean and the geodesic mean satisfy Jensen's inequality. For the Fréchet mean, we can apply Lemma 24. While Lemmas 19 and 26 ensure the geodesic mean satisfies the requirement.

Combining the lower and upper bound for $\ln W_T$, we see

$$
-\beta \min_{i\in[N]}\left(L_{T,i} + \frac{1}{\beta}\ln\frac{1}{w_{1,i}}\right) \leq \sum_{t=1}^{T}\left(-\beta f_t(\mathbf{x}_t) + \frac{\beta^2 G^2 D^2}{8}\right).
$$

After simplifying, we get

$$
\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{i\in[N]}\left(\sum_{t=1}^{T} f_t(\mathbf{x}_{t,i}) + \frac{1}{\beta}\ln\frac{1}{w_{1,i}}\right) \leq \frac{\beta G^2 D^2 T}{8}.
$$

Setting $\beta = \sqrt{\frac{8}{G^2 D^2 T}}$, we have

$$
\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_{t,i}) \leq \sqrt{\frac{G^2 D^2 T}{8}}\left(1 + \ln\frac{1}{w_{1,i}}\right).
$$

A.3. Proof of Theorem 2

Each expert performs R-OGD, so by Theorem 1 we have

$$
\sum_{t=1}^{T} f_t(\mathbf{x}_{t,i}) - f_t(\mathbf{u}_t) \leq \frac{D^2 + 2DP_T}{2\eta} + \frac{\eta\zeta G^2 T}{2}.
\tag{13}
$$

holds for any $i \in [N]$. Now it suffices to verify that there always exists $\eta_k \in \mathcal{H}$ which is close to the optimal stepsize

$$
\eta^\star = \sqrt{\frac{D^2 + 2DP_T}{\zeta G^2 T}}.
\tag{14}
$$

By Assumption 2,

$$
0 \leq P_T = \sum_{t=2}^{T} d(\mathbf{u}_{t-1}, \mathbf{u}_t) \leq TD.
$$

Thus

$$
\sqrt{\frac{D^2}{TG^2\zeta}} \leq \eta^* \leq \sqrt{\frac{D^2 + 2TD^2}{TG^2\zeta}}
$$

It is obvious that

$$\min \mathcal{H} = \sqrt{\frac{D^2}{TG^2\zeta}}, \text{ and } \max \mathcal{H} \geq 2\sqrt{\frac{D^2+2TD^2}{TG^2\zeta}}$$

Therefore, there exists $k \in [N-1]$ such that

$$\eta_k = 2^{k-1}\sqrt{\frac{D^2}{TG^2\zeta}} \leq \eta^* \leq 2\eta_k \tag{15}$$

The dynamic regret of the k-th expert is

$$\begin{aligned}
&\sum_{t=1}^{T} f_t(\mathbf{x}_{t,k}) - \sum_{t=1}^{T} f_t\left(\mathbf{u}_t\right)\\
\overset{(1)}{\leq}&\frac{D^2}{2\eta_k} + \frac{DP_T}{\eta_k} + \left(\frac{\eta_k TG^2\zeta}{2}\right)\\
\overset{(2)}{\leq}&\frac{D^2}{\eta^*} + \frac{2DP_T}{\eta^*} + \left(\frac{\eta^* TG^2\zeta}{2}\right)\\
=&\frac{3}{2}\sqrt{TG^2\zeta\left(D^2+2DP_T\right)}.
\end{aligned} \tag{16}$$

The second inequality follows from Equation (15) and we use Equation (14) to derive the last equality.

Since the initial weight of the k-th expert satisfies

$$w_{1,k} = \frac{N+1}{k(k+1)N} \geq \frac{1}{(k+1)^2},$$

the regret of the meta algorithm with respect to the k-th expert is bounded by

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_{t,k}) \leq \sqrt{\frac{G^2D^2T}{8}}\left(1+2\ln(k+1)\right) \tag{17}$$

in view of Lemma 1. Combining Equations (16) and (17), we have

$$\begin{aligned}
&\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}_t)\\
\leq&\frac{3}{2}\sqrt{TG^2\zeta\left(D^2+2DP_T\right)} + \sqrt{\frac{G^2D^2T}{8}}\left(1+2\ln(k+1)\right)\\
\leq&\sqrt{2\left(\frac{9}{4}TG^2\zeta\left(D^2+2DP_T\right) + \frac{G^2D^2T}{8}\left(1+2\ln(k+1)\right)^2\right)}\\
=&O\left(\sqrt{\zeta(1+P_T)T}\right),
\end{aligned} \tag{18}$$

where $\sqrt{a}+\sqrt{b} \leq \sqrt{2(a+b)}$ is applied to derive the second inequality, and for the equality follows from $\ln k = O(\log\log P_T) = o(\sqrt{P_T})$.

A.4. Minimax Dynamic Regret Lower Bound on Hadamard Manifolds

In this part, we first establish a $\Omega(\sqrt{T})$ minimax static regret lower bound on Hadamard manifolds following the classical work of Abernethy et al. (2008), then follow the reduction in Zhang et al. (2018) to get its dynamic counterpart. We focus on the manifold of SPD matrices (Bhatia, 2009) $\mathcal{S}_{++}^n = \{p : p \in \mathbb{R}^{n\times n}, p = p^T \text{and } p \succ 0^{n\times n}\}$, which becomes Hadamard when equipped with the affine-invariant metric $\langle U, V\rangle_p = tr(p^{-1}Up^{-1}V)$ where $tr(\cdot)$ is the trace operator. The tangent space of $\mathcal{S}_{++}^n$ is $\mathcal{S}^n = \{U : U \in \mathbb{R}^{n\times n}, U = U^T\}$. Under the affine-invariant metric, we have

$$\begin{aligned}
\mathrm{Exp}_p U &= p^{\frac{1}{2}} \exp\left(p^{-\frac{1}{2}} U p^{-\frac{1}{2}}\right) p^{\frac{1}{2}} \\
\mathrm{Exp}_p^{-1} q &= p^{\frac{1}{2}} \log\left(p^{-\frac{1}{2}} q p^{-\frac{1}{2}}\right) p^{\frac{1}{2}} \\
d(p, q) &= \sqrt{\sum_{i=1}^n (\log \lambda_i(q^{-\frac{1}{2}} p q^{-\frac{1}{2}}))^2}.
\end{aligned}$$

For technical reason, we restrict the manifold to be the manifold of SPD matrices with diagonal entries $\mathcal{D}_{++}^n = \{p : p \in \mathbb{R}^{n\times n}, p_{i,i} > 0, p \text{ is diagonal}\}$ and its tangent space is $\mathcal{D}^n = \{U : U \in \mathbb{R}^{n\times n}, U \text{ is diagonal}\}$. A key component in the proof is the Busemann function (Ballmann, 2012) on $\mathcal{D}_{++}^n$ equipped with affine-invariant metric has a closed form, which we describe as follows.

Definition 3 *(Ballmann, 2012) Suppose $\mathcal{M}$ is a Hadamard manifold and $c : [0, \infty)$ is a geodesic ray on $\mathcal{M}$ with $\|\dot{c}(0)\| = 1$. Then the Busemann function associated with c is defined as*

$$b_c(p) = \lim_{t\to\infty} (d(p, c(t)) - t).$$

Busemann functions enjoy the following useful properties.

Lemma 10 *(Ballmann, 2012) A Busemann function b_c satisfies*

1) b_c is gsc-convex;

2) $\nabla b_c(c(t)) = \dot{c}(t)$ for any $t \in [0, \infty)$;

3) $\|\nabla b_c(p)\| \leq 1$ for every $p \in \mathcal{M}$.

Lemma 11 *(Bridson and Haefliger, 2013, Chapter II.10) On $\mathcal{D}_{++}^n$, suppose $c(t) = \mathrm{Exp}_I(tX)$, $p = \mathrm{Exp}_I(Y)$ and $\|X\| = 1$, then the Busemann function is*

$$b_c(p) = -tr(XY).$$

Proof We first compute

$$\begin{aligned} d(p,c(t))^2 =& d(\mathrm{Exp}_I(Y), \mathrm{Exp}_I(tX))^2 \\ =& d(e^Y, e^{tX})^2 \\ =& \sum_{i=1}^n (\log \lambda_i(e^{-tX+Y}))^2 \\ =& d(I, e^{Y-tX})^2 \\ =& d(I, \mathrm{Exp}_I(Y-tX))^2 \\ =& \|Y-tX\|^2 \\ =& tr((Y-tX)(Y-tX)) \\ =& tr(Y^2) - 2t\cdot tr(XY) + t^2 \cdot tr(X^2), \end{aligned} \tag{19}$$

where we use facts that $\mathrm{Exp}_I(X) = e^X$ and $d(p,q) = \sqrt{\sum_{i=1}^n (\log \lambda_i(q^{-\frac{1}{2}} p q^{-\frac{1}{2}}))^2}$. Meanwhile,

$$\lim_{t\to\infty} \frac{d(p,c(t))}{t} = 1$$

due to the triangle inequality on $\triangle(p, c(0), c(t))$. Therefore,

$$\begin{aligned} b_c(p) =& \lim_{t\to\infty} (d(p,c(t)) - t) \\ =& \lim_{t\to\infty} \frac{d(p,c(t))^2 - t^2}{2t} \\ =& -tr(XY). \end{aligned}$$

■

Remark 12 *We consider $\mathcal{D}^n_{++}$ to ensure X and Y commute thus also e^Y and e^{tX} commute. This is necessary to get a closed form of the Busemann function.*

Now we describe the minimax game on $\mathcal{D}^n_{++}$. Each round t, the player chooses p_t from $\mathcal{N} = \{p : d(p, I) \leq \frac{D}{2}\} = \{p : p = \mathrm{Exp}_I(Y), \|Y\| \leq \frac{D}{2}\}$, and the adversary is allowed to pick a geodesic c^t, which determines a loss function in

$$\begin{aligned} \mathcal{F}_t =& \{\alpha_t G_t b_c(p) : \|\dot{c}(0)\| = 1, \alpha_t \in [0,1]\} \\ =& \{\alpha_t G_t \cdot -tr(X_t Y) : \|X_t\| = 1, \alpha_t \in [0,1]\} \\ =& \{-tr(X_t Y) : \|X_t\| \leq G_t\} \end{aligned} \tag{20}$$

The domain is gsc-convex by Lemmas 28 and 29. Each loss function is gsc-convex and has a gradient upper bound G_t using the second item of Lemma 10. WLOG, we assume $D = 2$, and the value of the game is

$$\mathcal{V}_T(\mathcal{N}, \{\mathcal{F}_t\}) := \inf_{\|Y_1\|\leq 1} \sup_{\|X_1\|\leq G_1} \ldots \inf_{\|Y_T\|\leq 1} \sup_{\|X_T\|\leq G_T} \left[\sum_{t=1}^T -tr(X_t Y_t) - \inf_{\|Y\|\leq 1} \sum_{t=1}^T -tr(X_t Y) \right]$$

Lemma 13 *The value of the minimax game $\mathcal{V}_T$ can be written as*

$$\mathcal{V}_T(\mathcal{N}, \{\mathcal{F}_t\}) = \inf_{\|Y_1\|\le 1} \sup_{\|X_1\|\le G_1} \dots \inf_{\|Y_T\|\le 1} \sup_{\|X_T\|\le G_T} \left[\sum_{t=1}^{T} -tr(X_tY_t) + \left\| \sum_{t=1}^{T} X_t \right\| \right]$$

Proof This is obvious due to Cauchy–Schwarz inequality:

$$\inf_{\|Y\|\le 1} \sum_{t=1}^{T} -tr(X_tY) = -\sup_{\|Y\|\le 1} tr\left(\sum_{t=1}^{T} X_tY \right) = -\left\| \sum_{t=1}^{T} X_t \right\|.$$

■

Lemma 14 *For $n > 2$, the adversary guarantees at least $\sqrt{\sum_{t=1}^{T} G_t^2}$ regardless of the player's strategy.*

Proof Each round, after the player chooses Y_t, the adversary chooses X_t such that $\|X_t\| = G_t$, $\langle X_t, Y_t \rangle = 0$ and $\left\langle X_t, \sum_{s=1}^{t-1} X_s \right\rangle = 0$. This is always possible when $n > 2$. Under this strategy, $\sum_{t=1}^{T} -tr(X_tY_t) = 0$ and we can show $\left\| \sum_{t=1}^{T} X_t \right\| = \sqrt{\sum_{t=1}^{T} G_t^2}$ by induction. The case for $T = 1$ is obvious. Assume $\left\| \sum_{s=1}^{t-1} X_s \right\| = \sqrt{\sum_{s=1}^{t-1} G_s^2}$, then

$$\left\| \sum_{s=1}^{t} X_s \right\| = \left\| \sum_{s=1}^{t-1} X_s + X_t \right\| = \sqrt{\left\| \sum_{s=1}^{t-1} X_s \right\|^2 + \|X_t\|^2} = \sqrt{\sum_{s=1}^{t} G_t^2}.$$

where the second equality is due to $\left\langle X_t, \sum_{s=1}^{t-1} X_s \right\rangle = 0$. ■

Lemma 15 *Let $X_0 = 0$. If the player plays*

$$Y_t = \frac{\sum_{s=1}^{t-1} X_s}{\sqrt{\left\| \sum_{s=1}^{t-1} X_s \right\|^2 + \sum_{s=t}^{T} G_s^2}},$$

then

$$\sup_{\|X_1\|\le G_1} \sup_{\|X_2\|\le G_2} \dots \sup_{\|X_T\|\le G_T} \left[\sum_{t=1}^{T} -tr(X_tY_t) + \left\| \sum_{t=1}^{T} X_t \right\| \right] \le \sqrt{\sum_{t=1}^{T} G_t^2}.$$

Proof Let $\Gamma_t^2 = \sum_{s=t}^{T} G_s^2$, $\Gamma_{T+1} = 0$ and $\tilde{X}_t = \sum_{s=1}^{t} X_s$. We define

$$\Phi_t(X_1, \dots, X_{t-1}) = \sum_{s=1}^{t-1} -tr(X_sY_s) + \sqrt{\left\| \sum_{s=1}^{t-1} X_s \right\|^2 + \Gamma_t^2}$$

and $\Phi_1 = \sqrt{\sum_{t=1}^T G_t^2}$. We further let

$$\Psi_t(X_1, \dots, X_{t-1}) = \sup_{\|X_t\| \leq G_t} \dots \sup_{\|X_T\| \leq G_T} \left[\sum_{s=1}^T -tr(X_s Y_s) + \left\| \sum_{s=1}^T X_s \right\| \right]$$

be the payoff of the adversary when he plays $X_1, \dots, X_{t-1}$ and then plays optimally.

We do backward induction for this argument, which means for all $t \in \{1, \dots, T+1\}$,

$$\Psi_t(X_1, \dots, X_{t-1}) \leq \Phi_t(X_1, \dots, X_{t-1}).$$

The case of $t = T+1$ is obvious because $\Psi_{T+1} = \Phi_{T+1}$. Assume the argument holds for $t+1$ and we try to show the case for t.

$$\begin{aligned}
&\Psi_t(X_1, \dots, X_{t-1}) \\
=& \sup_{\|X_t\| \leq G_t} \Psi_{t+1}(X_1, \dots, X_t) \\
\leq& \sup_{\|X_t\| \leq G_t} \Phi_{t+1}(X_1, \dots, X_t) \\
=& \sum_{s=1}^{t-1} -tr(X_s Y_s) + \sup_{\|X_t\| \leq G_t} \left[-tr(X_t Y_t) + \sqrt{\left\| \sum_{s=1}^t X_s \right\|^2 + \Gamma_{t+1}^2} \right].
\end{aligned} \tag{21}$$

Now it suffices to show

$$\sup_{\|X_t\| \leq G_t} \left[-tr(X_t Y_t) + \sqrt{\left\| \sum_{s=1}^t X_s \right\|^2 + \Gamma_{t+1}^2} \right] \leq \sqrt{\left\| \sum_{s=1}^{t-1} X_s \right\|^2 + \Gamma_t^2}$$

to establish our argument. Recall that

$$Y_t = \frac{\sum_{s=1}^{t-1} X_s}{\sqrt{\left\| \sum_{s=1}^{t-1} X_s \right\|^2 + \sum_{s=t}^T G_s^2}},$$

and denote $\tilde{X}_{t-1} = \sum_{s=1}^{t-1} X_s$. It turns out that what we need to show is

$$\sup_{\|X_t\| \leq G_t} -tr\left(\frac{\left\langle X_t, \tilde{X}_{t-1} \right\rangle}{\sqrt{\|\tilde{X}_{t-1}\|^2 + \Gamma_t^2}} \right) + \sqrt{\|\tilde{X}_{t-1} + X_t\|^2 + \Gamma_{t+1}^2} \leq \sqrt{\|\tilde{X}_{t-1}\|^2 + \Gamma_t^2}.$$

We use the Lagrange multiplier method to prove this argument. Let

$$g(X_t) = -tr\left(\frac{\left\langle X_t, \tilde{X}_{t-1} \right\rangle}{\sqrt{\|\tilde{X}_{t-1}\|^2 + \Gamma_t^2}} \right) + \sqrt{\|\tilde{X}_{t-1} + X_t\|^2 + \Gamma_{t+1}^2} + \lambda(\|X_t\|^2 - G_t^2),$$

then the stationary point of g satisfies

$$\frac{\partial g(X_t)}{\partial X_t} = -\frac{\tilde{X}_{t-1}}{\sqrt{\|\tilde{X}_{t-1}\|^2 + \Gamma_t^2}} + \frac{\tilde{X}_{t-1} + X_t}{\sqrt{\|\tilde{X}_{t-1} + X_t\|^2 + \Gamma_{t+1}^2}} + 2\lambda X_t = 0$$

and

$$\lambda(\|X_t\|^2 - G_t^2) = 0.$$

We first consider that $\tilde{X}_{t-1}$ is co-linear with X_t. When $\lambda = 0$, we have $X_t = c\tilde{X}_{t-1}$ where

$$c = \frac{\Gamma_{t+1}}{\Gamma_t} - 1.$$

If $\tilde{X}_{t-1}$ is co-linear with X_t and $\lambda \neq 0$, we know $\|X_t\| = G_t$ and again let $X_t = G_t \frac{\tilde{X}_{t-1}}{\|\tilde{X}_{t-1}\|}$ or $X_t = -G_t \frac{\tilde{X}_{t-1}}{\|\tilde{X}_{t-1}\|}$. Then we need to ensure

$$g(c\tilde{X}_{t-1}) \leq \sqrt{\|\tilde{X}_{t-1}\|^2 + \Gamma_t^2}$$

holds for $c = \frac{\Gamma_{t+1}}{\Gamma_t} - 1$, $\frac{G_t}{\|\tilde{X}_{t-1}\|}$ and $-\frac{G_t}{\|\tilde{X}_{t-1}\|}$.

By Lemma 16, it suffices to verify

$$(c^2 - 1)\|\tilde{X}_{t-1}\|^2\Gamma_t^2 + (\|\tilde{X}_{t-1}\|^2 + \Gamma_t^2)\Gamma_{t+1}^2 \leq \Gamma_t^4.$$

If $c = \frac{\Gamma_{t+1}}{\Gamma_t} - 1$, we have to ensure

$$\begin{aligned}
&(c^2 - 1)\|\tilde{X}_{t-1}\|^2\Gamma_t^2 + (\|\tilde{X}_{t-1}\|^2 + \Gamma_t^2)\Gamma_{t+1}^2 - \Gamma_t^4 \\
=&\left(\frac{\Gamma_{t+1}^2}{\Gamma_t^2} - 2\frac{\Gamma_{t+1}}{\Gamma_t}\right)\|\tilde{X}_{t-1}\|^2\Gamma_t^2 + \|\tilde{X}_{t-1}\|^2\Gamma_{t+1}^2 + \Gamma_t^2\Gamma_{t+1}^2 - \Gamma_t^4 \\
=&2(\Gamma_{t+1} - \Gamma_t)\Gamma_{t+1}\|\tilde{X}_{t-1}\|^2 + \Gamma_t^2(\Gamma_{t+1}^2 - \Gamma_t^2) \leq 0.
\end{aligned} \tag{22}$$

For the case where $c^2 = \frac{G_t^2}{\|\tilde{X}_{t-1}\|^2}$, we have

$$\begin{aligned}
&(c^2 - 1)\|\tilde{X}_{t-1}\|^2\Gamma_t^2 + (\|\tilde{X}_{t-1}\|^2 + \Gamma_t^2)\Gamma_{t+1}^2 - \Gamma_t^4 \\
=&\left(\frac{G_t^2}{\|\tilde{X}_{t-1}\|^2} - 1\right)\|\tilde{X}_{t-1}\|^2\Gamma_t^2 + (\|\tilde{X}_{t-1}\|^2 + \Gamma_t^2)\Gamma_{t+1}^2 - \Gamma_t^4 \\
=&\Gamma_t^2(G_t^2 + \Gamma_{t+1}^2 - \Gamma_t^2) - G_t^2\|\tilde{X}_{t-1}\|^2 \\
=&-G_t^2\|\tilde{X}_{t-1}\|^2 \leq 0.
\end{aligned} \tag{23}$$

The only case left is when X_t is not parallel to $\tilde{X}_{t-1}$. $\lambda = 0$ implies $X_t = 0$ and thus

$$g(0) = \sqrt{\|\tilde{X}_{t-1}\|^2 + \Gamma_{t+1}^2} \leq \sqrt{\|\tilde{X}_{t-1}\|^2 + \Gamma_t^2}.$$

If $\lambda \neq 0$ then $\|X_t\| = G$. We have

$$-\frac{\tilde{X}_{t-1}}{\sqrt{\|\tilde{X}_{t-1}\|^2 + \Gamma_t^2}} + \frac{\tilde{X}_{t-1}}{\sqrt{\|\tilde{X}_{t-1} + X_t\|^2 + \Gamma_{t+1}^2}} = 0$$

which in turn implies $\left\langle X_t, \tilde{X}_{t-1} \right\rangle = 0$. This is the maximum point of g as now

$$g(X_t) = \sqrt{\|\tilde{X}_{t-1}\|^2 + \Gamma_t^2}.$$

Thus we finished the induction step and the lemma was established. ■

Lemma 16

$$-tr\left(\frac{\left\langle X_t, \tilde{X}_{t-1} \right\rangle}{\sqrt{\|\tilde{X}_{t-1}\|^2 + \Gamma_t^2}}\right) + \sqrt{\|\tilde{X}_{t-1} + X_t\|^2 + \Gamma_{t+1}^2} \leq \sqrt{\|\tilde{X}_{t-1}\|^2 + \Gamma_t^2}.$$

holds for $X_t = c\tilde{X}_{t-1}$ iff

$$(c^2 - 1)\|\tilde{X}_{t-1}\|^2\Gamma_t^2 + (\|\tilde{X}_{t-1}\|^2 + \Gamma_t^2)\Gamma_{t+1}^2 \leq \Gamma_t^4.$$

Proof The statement we want to show is

$$-\frac{c\|\tilde{X}_{t-1}\|^2}{\sqrt{\|\tilde{X}_{t-1}\|^2 + \Gamma_t^2}} + \sqrt{\|\tilde{X}_{t-1}\|^2(1+c)^2 + \Gamma_{t+1}^2} \leq \sqrt{\|\tilde{X}_{t-1}\|^2 + \Gamma_t^2}.$$

Let $\alpha = \|\tilde{X}_{t-1}\|^2$, $\beta = \Gamma_t^2$ and $\gamma = \Gamma_{t+1}^2$. Following a series of algebraic manipulations, we get

$$(c^2 - 1)\alpha\beta + (\alpha + \beta)\gamma \leq \beta^2.$$

And the argument is proved after plugging back α, β, γ. ■

Theorem 10 *There exists a game on $\mathcal{D}_{++}^n$ such that we can exactly compute the value of the minimax regret. Specifically, the decision set of the player is $\mathcal{N} = \{p : p = \mathrm{Exp}_I(Y), \|Y\| \leq \frac{D}{2}\}$, and the adversary is allowed to pick a loss function in*

$$\mathcal{F}_t = \{\alpha_t G_t b_c(p) : \|\dot{c}(0)\| = 1, \alpha_t \in [0,1]\} = \{-tr(X_t Y) : \|X_t\| \leq G_t\}.$$

Then the minimax value of the game is

$$\mathcal{V}_T(\mathcal{N}, \{\mathcal{F}_t\}) = \frac{D}{2}\sqrt{\sum_{t=1}^{T} G_t^2}.$$

In addition, the optimal strategy of the player is

$$Y_t = \frac{\sum_{s=1}^{t-1} X_s}{\sqrt{\left\|\sum_{s=1}^{t-1} X_s\right\|^2 + \sum_{s=t}^{T} G_s^2}}.$$

Proof The proposition is a direct conclusion of Lemmas 13, 14, and 15. ■

Theorem 11 *There exists a comparator sequence which satisfies* $\sum_{t=2}^{T} d(\mathbf{u}_t, \mathbf{u}_{t-1}) \leq P_T$ *and the dynamic minimax regret lower bound on Hadamard manifolds is* $\Omega(G\sqrt{D^2 + DP_T})$.

Proof We combine Theorem 10 with a reduction in Zhang et al. (2018) to finish the proof. By Theorem 10 we have

$$\mathcal{V}_T = \inf_{\|Y_1\|\leq 1} \sup_{\|X_1\|\leq G} \dots \inf_{\|Y_T\|\leq 1} \sup_{\|X_T\|\leq G} \left[\sum_{t=1}^{T} -tr(X_t Y_t) - \inf_{\|Y\|\leq 1} \sum_{t=1}^{T} -tr(X_t Y) \right] = \frac{GD\sqrt{T}}{2}.$$

Note that the path-length is upper bounded by TD. For any $\tau \in [0, TD]$, we define the set of comparators with path-length bounded by τ as

$$C(\tau) = \left\{ \mathbf{u}_1, \dots, \mathbf{u}_T \in \mathcal{N} : \sum_{t=2}^{T} d(\mathbf{u}_t, \mathbf{u}_{t-1}) \leq \tau \right\}$$

where $\mathcal{N} = \{\mathbf{u} : d(I, \mathbf{u}) \leq \frac{D}{2}\}$ is a gsc-convex subset and the minimax dynamic regret w.r.t. $C(\tau)$ is

$$\mathcal{V}_T(C(\tau)) = \inf_{\|Y_1\|\leq 1} \sup_{\|X_1\|\leq G} \dots \inf_{\|Y_T\|\leq 1} \sup_{\|X_T\|\leq G} \left[\sum_{t=1}^{T} -tr(X_t Y_t) - \inf_{\mathbf{u}_1, \dots, \mathbf{u}_T \in C(\tau)} \sum_{t=1}^{T} -tr(X_t \mathrm{Exp}_I^{-1} \mathbf{u}_t) \right].$$

We distinguish two cases. When $\tau \leq D$, we invoke the minimax static regret directly to get

$$\mathcal{V}_T(C(\tau)) \geq \mathcal{V}_T = \frac{GD\sqrt{T}}{2}. \tag{24}$$

For the case of $\tau \geq D$, WLOG, we assume $\lceil \tau/D \rceil$ divides T and let L be the quotient. We construct a subset of $C(\tau)$, named $C'(\tau)$, which contains comparators that are fixed for each consecutive L rounds. Specifically,

$$C'(\tau) = \left\{ \mathbf{u}_1, \dots, \mathbf{u}_T \in \mathcal{N} : \mathbf{u}_{(i-1)L+1} = \dots \mathbf{u}_{iL}, \forall i \in [1, \lceil \tau/D \rceil] \right\}.$$

Note that the path-length of comparators in $C'(\tau)$ is at most $(\lceil \tau/D \rceil - 1)D \leq \tau$, which implies $C'(\tau)$ is a subset of $C(\tau)$. Thus we have

$$\mathcal{V}_T(C(\tau)) \geq \mathcal{V}_T(C'(\tau)). \tag{25}$$

The objective of introducing $C'(\tau)$ is we can set $\mathbf{u}_{(i-1)L+1} = \dots = \mathbf{u}_{(iL)}$ to be the offline minimizer of the i-th segment and invoke the minimax lower bound for the static regret for each segment. Thus we have

$$\begin{aligned} &\mathcal{V}_T(C'(\tau)) \\ =& \inf_{\|Y_1\|\leq 1} \sup_{\|X_1\|\leq G} \dots \inf_{\|Y_T\|\leq 1} \sup_{\|X_T\|\leq G} \left[\sum_{t=1}^{T} -tr(X_t Y_t) - \inf_{\mathbf{u}_1, \dots, \mathbf{u}_T \in C'(\tau)} \sum_{t=1}^{T} -tr(X_t \mathrm{Exp}_I^{-1} \mathbf{u}_t) \right] \\ =& \inf_{\|Y_1\|\leq 1} \sup_{\|X_1\|\leq G} \dots \inf_{\|Y_T\|\leq 1} \sup_{\|X_T\|\leq G} \left[\sum_{t=1}^{T} -tr(X_t Y_t) - \sum_{i=1}^{\lceil \tau/D \rceil} \inf_{\|Y\|\leq 1} \sum_{t=(i-1)L+1}^{iL} -tr(X_t Y) \right] \\ =& \lceil \tau/D \rceil \frac{GD\sqrt{L}}{2} = \frac{GD\sqrt{T\lceil \tau/D \rceil}}{2} \geq \frac{G\sqrt{TD\tau}}{2}. \end{aligned} \tag{26}$$

Combining Equations (24), (25) and (26) yields

$$\mathcal{V}_T(C(\tau)) \geq \frac{G}{2}\max(D\sqrt{T}, \sqrt{TD\tau}) = \Omega(G\sqrt{T(D^2 + D\tau)}).$$

■

Appendix B. Omitted Proof for Section 5

B.1. Proof of Theorem 4

We first argue $\mathcal{N}_c$ is gsc-convex for any $c \geq 0$ to ensure the algorithm is well-defined. By Lemma 28, $d(\mathbf{x}, \mathcal{N})$ is gsc-convex on Hadamard manifolds. The sub-level set of a gsc-convex function is a gsc-convex set due to Lemma 29, which implies $\mathcal{N}_c$ is gsc-convex. We notice that

$$f_t(\mathbf{x}_t) - f_t(\mathbf{u}_t) = (f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t')) + (f_t(\mathbf{x}_t') - f_t(\mathbf{u}_t))$$

and derive upper bounds for two terms individually. If $\mathbf{x}_t' \in \mathcal{N}_{\delta M}$ then $\mathbf{x}_t' = \mathbf{x}_t$ and $f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t') = 0$. If this is not the case, by Lemma 23, we have

$$d(\mathbf{x}_t', \mathbf{x}_t) \leq d(\mathbf{x}_t', \mathbf{z}) \leq d(\mathbf{x}_t', \mathbf{y}_t)$$

where $\mathbf{z}$ is the intersection of $\mathcal{N}_\delta$ and the geodesic segment connecting $\mathbf{x}_t'$ and $\mathbf{y}_t$. Thus

$$f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t') \leq \left\langle \nabla f_t(\mathbf{x}_t), -\mathrm{Exp}^{-1}_{\mathbf{x}_t}\mathbf{x}_t' \right\rangle \leq \|\nabla f_t(\mathbf{x}_t)\| \cdot d(\mathbf{x}_t', \mathbf{x}_t) \leq G \cdot d(\mathbf{x}_t', \mathbf{y}_t), \tag{27}$$

where we notice $\mathbf{x}_t \in \mathcal{N}_{\delta M}$ and use Assumption 3. Let $\mathbf{y}_{t+1}' = \mathrm{Exp}_{\mathbf{x}_t'}\left(-\eta \nabla f_t(\mathbf{x}_t') + \mathrm{Exp}^{-1}_{\mathbf{x}_t'}(\mathbf{y}_t)\right)$. The second term $f_t(\mathbf{x}_t') - f_t(\mathbf{u}_t)$ can be bounded by

$$\begin{aligned}
&f_t(\mathbf{x}_t') - f_t(\mathbf{u}_t) \overset{(1)}{\leq} -\langle \mathrm{Exp}^{-1}_{\mathbf{x}_t'}(\mathbf{u}_t), \nabla f_t(\mathbf{x}_t') \rangle \\
=&\frac{1}{\eta}\langle \mathrm{Exp}^{-1}_{\mathbf{x}_t'}(\mathbf{u}_t), \mathrm{Exp}^{-1}_{\mathbf{x}_t'}(\mathbf{y}_{t+1}') - \mathrm{Exp}^{-1}_{\mathbf{x}_t'}(\mathbf{y}_t) \rangle \\
\overset{(2)}{\leq}&\frac{1}{2\eta}\left(\zeta d(\mathbf{x}_t', \mathbf{y}_{t+1}')^2 + d(\mathbf{x}_t', \mathbf{u}_t)^2 - d(\mathbf{y}_{t+1}', \mathbf{u}_t)^2\right) - \frac{1}{2\eta}\left(d(\mathbf{x}_t', \mathbf{y}_t)^2 + d(\mathbf{x}_t', \mathbf{u}_t)^2 - d(\mathbf{y}_t, \mathbf{u}_t)^2\right) \\
=&\frac{1}{2\eta}\left(\zeta d(\mathbf{x}_t', \mathbf{y}_{t+1}')^2 - d(\mathbf{x}_t', \mathbf{y}_t)^2 + d(\mathbf{y}_t, \mathbf{u}_t)^2 - d(\mathbf{y}_{t+1}', \mathbf{u}_t)^2\right) \\
=&\frac{1}{2\eta}\left(\zeta\| -\eta\nabla f_t(\mathbf{x}_t') + \mathrm{Exp}^{-1}_{\mathbf{x}_t'}(\mathbf{y}_t)\|^2 - d(\mathbf{x}_t', \mathbf{y}_t)^2 + d(\mathbf{y}_t, \mathbf{u}_t)^2 - d(\mathbf{y}_{t+1}', \mathbf{u}_t)^2\right) \\
=&\frac{1}{2\eta}\left(\zeta\| -\eta\nabla f_t(\mathbf{x}_t') + \eta\Gamma_{\mathbf{y}_t}^{\mathbf{x}_t'} M_t\|^2 - d(\mathbf{x}_t', \mathbf{y}_t)^2 + d(\mathbf{y}_t, \mathbf{u}_t)^2 - d(\mathbf{y}_{t+1}', \mathbf{u}_t)^2\right) \\
\overset{(3)}{\leq}&\frac{1}{2\eta}\left(\zeta\| -\eta\nabla f_t(\mathbf{x}_t') + \eta\Gamma_{\mathbf{y}_t}^{\mathbf{x}_t'} M_t\|^2 - d(\mathbf{x}_t', \mathbf{y}_t)^2 + d(\mathbf{y}_t, \mathbf{u}_t)^2 - d(\mathbf{y}_{t+1}, \mathbf{u}_t)^2\right)
\end{aligned} \tag{28}$$

where the second inequality follows from Lemmas 21 and 22 and the third one is due to the non-expansive property of projection onto Hadamard manifolds. We apply $\Gamma_{\mathbf{x}}^{\mathbf{y}}\mathrm{Exp}_{\mathbf{x}}^{-1}\mathbf{y} = -\mathrm{Exp}_{\mathbf{y}}^{-1}\mathbf{x}$ to derive the last equality.

Now we can get the desired squared term $\|\nabla f_t(\mathbf{y}_t) - M_t\|^2$ by considering

$$\begin{aligned}
&\|\nabla f_t(\mathbf{x}_t') - \Gamma_{\mathbf{y}_t}^{\mathbf{x}_t'} M_t\|^2 \\
=&\|\nabla f_t(\mathbf{x}_t') - \Gamma_{\mathbf{y}_t}^{\mathbf{x}_t'} \nabla f_t(\mathbf{y}_t) + \Gamma_{\mathbf{y}_t}^{\mathbf{x}_t'} \nabla f_t(\mathbf{y}_t) - \Gamma_{\mathbf{y}_t}^{\mathbf{x}_t'} M_t\|^2 \\
\leq&2\left(\|\nabla f_t(\mathbf{x}_t') - \Gamma_{\mathbf{y}_t}^{\mathbf{x}_t'} \nabla f_t(\mathbf{y}_t)\|^2 + \|\Gamma_{\mathbf{y}_t}^{\mathbf{x}_t'} \nabla f_t(\mathbf{y}_t) - \Gamma_{\mathbf{y}_t}^{\mathbf{x}_t'} M_t\|^2\right) \\
\leq&2L^2 d(\mathbf{x}_t', \mathbf{y}_t)^2 + 2\|\nabla f_t(\mathbf{y}_t) - M_t\|^2,
\end{aligned} \tag{29}$$

where in the first inequlity we use $\|\mathbf{x}+\mathbf{y}\|^2 \leq 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2$ holds for any SPD norm $\|\cdot\|$, and the second inequality is due to the smoothness of f and parallel transport is an isometry. Combining Equations (27), (28) and (29), we have

$$\begin{aligned}
&f_t(\mathbf{x}_t) - f_t(\mathbf{u}_t) \\
\leq&Gd(\mathbf{x}_t', \mathbf{y}_t) + \frac{\eta\zeta}{2}\left(2\|\nabla f_t(\mathbf{y}_t) - M_t\|^2 + 2L^2 d(\mathbf{x}_t', \mathbf{y}_t)^2\right) \\
&- \frac{1}{2\eta} d(\mathbf{x}_t', \mathbf{y}_t)^2 + \frac{1}{2\eta}(d(\mathbf{y}_t, \mathbf{u}_t)^2 - d(\mathbf{y}_{t+1}, \mathbf{u}_t)^2) \\
\leq&\eta\zeta\|\nabla f_t(\mathbf{y}_t) - M_t\|^2 + \frac{1}{2\eta}\left(2\eta G + 2\eta^2\zeta L^2 d(\mathbf{x}_t', \mathbf{y}_t) - d(\mathbf{x}_t', \mathbf{y}_t)\right) d(\mathbf{x}_t', \mathbf{y}_t) \\
&+ \frac{1}{2\eta}(d(\mathbf{y}_t, \mathbf{u}_t)^2 - d(\mathbf{y}_{t+1}, \mathbf{u}_t)^2).
\end{aligned} \tag{30}$$

Now we show

$$\frac{1}{2\eta}\left(2\eta G + 2\eta^2\zeta L^2 d(\mathbf{x}_t', \mathbf{y}_t) - d(\mathbf{x}_t', \mathbf{y}_t)\right) d(\mathbf{x}_t', \mathbf{y}_t) \leq 0 \tag{31}$$

holds for any $t \in [T]$. First we consider the case that $d(\mathbf{x}_t', \mathbf{y}_t) \leq \delta M$, which means $\mathbf{x}_t' \in \mathcal{N}_{\delta M}$ and $f_t(\mathbf{x}_t) = f_t(\mathbf{x}_t')$. Thus Equation (31) is implied by

$$2\eta^2\zeta L^2 d(\mathbf{x}_t', \mathbf{y}_t) - d(\mathbf{x}_t', \mathbf{y}_t) \leq 0,$$

which is obviously true by considering our assumption on η:

$$\eta \leq \frac{\delta M}{G + (G^2 + 2\zeta\delta^2 M^2 L^2)^{\frac{1}{2}}} \leq \frac{1}{\sqrt{2\zeta L^2}}.$$

When $d(\mathbf{x}_t', \mathbf{y}_t) \geq \delta M$, to simplify the proof, we denote $\lambda = d(\mathbf{x}_t', \mathbf{y}_t)$ and try to find η such that

$$h(\eta; \lambda) := 2\eta G + 2\eta^2\zeta L^2\lambda - \lambda \leq 0 \tag{32}$$

holds for any $\lambda \geq \delta M$. We denote the only non-negative root of $h(\eta; \lambda)$ as $\eta(\lambda)$, which can be solved explicitly as

$$\eta(\lambda) = \frac{-G + (G^2 + 2\zeta\lambda^2 L^2)^{\frac{1}{2}}}{2\zeta\lambda L^2}.$$

Applying Lemma 25 with $a = G$ and $b = 2\zeta L^2$, we know $\eta(\lambda)$ increases on $[0, \infty)$. Thus $\eta(\lambda) \geq \eta(\delta M)$ holds for any $\lambda \geq \delta M$. Combining with the fact that $h(0; \lambda) = -\lambda < 0$, we know $h(\eta; \lambda) \leq 0$ holds for any $\eta \leq \eta(\lambda)$, so we can simply set

$$\eta \leq \min_{\lambda} \eta(\lambda) = \eta(\delta M) = \frac{\delta M}{G + (G^2 + 2\zeta\delta^2 M^2 L^2)^{\frac{1}{2}}}.$$

to ensure $h(\eta;\lambda)\leq 0$ always holds.

Now it suffices to bound

$$\begin{aligned}&\sum_{t=1}^{T}\frac{1}{2\eta}\left(d(\mathbf{y}_t,\mathbf{u}_t)^2-d(\mathbf{y}_{t+1},\mathbf{u}_t)^2\right)\\ \leq&\frac{d(\mathbf{y}_1,\mathbf{u}_1)^2}{2\eta}+\sum_{t=2}^{T}\frac{\left(d(\mathbf{y}_t,\mathbf{u}_t)^2-d(\mathbf{y}_t,\mathbf{u}_{t-1})^2\right)}{2\eta}\\ \leq&\frac{D^2}{2\eta}+2D\frac{\sum_{t=2}^{T}d(\mathbf{u}_t,\mathbf{u}_{t-1})}{2\eta}=\frac{D^2+2DP_T}{2\eta}.\end{aligned} \tag{33}$$

Finally, we apply the telescoping-sum on Equation (30),

$$\sum_{t=1}^{T}f_t(\mathbf{x}_t)-\sum_{t=1}^{T}f_t(\mathbf{u}_t)\leq\eta\zeta\sum_{t=1}^{T}\|\nabla f_t(\mathbf{y}_t)-M_t\|^2+\frac{D^2+2DP_T}{2\eta}. \tag{34}$$

B.2. Extension of Theorem 4 to CAT(κ) Spaces

In this part, we show how to get optimistic regret bound on CAT(κ) spaces, the sectional curvature of which is upper bounded by κ. Note that Hadamard manifolds are complete CAT(0) spaces. To proceed, we make the following assumption.

Assumption 4 *The sectional curvature of manifold $\mathcal{M}$ satisfies $-\kappa_1\leq\kappa\leq\kappa_2$ where $\kappa_1\geq 0$. We define*

$$\mathcal{D}(\kappa):=\begin{cases}\infty, & \kappa\leq 0\\ \frac{\pi}{\sqrt{\kappa}}, & \kappa>0.\end{cases}$$

The diameter of the gsc-convex set $\mathcal{N}\subset\mathcal{M}$ is D and we assume $D+2\delta M\leq\mathcal{D}(\kappa_2)$. The gradient satisfies $\sup_{\mathbf{x}\in\mathcal{N}_{\delta M}}\|\nabla f_t(\mathbf{x})\|\leq G$.

Lemma 17 *(Alimisis et al., 2020, Corollary 2.1) Let $\mathcal{M}$ be a Riemannian manifold with sectional curvature upper bounded by κ_2 and $\mathcal{N}\subset\mathcal{M}$ be a gsc-convex set with diameter upper bounded by $\mathcal{D}(\kappa_2)$. For a geodesic triangle fully lies within $\mathcal{N}$ with side lengths a,b,c, we have*

$$a^2\geq\xi(\kappa_2,D)b^2+c^2-2bc\cos A$$

where $\xi(\kappa_2,D)=\sqrt{-\kappa_2}D\coth(\sqrt{-\kappa_2}D)$ when $\kappa_2\leq 0$ and $\xi(\kappa_2,D)=\sqrt{\kappa_2}D\cot(\sqrt{\kappa_2}D)$ when $\kappa_2>0$.

Definition 4 *We define $\zeta=\zeta(-\kappa_1,D+2\delta M)$ and $\xi=\xi(\kappa_2,D+2\delta M)$ where $\xi(\cdot,\cdot)$ and $\zeta(\cdot,\cdot)$ are defined in Lemmas 17 and 21, respectively.*

Lemma 18 *(Bridson and Haefliger, 2013) For a CAT(κ) space $\mathcal{M}$, a ball of diameter smaller than $\mathcal{D}(\kappa)$ is convex. Let C be a convex subset in $\mathcal{M}$. If $d(\mathbf{x},C)\leq\frac{\mathcal{D}(\kappa)}{2}$ then $d(\mathbf{x},C)$ is convex and there exists a unique point $\Pi_C(\mathbf{x})\in\mathcal{C}$ such that $d(\mathbf{x},\Pi_C(\mathbf{x}))=d(\mathbf{x},C)=\inf_{\mathbf{y}\in C}d(\mathbf{x},\mathbf{y})$.*

Theorem 12 *Suppose all losses f_t are L-gsc-smooth on $\mathcal{M}$. Under Assumptions 4, the iterates*

$$\begin{aligned} \mathbf{x}_t' &= \mathrm{Exp}_{\mathbf{y}_t}(-\eta M_t) \\ \mathbf{x}_t &= \Pi_{\mathcal{N}_{\delta M}} \mathbf{x}_t' \\ \mathbf{y}_{t+1} &= \Pi_{\mathcal{N}} \mathrm{Exp}_{\mathbf{x}_t'} \left(-\eta \nabla f_t(\mathbf{x}_t') + \mathrm{Exp}_{\mathbf{x}_t'}^{-1}(\mathbf{y}_t)\right). \end{aligned} \tag{35}$$

satisfies

$$\textstyle\sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}_t) \le \eta\zeta \sum_{t=1}^T \|\nabla f_t(\mathbf{y}_t) - M_t\|^2 + \frac{D^2+2DP_T}{2\eta}.$$

for any $\mathbf{u}_1, \dots, \mathbf{u}_T \in \mathcal{N}$ and $\eta \le \min\left\{\frac{\xi\delta M}{G+(G^2+2\zeta\xi\delta^2 M^2 L^2)^{\frac{1}{2}}}, \frac{\mathcal{D}(\kappa_2)}{2(G+2M)}\right\}$.

Proof We highlight key differences between the proof of Theorem 4. Again we let $\mathbf{y}_{t+1}' = \mathrm{Exp}_{\mathbf{x}_t'}\left(-\eta\nabla f_t(\mathbf{x}_t') + \mathrm{Exp}_{\mathbf{x}_t'}^{-1}(\mathbf{y}_t)\right)$. First, we need to argue the algorithm is well-defined. The diameter of $\mathcal{N}_{\delta M}$ is at most $D + 2\delta M$ by triangle inequality, so $\mathcal{N}_{\delta M}$ is gsc-convex by Assumption 4 and Lemma 18. We also need to ensure that $d(\mathbf{x}_t', \mathcal{N}_{\delta M}) \le \frac{\mathcal{D}(\kappa_2)}{2}$ and $d(\mathbf{y}_{t+1}', \mathcal{N}) \le \frac{\mathcal{D}(\kappa_2)}{2}$ and apply Lemma 18 to show the projection is unique and non-expansive. For $d(\mathbf{x}_t', \mathcal{N}_{\delta M})$, we have

$$d(\mathbf{x}_t', \mathcal{N}_{\delta M}) \le d(\mathbf{x}_t', \mathbf{y}_t) \le \eta M \le \frac{\mathcal{D}(\kappa_2)}{2}.$$

by $\eta \le \frac{\mathcal{D}(\kappa_2)}{2(G+2M)}$. Similarly, for $d(\mathbf{y}_{t+1}', \mathcal{N})$

$$\begin{aligned} & d(\mathbf{y}_{t+1}', \mathcal{N}) \le d(\mathbf{y}_{t+1}', \mathbf{y}_t) \le d(\mathbf{y}_{t+1}', \mathbf{x}_t') + d(\mathbf{y}_t, \mathbf{x}_t') \\ \le & \| -\eta\nabla f_t(\mathbf{x}_t') + \eta\Gamma_{\mathbf{y}_t}^{\mathbf{x}_t'} M_t\| + \eta M \\ \le & \eta(G+2M) \le \frac{\mathcal{D}(\kappa_2)}{2}. \end{aligned}$$

We can bound $f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t')$ in the same way as Theorem 4, but we now use Lemmas 17 and 21 to bound $f_t(\mathbf{x}_t') - f_t(\mathbf{u}_t)$.

$$\begin{aligned} & f_t(\mathbf{x}_t') - f_t(\mathbf{u}_t) \le -\langle \mathrm{Exp}_{\mathbf{x}_t'}^{-1}(\mathbf{u}_t), \nabla f_t(\mathbf{x}_t')\rangle \\ = & \frac{1}{\eta}\langle \mathrm{Exp}_{\mathbf{x}_t'}^{-1}(\mathbf{u}_t), \mathrm{Exp}_{\mathbf{x}_t'}^{-1}(\mathbf{y}_{t+1}') - \mathrm{Exp}_{\mathbf{x}_t'}^{-1}(\mathbf{y}_t)\rangle \\ \le & \frac{1}{2\eta}\left(\zeta d(\mathbf{x}_t', \mathbf{y}_{t+1}')^2 + d(\mathbf{x}_t', \mathbf{u}_t)^2 - d(\mathbf{y}_{t+1}', \mathbf{u}_t)^2\right) - \frac{1}{2\eta}\left(\xi d(\mathbf{x}_t', \mathbf{y}_t)^2 + d(\mathbf{x}_t', \mathbf{u}_t)^2 - d(\mathbf{y}_t, \mathbf{u}_t)^2\right). \end{aligned} \tag{36}$$

Finally, we need to show

$$2\eta G + 2\eta^2\zeta L^2 d(\mathbf{x}_t', \mathbf{y}_t) - \xi d(\mathbf{x}_t', \mathbf{y}_t) \le 0 \tag{37}$$

Following the proof of Theorem 4, we find

$$\eta \le \frac{\xi\delta M}{G + (G^2 + 2\zeta\xi\delta^2 M^2 L^2)^{\frac{1}{2}}}$$

satisfies the required condition. The guarantee is thus established. ■

B.3. Proof of Lemma 4

We first show

$$\sum_{t=1}^{T}\langle \boldsymbol{\ell}_t, \mathbf{w}_t-\mathbf{w}^*\rangle \le \frac{\ln N + R(\mathbf{w}^*)}{\beta} + \beta\sum_{t=1}^{T}\|\boldsymbol{\ell}_t-\mathbf{m}_t\|_\infty^2 - \frac{1}{2\beta}\sum_{t=1}^{T}\left(\|\mathbf{w}_t-\mathbf{w}_t'\|_1^2+\|\mathbf{w}_t-\mathbf{w}_{t-1}'\|_1^2\right) \tag{38}$$

holds for any $\mathbf{w}^*\in\Delta_N$, where

$$\mathbf{w}_t = \operatorname*{argmin}_{\mathbf{w}\in\Delta_N} \beta\left\langle \sum_{s=1}^{t-1}\boldsymbol{\ell}_s+\mathbf{m}_t, \mathbf{w}\right\rangle + R(\mathbf{w}), \quad t\ge 1$$

and

$$\mathbf{w}_t' = \operatorname*{argmin}_{\mathbf{w}\in\Delta_N} \beta\left\langle \sum_{s=1}^{t}\boldsymbol{\ell}_s, \mathbf{w}\right\rangle + R(\mathbf{w}), \quad t\ge 0.$$

Note that $R(\mathbf{w})=\sum_{i\in[N]} w_i\ln w_i$ is the negative entropy. According to the equivalence between Hedge and follow the regularized leader with the negative entropy regularizer, we have $w_{t,i}\propto e^{-\beta(\sum_{s=1}^{t-1}\ell_{s,i}+m_{t,i})}$ and $w_{t,i}'\propto e^{-\beta(\sum_{s=1}^{t}\ell_{s,i})}$. To prove Equation (38), we consider the following decomposition:

$$\langle \boldsymbol{\ell}_t, \mathbf{w}_t-\mathbf{w}^*\rangle = \langle \boldsymbol{\ell}_t-\mathbf{m}_t, \mathbf{w}_t-\mathbf{w}_t'\rangle + \langle \mathbf{m}_t, \mathbf{w}_t-\mathbf{w}_t'\rangle + \langle \boldsymbol{\ell}_t, \mathbf{w}_t'-\mathbf{w}^*\rangle .$$

Since $R(\cdot)$ is 1-strongly convex w.r.t. the $\boldsymbol{\ell}_1$ norm, by Lemma 31, we have $\|\mathbf{w}_t-\mathbf{w}_t'\|_1\le\beta\|\boldsymbol{\ell}_t-\mathbf{m}_t\|_\infty$ and

$$\langle \boldsymbol{\ell}_t-\mathbf{m}_t, \mathbf{w}_t-\mathbf{w}_t'\rangle \le \|\boldsymbol{\ell}_t-\mathbf{m}_t\|_\infty\|\mathbf{w}_t-\mathbf{w}_t'\|_1 \le \beta\|\boldsymbol{\ell}_t-\mathbf{m}_t\|_\infty^2$$

by Hölder's inequality. Hence it suffices to show

$$\sum_{t=1}^{T}\langle \mathbf{m}_t, \mathbf{w}_t-\mathbf{w}_t'\rangle + \langle \boldsymbol{\ell}_t, \mathbf{w}_t'-\mathbf{w}^*\rangle \le \frac{\ln N + R(\mathbf{w}^*)}{\beta} - \frac{1}{2\beta}\sum_{t=1}^{T}\left(\|\mathbf{w}_t-\mathbf{w}_t'\|_1^2+\|\mathbf{w}_t-\mathbf{w}_{t-1}'\|_1^2\right) \tag{39}$$

to prove Equation (38).

Equation (39) holds for $T=0$ because $R(\mathbf{w}^*)\ge -\ln N$ holds for any $\mathbf{w}^*\in\Delta_N$. To proceed, we need the following proposition:

$$g(\mathbf{w}^*) + \frac{c}{2}\|\mathbf{w}-\mathbf{w}^*\|^2 \le g(\mathbf{w}) \tag{40}$$

holds for any c-strongly convex $g(\cdot):\mathcal{W}\to\mathbb{R}$ where $\mathbf{w}^*=\operatorname{argmin}_{\mathbf{w}\in\mathcal{W}} g(\mathbf{w})$. This fact can be easily seen by combining the strong convexity and the first-order optimality condition for convex functions.

Assume Equation (39) holds for round $T-1$ ($T\ge 1$) and we denote

$$C_T = \frac{1}{2\beta}\sum_{t=1}^{T}\left(\|\mathbf{w}_t-\mathbf{w}_t'\|_1^2+\|\mathbf{w}_t-\mathbf{w}_{t-1}'\|_1^2\right).$$

Now for round T,

$$\begin{aligned}
&\sum_{t=1}^{T}\left(\left\langle \mathbf{m}_t, \mathbf{w}_t - \mathbf{w}'_t\right\rangle + \left\langle \boldsymbol{\ell}_t, \mathbf{w}'_t\right\rangle\right)\\
\overset{(1)}{\leq}&\sum_{t=1}^{T-1}\left\langle \boldsymbol{\ell}_t, \mathbf{w}'_{T-1}\right\rangle + \frac{\ln N + R(\mathbf{w}'_{T-1})}{\beta} - C_{T-1} + \left\langle \mathbf{m}_T, \mathbf{w}_T - \mathbf{w}'_T\right\rangle + \left\langle \boldsymbol{\ell}_T, \mathbf{w}'_T\right\rangle\\
\overset{(2)}{\leq}&\sum_{t=1}^{T-1}\left\langle \boldsymbol{\ell}_t, \mathbf{w}_T\right\rangle + \frac{\ln N + R(\mathbf{w}_T)}{\beta} - C_{T-1} + \left\langle \mathbf{m}_T, \mathbf{w}_T - \mathbf{w}'_T\right\rangle + \left\langle \boldsymbol{\ell}_T, \mathbf{w}'_T\right\rangle - \frac{1}{2\beta}\|\mathbf{w}_T - \mathbf{w}'_{T-1}\|_1^2\\
=&\left(\sum_{t=1}^{T-1}\left\langle \boldsymbol{\ell}_t, \mathbf{w}_T\right\rangle + \left\langle \mathbf{m}_T, \mathbf{w}_T\right\rangle + \frac{\ln N + R(\mathbf{w}_T)}{\beta}\right) + \left\langle \boldsymbol{\ell}_T - \mathbf{m}_T, \mathbf{w}'_T\right\rangle - C_{T-1} - \frac{1}{2\beta}\|\mathbf{w}_T - \mathbf{w}'_{T-1}\|_1^2\\
\overset{(3)}{\leq}&\left(\sum_{t=1}^{T-1}\left\langle \boldsymbol{\ell}_t, \mathbf{w}'_T\right\rangle + \left\langle \mathbf{m}_T, \mathbf{w}'_T\right\rangle + \frac{\ln N + R(\mathbf{w}'_T)}{\beta}\right) + \left\langle \boldsymbol{\ell}_T - \mathbf{m}_T, \mathbf{w}'_T\right\rangle\\
&- C_{T-1} - \frac{1}{2\beta}\|\mathbf{w}_T - \mathbf{w}'_{T-1}\|_1^2 - \frac{1}{2\beta}\|\mathbf{w}_T - \mathbf{w}'_T\|_1^2\\
=&\sum_{t=1}^{T}\left\langle \boldsymbol{\ell}_t, \mathbf{w}'_T\right\rangle + \frac{\ln N + R(\mathbf{w}'_T)}{\beta} - C_T\\
\overset{(4)}{\leq}&\sum_{t=1}^{T}\left\langle \boldsymbol{\ell}_t, \mathbf{w}^*\right\rangle + \frac{\ln N + R(\mathbf{w}^*)}{\beta} - C_T.
\end{aligned} \tag{41}$$

The first inequality is due to the induction hypothesis with $\mathbf{w}^\star = \mathbf{w}'_{T-1}$. The second and the third ones are applications of Equation (40). Specifically, $\mathbf{w}'_{T-1}$ and $\mathbf{w}_T$ minimize $\sum_{t=1}^{T-1}\beta\left\langle \boldsymbol{\ell}_t, \mathbf{w}\right\rangle + R(\mathbf{w})$ and $\sum_{t=1}^{T-1}\beta\left\langle \boldsymbol{\ell}_t, \mathbf{w}\right\rangle + \beta\left\langle \mathbf{m}_T, \mathbf{w}\right\rangle + R(\mathbf{w})$ respectively. The forth inequality follows from $\mathbf{w}'_T$ minimizes $\beta\sum_{t=1}^{T}\left\langle \boldsymbol{\ell}_t, \mathbf{w}\right\rangle + R(\mathbf{w})$.

We now demonstrate how to remove the dependence on $\mathbf{w}'_t$:

$$\begin{aligned}
&\frac{1}{2\beta}\sum_{t=1}^{T}\left(\|\mathbf{w}_t - \mathbf{w}'_t\|_1^2 + \|\mathbf{w}_t - \mathbf{w}'_{t-1}\|_1^2\right)\\
\geq&\frac{1}{2\beta}\sum_{t=1}^{T}\left(\|\mathbf{w}_t - \mathbf{w}'_t\|_1^2 + \|\mathbf{w}_{t+1} - \mathbf{w}'_t\|_1^2\right) - \frac{1}{2\beta}\|\mathbf{w}_{T+1} - \mathbf{w}'_T\|_1^2\\
\geq&\frac{1}{4\beta}\sum_{t=2}^{T}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_1^2 - \frac{2}{\beta},
\end{aligned} \tag{42}$$

where the last inequality follows from $\|\mathbf{x} + \mathbf{y}\|^2 \leq 2(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2)$ holds for any norm. Now the proof is completed by combining Equations (38) and (42).

B.4. Proof of Theorem 5

We first show $f_t(\mathbf{x}_t) - f_t(\mathbf{x}_{t,i}) \leq \langle \mathbf{w}_t, \boldsymbol{\ell}_t \rangle - \ell_{t,i}$ so that Lemma 4 can be invoked to bound the regret of the meta algorithm. We start from gsc-convexity,

$$\begin{aligned} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_{t,i}) &\leq - \left\langle \nabla f_t(\mathbf{x}_t), \mathrm{Exp}_{\mathbf{x}_t}^{-1}\mathbf{x}_{t,i} \right\rangle \\ =& \left\langle \nabla f_t(\mathbf{x}_t), \sum_{j=1}^N w_{t,j} \mathrm{Exp}_{\mathbf{x}_t}^{-1}\mathbf{x}_{t,j} \right\rangle - \left\langle \nabla f_t(\mathbf{x}_t), \mathrm{Exp}_{\mathbf{x}_t}^{-1}\mathbf{x}_{t,i} \right\rangle \\ =& \sum_{j=1}^N w_{t,j}\ell_{t,j} - \ell_{t,i} = \langle \mathbf{w}_t, \boldsymbol{\ell}_t \rangle - \ell_{t,i}, \end{aligned} \tag{43}$$

where the first equality is due to the gradient of $\sum_{i=1}^N w_{t,i} d(\mathbf{x}, \mathbf{x}_{t,i})^2$ vanishes at $\mathbf{x}_t$. Now we can bound the regret as

$$\begin{aligned} \sum_{t=1}^T f_t(\mathbf{x}_t) - f_t(\mathbf{x}_{t,i}) &\leq \sum_{t=1}^T \langle \mathbf{w}_t, \boldsymbol{\ell}_t \rangle - \ell_{t,i} \\ \leq& \frac{2 + \ln N}{\beta} + \beta \sum_{t=1}^T \|\boldsymbol{\ell}_t - \mathbf{m}_t\|_\infty^2 - \frac{1}{4\beta} \sum_{t=2}^T \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_1^2 \end{aligned} \tag{44}$$

It now suffices to bound $\|\boldsymbol{\ell}_t - \mathbf{m}_t\|_\infty^2$ in terms of the gradient-variation V_T and $\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_1^2$. We start with the definition of the infinity norm.

$$\begin{aligned} & \|\boldsymbol{\ell}_t - \mathbf{m}_t\|_\infty^2 \\ =& \max_{i \in [N]} (\ell_{t,i} - m_{t,i})^2 \\ =& \max_{i \in [N]} \left(\left\langle \nabla f_t(\mathbf{x}_t), \mathrm{Exp}_{\mathbf{x}_t}^{-1}\mathbf{x}_{t,i} \right\rangle - \left\langle \nabla f_{t-1}(\bar{\mathbf{x}}_t), \mathrm{Exp}_{\bar{\mathbf{x}}_t}^{-1}\mathbf{x}_{t,i} \right\rangle \right)^2 \\ =& \max_{i \in [N]} \Big(\left\langle \nabla f_t(\mathbf{x}_t), \mathrm{Exp}_{\mathbf{x}_t}^{-1}\mathbf{x}_{t,i} \right\rangle - \left\langle \nabla f_{t-1}(\mathbf{x}_t), \mathrm{Exp}_{\mathbf{x}_t}^{-1}\mathbf{x}_{t,i} \right\rangle + \left\langle \nabla f_{t-1}(\mathbf{x}_t), \mathrm{Exp}_{\mathbf{x}_t}^{-1}\mathbf{x}_{t,i} \right\rangle \\ & - \left\langle \Gamma_{\bar{\mathbf{x}}_t}^{\mathbf{x}_t} \nabla f_{t-1}(\bar{\mathbf{x}}_t), \mathrm{Exp}_{\mathbf{x}_t}^{-1}\mathbf{x}_{t,i} \right\rangle + \left\langle \Gamma_{\bar{\mathbf{x}}_t}^{\mathbf{x}_t} \nabla f_{t-1}(\bar{\mathbf{x}}_t), \mathrm{Exp}_{\mathbf{x}_t}^{-1}\mathbf{x}_{t,i} \right\rangle - \left\langle \Gamma_{\bar{\mathbf{x}}_t}^{\mathbf{x}_t} \nabla f_{t-1}(\bar{\mathbf{x}}_t), \Gamma_{\bar{\mathbf{x}}_t}^{\mathbf{x}_t} \mathrm{Exp}_{\bar{\mathbf{x}}_t}^{-1}\mathbf{x}_{t,i} \right\rangle \Big)^2 \\ \overset{(1)}{\leq}& 3 \max_{i \in [N]} \Big(\left\langle \nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_t), \mathrm{Exp}_{\mathbf{x}_t}^{-1}\mathbf{x}_{t,i} \right\rangle^2 + \left\langle \nabla f_{t-1}(\mathbf{x}_t) - \Gamma_{\bar{\mathbf{x}}_t}^{\mathbf{x}_t} \nabla f_{t-1}(\bar{\mathbf{x}}_t), \mathrm{Exp}_{\mathbf{x}_t}^{-1}\mathbf{x}_{t,i} \right\rangle^2 \\ & + \left\langle \Gamma_{\bar{\mathbf{x}}_t}^{\mathbf{x}_t} \nabla f_{t-1}(\bar{\mathbf{x}}_t), \mathrm{Exp}_{\mathbf{x}_t}^{-1}\mathbf{x}_{t,i} - \Gamma_{\bar{\mathbf{x}}_t}^{\mathbf{x}_t} \mathrm{Exp}_{\bar{\mathbf{x}}_t}^{-1}\mathbf{x}_{t,i} \right\rangle^2 \Big) \\ \overset{(2)}{\leq}& 3 \left(D^2 \sup_{\mathbf{x} \in \mathcal{N}} \|\nabla f_t(\mathbf{x}) - \nabla f_{t-1}(\mathbf{x})\|^2 + D^2 L^2 d(\mathbf{x}_t, \bar{\mathbf{x}}_t)^2 + G^2 \|\mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}_{t,i}) - \Gamma_{\bar{\mathbf{x}}_t}^{\mathbf{x}_t} \mathrm{Exp}_{\bar{\mathbf{x}}_t}^{-1}(\mathbf{x}_{t,i})\|^2 \right) \end{aligned} \tag{45}$$

where the first inequality relies on fact that $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ holds for any $a, b, c \in \mathbb{R}$, and the second one follows from Assumptions 2, 3, L-gsc-smoothness and Hölder's inequality.

By Lemma 27, for a Hadamard manifold with sectional curvature lower bounded by κ, $h(\mathbf{x}) := \frac{1}{2} d(\mathbf{x}, \mathbf{x}_{t,i})^2$ is $\frac{\sqrt{\kappa}D}{\tanh(\sqrt{\kappa}D)}$-smooth (which is exactly ζ as in Definition 1) on Hadamard manifolds. Thus

$$\|\mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}_{t,i}) - \Gamma_{\bar{\mathbf{x}}_t}^{\mathbf{x}_t} \mathrm{Exp}_{\bar{\mathbf{x}}_t}^{-1}(\mathbf{x}_{t,i})\| = \| - \nabla h(\mathbf{x}_t) + \Gamma_{\bar{\mathbf{x}}_t}^{\mathbf{x}_t} \nabla h(\bar{\mathbf{x}}_t)\| \leq \zeta d(\mathbf{x}_t, \bar{\mathbf{x}}_t). \tag{46}$$

We need to bound $d(\mathbf{x}_t, \bar{\mathbf{x}}_t)$ in terms of $\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_1$ to make full use of the negative term in Lemma 4. By Lemma 20

$$d(\mathbf{x}_t, \bar{\mathbf{x}}_t) \leq \sum_{i=1}^{N} w_{t,i} \cdot d(\mathbf{x}_{t,i}, \mathbf{x}_{t,i}) + D\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_1 = D\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_1. \tag{47}$$

Combining Equations (44), (45), (46) and (47), we have

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_{t,i}) \leq \frac{2 + \ln N}{\beta} + 3\beta D^2(V_T + G^2) + \sum_{t=2}^{T} \left(3\beta(D^4L^2 + D^2G^2\zeta^2) - \frac{1}{4\beta}\right) \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_1^2, \tag{48}$$

where the $3\beta D^2G^2$ term is due to the calculation of V_T starts from $t = 2$, while $\mathbf{w}_0 = \mathbf{w}_1$ ensures $\mathbf{x}_1 = \bar{\mathbf{x}}_1$ and thus $d(\mathbf{x}_1, \bar{\mathbf{x}}_1) = 0$.

B.5. Proof of Theorem 6

The optimal step size, according to Theorem 4 is

$$\eta^\star = \min\left\{\frac{\delta}{1 + (1 + 2\zeta\delta^2L^2)^{\frac{1}{2}}}, \sqrt{\frac{D^2 + 2DP_T}{2\zeta V_T}}\right\}.$$

Based on Assumption 3, we know V_T has an upper bound $V_T = \sum_{t=2}^{T} \sup_{\mathbf{x}\in\mathcal{N}} \|\nabla f_t(\mathbf{x}) - \nabla f_{t-1}(\mathbf{x})\|^2 \leq 4G^2T$. Therefore, $\eta^\star$ can be bounded by

$$\sqrt{\frac{D^2}{8\zeta G^2T}} \leq \eta^\star \leq \frac{\delta}{1 + (1 + 2\zeta\delta^2L^2)^{\frac{1}{2}}}.$$

According to the construction of $\mathcal{H}$,

$$\min\mathcal{H} = \sqrt{\frac{D^2}{8\zeta G^2T}}, \qquad \max\mathcal{H} \geq 2\frac{\delta}{1 + (1 + 2\zeta\delta^2L^2)^{\frac{1}{2}}}.$$

Therefore, there always exists $k \in [N]$ such that $\eta_k \leq \eta^\star \leq 2\eta_k$. We can bound the regret of the k-th expert as

$$\begin{aligned}
\sum_{t=1}^{T} f_t(\mathbf{x}_{t,k}) - \sum_{t=1}^{T} f_t(\mathbf{u}_t) &\leq \eta_k\zeta V_T + \frac{D^2 + 2DP_T}{2\eta_k} \leq \eta^\star\zeta V_T + \frac{D^2 + 2DP_T}{\eta^\star} \\
&\leq \zeta V_T\sqrt{\frac{D^2 + 2DP_T}{2\zeta V_T}} + (D^2 + 2DP_T) \cdot \max\left\{\sqrt{\frac{2\zeta V_T}{D^2 + 2DP_T}}, \frac{1 + (1 + 2\zeta\delta^2L^2)^{\frac{1}{2}}}{\delta}\right\} \\
&= \frac{3}{2}\sqrt{2(D^2 + 2DP_T)\zeta V_T} + (D^2 + 2DP_T)\frac{1 + (1 + 2\zeta\delta^2L^2)^{\frac{1}{2}}}{\delta}.
\end{aligned} \tag{49}$$

Since the dynamic regret can be decomposed as the sum of the meta-regret and the expert-regret,

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}_t) = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_{t,k}) + \sum_{t=1}^{T} f_t(\mathbf{x}_{t,k}) - \sum_{t=1}^{T} f_t(\mathbf{u}_t).$$

Applying Theorem 5 with $\beta \leq \frac{1}{\sqrt{12(D^4L^2+D^2G^2\zeta^2)}}$, we have

$$\left(3\beta(D^4L^2 + D^2G^2\zeta^2) - \frac{1}{4\beta}\right) \leq 0$$

and

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_{t,i}) \leq \frac{2+\ln N}{\beta} + 3\beta D^2(V_T + G^2).$$

We need to consider two cases based on the value of β.

If $\sqrt{\frac{2+\ln N}{3D^2(V_T+G^2)}} \leq \frac{1}{\sqrt{12(D^4L^2+D^2G^2\zeta^2)}}$, then

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_{t,i}) \leq 2\sqrt{3D^2(V_T + G^2)(2+\ln N)}.$$

Otherwise, we have

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_{t,i}) \leq 2(2+\ln N)\sqrt{12(D^4L^2 + D^2G^2\zeta^2)}.$$

In sum,

$$\begin{aligned}
&\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_{t,i}) \\
\leq& \max\left\{2\sqrt{3D^2(V_T + G^2)(2+\ln N)}, 2(2+\ln N)\sqrt{12(D^4L^2 + D^2G^2\zeta^2)}\right\}.
\end{aligned} \tag{50}$$

Combining Equations (50) and (49), we have

$$\begin{aligned}
&\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}_t) \\
\leq& \max\left\{2\sqrt{3D^2(V_T + G^2)(2+\ln N)}, 2(2+\ln N)\sqrt{12(D^4L^2 + D^2G^2\zeta^2)}\right\} \\
&+ \frac{3}{2}\sqrt{2(D^2+2DP_T)\zeta V_T} + (D^2+2DP_T)\frac{1+(1+2\zeta\delta^2L^2)^{\frac{1}{2}}}{\delta} \\
=& O\left(\sqrt{(V_T + \zeta^2\ln N)\ln N}\right) + O\left(\sqrt{\zeta(V_T + (1+P_T)/\delta^2)(1+P_T)}\right) \\
=& O\left(\sqrt{\zeta(V_T + (1+P_T)/\delta^2)(1+P_T)}\right),
\end{aligned} \tag{51}$$

where we use $O(\cdot)$ to hide $O(\log\log T)$ following Luo and Schapire (2015) and Zhao et al. (2020) and $N = O(\log T)$ leads to $\ln N = O(\log\log T)$.

Appendix C. Omitted Proof for Section 6

C.1. Proof of Lemma 5

By L-gsc-smoothness, we have

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \left\langle \nabla f(\mathbf{x}), \mathrm{Exp}_{\mathbf{x}}^{-1}(\mathbf{y}) \right\rangle + \frac{L \cdot d(\mathbf{x}, \mathbf{y})^2}{2}.$$

Setting $\mathbf{y} = \mathrm{Exp}_{\mathbf{x}}\left(-\frac{1}{L}\nabla f(\mathbf{x})\right)$, we have

$$\begin{aligned} 0 \leq & f(\mathbf{y}) \leq f(\mathbf{x}) - \frac{1}{L}\|\nabla f(\mathbf{x})\|^2 + \frac{L}{2} \cdot \frac{1}{L^2}\|\nabla f(\mathbf{x})\|^2 \\ = & f(\mathbf{x}) - \frac{1}{2L}\|\nabla f(\mathbf{x})\|^2, \end{aligned} \tag{52}$$

where we use the non-negativity of f. The above inequality is equivalent to

$$\|\nabla f(\mathbf{x})\|^2 \leq 2L \cdot f(\mathbf{x}),$$

in which the constant is two times better than that of Srebro et al. (2010).

C.2. Proof of Lemma 6

The proof is similar to the proof of Theorem 1. Let $\mathbf{x}'_{t+1} = \mathrm{Exp}_{\mathbf{x}_t}\left(-\eta \nabla f_t(\mathbf{x}_t)\right)$, then analog to Equation (10), we have

$$\begin{aligned} f_t(\mathbf{x}_t) - f_t(\mathbf{u}_t) \leq & \frac{1}{2\eta}\left(\|\mathrm{Exp}_{\mathbf{x}_t}^{-1}\mathbf{u}_t\|^2 - \|\mathrm{Exp}_{\mathbf{x}_{t+1}}^{-1}\mathbf{u}_{t+1}\|^2 + 2D\|\mathrm{Exp}_{\mathbf{u}_t}^{-1}\mathbf{u}_{t+1}\|\right) + \frac{\eta\zeta\|\nabla f_t(\mathbf{x}_t)\|^2}{2} \\ \leq & \frac{1}{2\eta}\left(\|\mathrm{Exp}_{\mathbf{x}_t}^{-1}\mathbf{u}_t\|^2 - \|\mathrm{Exp}_{\mathbf{x}_{t+1}}^{-1}\mathbf{u}_{t+1}\|^2 + 2D\|\mathrm{Exp}_{\mathbf{u}_t}^{-1}\mathbf{u}_{t+1}\|\right) + \eta\zeta L f_t(\mathbf{x}_t), \end{aligned} \tag{53}$$

where for the second inequality we apply Lemma 5. WLOG, we can assume $\mathbf{u}_{T+1} = \mathbf{u}_T$ and sum from $t = 1$ to T:

$$\sum_{t=1}^{T}(f_t(\mathbf{x}_t) - f_t(\mathbf{u}_t)) \leq \frac{D^2 + 2DP_T}{2\eta} + \eta\zeta L\sum_{t=1}^{T} f_t(\mathbf{x}_t).$$

After simplifying, we get

$$\begin{aligned} \sum_{t=1}^{T}(f_t(\mathbf{x}_t) - f_t(\mathbf{u}_t)) \leq & \frac{D^2 + 2DP_T}{2\eta(1-\eta\zeta L)} + \frac{\eta\zeta L\sum_{t=1}^{T} f_t(\mathbf{u}_t)}{1-\eta\zeta L} \\ = & \frac{D^2 + 2DP_T}{2\eta(1-\eta\zeta L)} + \frac{\eta\zeta L F_T}{1-\eta\zeta L} \\ \leq & \frac{D^2 + 2DP_T}{\eta} + 2\eta\zeta L F_T \\ = & O\left(\frac{1+P_T}{\eta} + \eta F_T\right). \end{aligned} \tag{54}$$

where $\eta \leq \frac{1}{2\zeta L}$ is used to obtain the second inequality.

C.3. Proof of Lemma 7

We again apply Lemma 4, with the surrogate loss $\ell_{t,i} = \left\langle \nabla f_t(\mathbf{x}_t), \operatorname{Exp}_{\mathbf{x}_t}^{-1} \mathbf{x}_{t,i} \right\rangle$ and the optimism $m_{t,i} = 0$ for any $i \in [N]$. In this way,

$$\begin{aligned}
&\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_{t,i}) \\
\leq& -\sum_{t=1}^{T} \left\langle \nabla f_t(\mathbf{x}_t), \operatorname{Exp}_{\mathbf{x}_t}^{-1} \mathbf{x}_{t,i} \right\rangle = \sum_{t=1}^{T} \langle \mathbf{w}_t, \boldsymbol{\ell}_t \rangle - w_{t,i} \\
\leq& \frac{2+\ln N}{\beta} + \beta \sum_{t=1}^{T} \|\boldsymbol{\ell}_t\|_\infty^2 - \frac{1}{4\beta} \sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_1^2 \qquad (55) \\
\leq& \frac{2+\ln N}{\beta} + \beta D^2 \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t)\|^2 \\
\leq& \frac{2+\ln N}{\beta} + 2\beta D^2 L \sum_{t=1}^{T} f_t(\mathbf{x}_t) = \frac{2+\ln N}{\beta} + 2\beta D^2 L \bar{F}_T,
\end{aligned}$$

where the second inequality follows from Lemma 4, the third one follows from Assumption 2 and Hölder's inequality, while the last inequality is due to Lemma 5. By setting $\beta = \sqrt{\frac{2+\ln N}{2LD^2 \bar{F}_T}}$, the regret of the meta algorithm is upper bounded by

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_{t,i}) \leq \sqrt{8D^2 L(2+\ln N)\bar{F}_T} = \sqrt{8D^2 L(2+\ln N) \sum_{t=1}^{T} f_t(\mathbf{x}_t)}. \qquad (56)$$

Although $\bar{F}_T$ is unknown similar to the case of Optimistic Hedge, we can use techniques like the doubling trick or a time-varying step size $\beta_t = O\left(\frac{1}{\sqrt{1+\bar{F}_t}}\right)$ to overcome this hurdle.

The RHS of Equation (56) depends on the cumulative loss of $\mathbf{x}_t$, which remains elusive. Here we apply an algebraic fact that $x - y \leq \sqrt{ax}$ implies $x - y \leq a + \sqrt{ay}$ holds for any non-negative x, y and a. Then

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_{t,i}) \leq 8D^2 L(2+\ln N) + \sqrt{8D^2 L(2+\ln N)\bar{F}_{T,i}} \qquad (57)$$

where we remind $\bar{F}_{T,i} = \sum_{t=1}^{T} f_t(\mathbf{x}_{t,i})$.

C.4. Proof of Theorem 7

Recall the regret of the meta algorithm as in Lemma 7:

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_{t,i}) \leq 8D^2 L(2+\ln N) + \sqrt{8D^2 L(2+\ln N)\bar{F}_{T,i}}. \qquad (58)$$

On the other hand, we know the regret can be decomposed as

$$\begin{aligned}\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}_t) =& \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_{t,i}) + \sum_{t=1}^{T} f_t(\mathbf{x}_{t,i}) - \sum_{t=1}^{T} f_t(\mathbf{u}_t) \\ =& \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_{t,i}) + \bar{F}_{T,i} - F_T.\end{aligned} \tag{59}$$

We can first show there exists an almost optimal step size and bound the regret of the corresponding expert. That regret immediately provides an upper bound of $\bar{F}_{T,i}$ in terms of F_T. This argument eliminates the dependence on $\bar{F}_{T,i}$ and leads to a regret bound solely depending on F_T and P_T.

Now we bound the regret of the best expert. Note that due to Lemma 6, the optimal step size is $\eta = \min\left\{\frac{1}{2\zeta L}, \sqrt{\frac{D^2+2DP_T}{2\zeta L F_T}}\right\}$. According to Assumptions 2, 3, $F_T \le GDT$. Thus the optimal step size $\eta^\star$ is bounded by

$$\sqrt{\frac{D}{4LGT}} \le \eta^\star \le \frac{1}{2\zeta L}.$$

Due to our construction of $\mathcal{H}$, there exists $k \in [N]$ such that $\eta_k \le \eta^\star \le 2\eta_k$.

According to Lemma 6,

$$\begin{aligned}&\sum_{t=1}^{T} f_t(\mathbf{x}_{t,k}) - \sum_{t=1}^{T} f_t(\mathbf{u}_t) \\ \le& \frac{D^2+2DP_T}{2\eta_k(1-\eta_k\zeta L)} + \frac{\eta_k \zeta L F_T}{1-\eta_k\zeta L} \le \frac{D^2+2DP_T}{\eta_k} + 2\eta_k\zeta L F_T \\ \le& \frac{2(D^2+2DP_T)}{\eta^\star} + 2\eta^\star \zeta L F_T \\ \le& 2(D^2+2DP_T)\left(2\zeta L + \sqrt{\frac{2\zeta L F_T}{D^2+2DP_T}}\right) + 2\zeta L F_T \cdot \sqrt{\frac{D^2+2DP_T}{2\zeta L F_T}} \\ =& 4\zeta L(D^2+2DP_T) + 3\sqrt{2\zeta L F_T(D^2+2DP_T)} \\ \le& \sqrt{2\left(16\zeta^2L^2(D^2+2DP_T)^2 + 18\zeta L F_T(D^2+2DP_T)\right)}\end{aligned} \tag{60}$$

where we apply $\eta^\star \le \sqrt{\frac{D^2+2DP_T}{2\zeta L F_T}}$, $\frac{1}{\eta^\star} \le 2\zeta L + \sqrt{\frac{2\zeta L F_T}{D^2+2DP_T}}$ and $\sqrt{a}+\sqrt{b} \le \sqrt{2(a+b)}$.

Now as we combine Equations (58), (59), (60),

$$\begin{aligned}&\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}_t) \\ \le& 8D^2L(2+\ln N) + \sqrt{8D^2L(2+\ln N)\bar{F}_{T,k}} + \sqrt{2\left(16\zeta^2L^2(D^2+2DP_T)^2 + 18\zeta L F_T(D^2+2DP_T)\right)} \\ \le& 8D^2L(2+\ln N) + \sqrt{8D^2L(2+\ln N)\left(F_T + \sqrt{2\left(16\zeta^2L^2(D^2+2DP_T)^2 + 18\zeta L F_T(D^2+2DP_T)\right)}\right)} \\ &+ \sqrt{2\left(16\zeta^2L^2(D^2+2DP_T)^2 + 18\zeta L F_T(D^2+2DP_T)\right)} \\ =& O(\sqrt{\zeta(\zeta(1+P_T)+F_T)(P_T+1)}).\end{aligned} \tag{61}$$

where we again use $O(\cdot)$ to hide the $\log\log T$ term.

Appendix D. Omitted Proof for Section 7

D.1. Necessity of Best-of-both-worlds Bound

We highlight the necessity of achieving a best-of-both-worlds bound by computing the Fréchet mean in the online setting on the d-dimensional unit Poincaré disk. The Poincaré disk looks like a unit ball in Euclidean space, but its Riemannian metric blows up near the boundary:

$$\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{x}} = \frac{4 \langle \mathbf{u}, \mathbf{v} \rangle_2}{(1 - \|\mathbf{x}\|_2^2)^2}$$

and has constant sectional curvature -1. We use $\mathbf{0}$ to denote the origin of the Poincaré ball and $\mathbf{e}_i$ to be the i-th unit vector in the standard basis. The Poincaré ball has the following property (Lou et al., 2020):

$$\begin{aligned} d(\mathbf{x}, \mathbf{y}) &= \operatorname{arcosh}\left(1 + \frac{2\|\mathbf{x} - \mathbf{y}\|_2^2}{(1 - \|\mathbf{x}\|_2^2)(1 - \|\mathbf{y}\|_2^2)}\right) \\ \operatorname{Exp}_{\mathbf{0}}^{-1} \mathbf{y} &= \operatorname{arctanh}(\|\mathbf{y}\|_2) \frac{\mathbf{y}}{\|\mathbf{y}\|_2}. \end{aligned}$$

Now consider the following loss function

$$f_t(\mathbf{x}) = \sum_{i=1}^{2d} \frac{d(\mathbf{x}, \mathbf{x}_{t,i})^2}{2d}$$

where $\mathbf{x}_{t,i} = \frac{t}{2T}\mathbf{e}_i$ for $1 \leq i \leq d$ and $\mathbf{x}_{t,i} = -\frac{t}{2T}\mathbf{e}_{i-d}$ for $d + 1 \leq i \leq 2d$. We choose $\mathcal{N}$ to be the convex hull of $\pm\frac{1}{2}\mathbf{e}_i$, $i = 1, \ldots, d$. And the comparator is $\mathbf{u}_t = \operatorname{argmin}_{\mathbf{u}_t \in \mathcal{N}} f_t(\mathbf{u}_t)$ which is indeed the origin $\mathbf{0}$ due to symmetry. Now we can bound V_T by

$$\begin{aligned} V_T =& \sum_{t=2}^{T} \sup_{\mathbf{x} \in \mathcal{N}} \|\nabla f_t(\mathbf{x}) - \nabla f_{t-1}(\mathbf{x})\|^2 \\ =& \frac{1}{4d^2} \sum_{t=2}^{T} \sup_{\mathbf{x} \in \mathcal{N}} \left\| \sum_{i=1}^{2d} \left(\operatorname{Exp}_{\mathbf{x}}^{-1} \mathbf{x}_{t,i} - \operatorname{Exp}_{\mathbf{x}}^{-1} \mathbf{x}_{t-1,i} \right) \right\|^2 \\ \leq& \frac{1}{4d^2} \sum_{t=2}^{T} c \cdot \left(\sum_{i=1}^{2d} d(\mathbf{x}_{t,i}, \mathbf{x}_{t-1,i}) \right)^2 = \sum_{t=2}^{T} c \left(\operatorname{arcosh}\left(1 + \frac{2(\frac{t}{2T} - \frac{t-1}{2T})^2}{(1 - (\frac{t}{2T})^2)(1 - (\frac{t-1}{2T})^2)} \right) \right)^2 \\ \leq& \sum_{t=2}^{T} c \left(\operatorname{arcosh}\left(1 + \frac{8}{9T^2} \right) \right)^2 \\ \leq& \sum_{t=2}^{T} c \cdot \left(\frac{8}{9T^2} + \sqrt{\frac{64}{81T^4} + \frac{16}{9T^2}} \right)^2 = \sum_{t=2}^{T} c \cdot O\left(\frac{1}{T^2}\right) = O\left(\frac{1}{T}\right), \end{aligned} \tag{62}$$

where the first inequality is due to triangle inequality and Lemma 30. We note that c is a constant depending on the diamater of $\mathcal{N}$ and the sectional curvature of $\mathcal{M}$. The second one is due to $t \leq T$, while the third inequality follows from $\operatorname{arcosh}(1+x) \leq x + \sqrt{x^2+2x}$.

Similarly, we can evaluate F_T

$$\begin{aligned}
F_T =& \sum_{t=1}^{T} f_t(\mathbf{u}_t) = \sum_{t=1}^{T} f_t(\mathbf{0}) = \sum_{t=1}^{T}\sum_{i=1}^{2d} \frac{d(\mathbf{0}, \mathbf{x}_{t,i})^2}{2d} \\
=& \sum_{t=1}^{T} \operatorname{arcosh}\left(\frac{1+\frac{t^2}{4T^2}}{1-\frac{t^2}{4T^2}}\right) = \int_0^T \operatorname{arcosh}\left(\frac{1+\frac{t^2}{4T^2}}{1-\frac{t^2}{4T^2}}\right) dt + O(1) \\
=& 2T \int_0^{\frac{1}{2}} \operatorname{arcosh} \frac{1+a^2}{1-a^2} da + O(1) \\
=& 2T \left(a \cdot \operatorname{arcosh} \frac{1+a^2}{1-a^2} + \ln(1-a^2)\right) |_0^{\frac{1}{2}} + O(1) = \Theta(T).
\end{aligned} \tag{63}$$

When the input losses change smoothly, $V_T \ll F_T$ and the gradient-variation bound is much tighter than the small-loss bound.

There also exist scenarios in which the small-loss bound is tighter. We still consider computing the Fréchet mean on the Poincaré disk

$$f_t(\mathbf{x}) = \sum_{i=1}^{n} d(\mathbf{x}, \mathbf{x}_{t,i})^2/n,$$

but assume $\mathbf{x}_{t,i} = \mathbf{y}_i$ when t is odd and $x_{t,i} = -\mathbf{y}_i$ when t is even. We restrict $\mathbf{y}_1, \dots, \mathbf{y}_n \in \mathbb{B}(\frac{\mathbf{e}_1}{2}, T^{-\alpha})$ where $\mathbb{B}(\mathbf{p}, r)$ is the geodesic ball centered at $\mathbf{p}$ and with radius r. $\mathcal{N}$ is the convex hull of $\mathbb{B}(\frac{\mathbf{e}_1}{2}, T^{-\alpha}) \cup \mathbb{B}(-\frac{\mathbf{e}_1}{2}, T^{-\alpha})$. Since the input sequence is alternating, $\sup_{\mathbf{x}\in\mathcal{N}} \|\nabla f_t(\mathbf{x}) - \nabla f_{t-1}(\mathbf{x})\|^2$ is a constant over time, and we can lower bounded it by

$$\begin{aligned}
& \sup_{\mathbf{x}\in\mathcal{N}} \|\nabla f_t(\mathbf{x}) - \nabla f_{t-1}(\mathbf{x})\|^2 \\
=& \sup_{\mathbf{x}\in\mathcal{N}} \left\| \frac{1}{n} \left(\sum_{i=1}^{n} \operatorname{Exp}_{\mathbf{x}}^{-1} \mathbf{x}_{t,i} - \sum_{i=1}^{n} \operatorname{Exp}_{\mathbf{x}}^{-1} \mathbf{x}_{t-1,i} \right) \right\|^2 \\
=& \sup_{\mathbf{x}\in\mathcal{N}} \left\| \frac{2}{n} \sum_{i=1}^{n} \operatorname{Exp}_{\mathbf{x}}^{-1} \mathbf{y}_i \right\|^2 \geq \frac{4}{n^2} \left\| \sum_{i=1}^{n} \operatorname{Exp}_{\mathbf{0}}^{-1} \mathbf{y}_i \right\|^2 \\
=& \frac{4}{n^2} \left(\left\| \left(\sum_{i=1}^{n} \operatorname{Exp}_{\mathbf{0}}^{-1} \mathbf{y}_i \right)^{\|} \right\|^2 + \left\| \left(\sum_{i=1}^{n} \operatorname{Exp}_{\mathbf{0}}^{-1} \mathbf{y}_i \right)^{\perp} \right\|^2 \right) \\
\geq& \frac{4}{n^2} \left\| \left(\sum_{i=1}^{n} \operatorname{Exp}_{\mathbf{0}}^{-1} \mathbf{y}_i \right)^{\|} \right\|^2 \geq \frac{4}{n^2} \left\| n \cdot \operatorname{arctanh}\left(\frac{1}{2} - T^{-\alpha}\right) \mathbf{e}_1 \right\|_{\mathbf{0}}^2 \\
=& 16 \operatorname{arctanh}^2 \left(\frac{1}{2} - T^{-\alpha}\right) \\
\geq& 16 \left(\frac{\frac{1}{2} - T^{-\alpha}}{\frac{3}{2} - T^{-\alpha}} \right)^2 = \Omega(1).
\end{aligned} \tag{64}$$

where we use $\mathbf{a}^{\|}$ and $\mathbf{a}^{\perp}$ to denote components parallel and orthogonal to the direction of $\mathbf{e}_1$, respectively. The key observation is the lower bound attains when $\left(\sum_{i=1}^n \mathrm{Exp}_{\mathbf{0}}^{-1}\mathbf{y}_i\right)^{\perp}$ is zero, and each $\mathbf{y}_i$ has the smallest component along $\mathbf{e}_1$, i.e., $\mathbf{y}_i = \left(\frac{1}{2} - T^{-\alpha}\right)\mathbf{e}_1$. We also use $\operatorname{arctanh}(x) \geq \frac{x}{1+x}$. Thus we have $V_T = \Omega(T)$. Now we consider F_T. By Lemma 24, we know that $\mathbf{u}_t$ lies within the same geodesic ball as $\mathbf{x}_{t,i}$, $i \in [n]$. Thus

$$F_T = \sum_{t=1}^T f_t(\mathbf{u}_t) = \sum_{t=1}^T \sum_{i=1}^n d(\mathbf{u}_t, \mathbf{x}_{t,i})^2/n \leq T \cdot \left(\frac{2}{T^\alpha}\right)^2 = O(T^{1-2\alpha}).$$

We can see whenever $\alpha > 0$, $F_T = o(T)$ but $V_T = \Omega(T)$.

D.2. Proof of Theorem 8

By Lemma 4 we have

$$\sum_{t=1}^T f_t(\mathbf{x}_t) - f_t(\mathbf{x}_{t,i}) \leq \frac{2+\ln N}{\beta} + \beta \sum_{t=1}^T \|\boldsymbol{\ell}_t - \mathbf{m}_t\|_\infty^2 - \frac{1}{4\beta}\sum_{t=2}^T \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_1^2 \tag{65}$$

We bound $\sum_{t=1}^T \|\boldsymbol{\ell}_t - \mathbf{m}_t\|_\infty^2$ in terms of $\sum_{t=1}^T \|\boldsymbol{\ell}_t - \mathbf{m}_t^v\|_\infty^2$ and $\sum_{t=1}^T \|\boldsymbol{\ell}_t - \mathbf{m}_t^s\|_\infty^2$ as follows. By Assumptions 2 and 3, $\|\boldsymbol{\ell}_t - \mathbf{m}_t\|_2^2 \leq 4NG^2D^2$ and we can compute $d_t(\mathbf{m}) = \|\boldsymbol{\ell}_t - \mathbf{m}\|_2^2$ is $\frac{1}{8NG^2D^2}$-exp-concave. . We have

$$\begin{aligned}
&\sum_{t=1}^T \|\boldsymbol{\ell}_t - \mathbf{m}_t\|_\infty^2 \leq \sum_{t=1}^T \|\boldsymbol{\ell}_t - \mathbf{m}_t\|_2^2 \\
\leq& \min\left\{\sum_{t=1}^T \|\boldsymbol{\ell}_t - \mathbf{m}_t^v\|_2^2, \sum_{t=1}^T \|\boldsymbol{\ell}_t - \mathbf{m}_t^s\|_2^2\right\} + 8NG^2D^2\ln 2 \qquad (66)\\
\leq& N\min\left\{\sum_{t=1}^T \|\boldsymbol{\ell}_t - \mathbf{m}_t^v\|_\infty^2, \sum_{t=1}^T \|\boldsymbol{\ell}_t - \mathbf{m}_t^s\|_\infty^2\right\} + 8NG^2D^2\ln 2,
\end{aligned}$$

where for the second inequality we use Lemma 32 and for the first and the third one the norm inequality $\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \sqrt{N}\|\mathbf{x}\|_\infty$ is used.

Combining Equations (50), (65), (66) and Lemma 7, we have

$$\sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{x}_{t,i}) \leq \frac{2+\ln T}{\beta} + \beta N\left(D^2\min\{3V_T, \bar{F}_T\} + 8G^2D^2\ln 2\right)$$

holds for any $\beta \leq \frac{1}{\sqrt{12(D^4L^2+D^2G^2\zeta^2)}}$ and $i \in [N]$.

Suppose $\beta^\star = \sqrt{\frac{2+\ln N}{N(D^2\min\{3(V_T+G^2), \bar{F}_T\}+8G^2D^2\ln 2)}} \leq \frac{1}{\sqrt{12(D^4L^2+D^2G^2\zeta^2)}}$, then

$$\sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{x}_{t,i}) \leq 2\sqrt{(2+\ln N)N(D^2\min\{3(V_T+G^2), \bar{F}_T\} + 8G^2D^2\ln 2)}.$$

Otherwise

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_{t,i}) \leq 2(2+\ln N)\sqrt{12(D^4L^2 + D^2G^2\zeta^2)}.$$

In sum, we have

$$\begin{aligned}&\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_{t,i})\\ \leq& \max\left\{2\sqrt{(2+\ln N)N(D^2\min\{3(V_T+G^2), \bar{F}_T\} + 8G^2D^2\ln 2)}, 2(2+\ln N)\sqrt{12(D^4L^2+D^2G^2\zeta^2)}\right\}\\ =&O(\log T\cdot \min\{V_T, \bar{F}_T\}).\end{aligned} \tag{67}$$

D.3. Proof of Theorem 9

By Theorem 8, we know

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_{t,i}) = O\left(\sqrt{\ln T(\min\{V_T, \bar{F}_T\})}\right)$$

holds for any $i \in [N^v + N^s]$. WLOG, we assume k and k' to be indexes of the best experts for the gradient-variation bound and the small-loss bound, respectively. Then by Theorem 6,

$$\sum_{t=1}^{T} f_t(\mathbf{x}_{t,k}) - f_t(\mathbf{u}_t) \leq \frac{3}{2}\sqrt{2(D^2+2DP_T)\zeta V_T} + (D^2+2DP_T)\frac{1+(1+2\zeta\delta^2L^2)^{\frac{1}{2}}}{\delta} \tag{68}$$

while by Theorem 7,

$$\sum_{t=1}^{T} f_t(\mathbf{x}_{t,k'}) - f_t(\mathbf{u}_t) \leq \sqrt{2\left(16\zeta^2L^2(D^2+2DP_T)^2 + 18\zeta LF_T(D^2+2DP_T)\right)}. \tag{69}$$

Since the regret admits the following decompositions

$$\begin{aligned}&\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}_t)\\ =&\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_{t,k}) + \sum_{t=1}^{T} f_t(\mathbf{x}_{t,k}) - \sum_{t=1}^{T} f_t(\mathbf{u}_t)\\ =&\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_{t,k'}) + \sum_{t=1}^{T} f_t(\mathbf{x}_{t,k'}) - \sum_{t=1}^{T} f_t(\mathbf{u}_t)\\ &,\end{aligned} \tag{70}$$

we indeed have

$$\begin{aligned}
&\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}_t)\\
\le& O\left(\sqrt{\ln T(\min\{V_T, \bar{F}_T\})}\right) + \min\left\{\frac{3}{2}\sqrt{2(D^2+2DP_T)\zeta V_T} + (D^2+2DP_T)\frac{1+(1+2\zeta\delta^2L^2)^{\frac{1}{2}}}{\delta},\right.\\
&\left.\sqrt{2\left(16\zeta^2L^2(D^2+2DP_T)^2+18\zeta LF_T(D^2+2DP_T)\right)}\right\}\\
=& O\left(\sqrt{\ln T(\min\{V_T, \bar{F}_T\})}\right)\\
&+O\left(\min\left\{\sqrt{\zeta(1+P_T)((1+P_T)/\delta^2+V_T)}, \sqrt{\zeta(1+P_T)(\zeta(1+P_T)+F_T)}\right\}\right)
\end{aligned} \tag{71}$$

Note that $\bar{F}_T$ can be processed similarly as in Lemma 7 and Theorem 7 to get F_T. In sum, the regret is bounded by

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}_t) = O\left(\sqrt{\zeta(P_T(\zeta+1/\delta^2)+\min\{V_T, F_T\}+1)(1+P_T)+\ln T\min\{V_T, F_T\}}\right).$$

Appendix E. Technical Lemmas

We need the following technical lemmas.

Lemma 19 *(Bento et al., 2021, Theorem 2.1) Suppose $f : \mathcal{M} \to \mathbb{R}$ is geodesically convex and $\mathcal{M}$ is Hadamard. The geodesic mean $\bar{\mathbf{x}}_k$ w.r.t coefficients $a_1, \ldots, a_N$ ($\sum_{i=1}^N a_i = 1$, $a_i \ge 0$) is defined as:*

$$\begin{aligned}
\bar{\mathbf{x}}_1 &= \mathbf{x}_1\\
\bar{\mathbf{x}}_k &= \mathrm{Exp}_{\bar{\mathbf{x}}_{k-1}}\left(\frac{a_k}{\sum_{i=1}^k a_i}\mathrm{Exp}_{\bar{\mathbf{x}}_{k-1}}^{-1}\mathbf{x}_k\right), \quad k > 1.
\end{aligned} \tag{72}$$

Then we have

$$f(\bar{\mathbf{x}}_N) \le \sum_{i=1}^{N} a_i f(\mathbf{x}_i). \tag{73}$$

Proof

We use induction to show a stronger statement

$$f(\bar{\mathbf{x}}_k) \le \sum_{i=1}^{k} \frac{a_i}{\sum_{j=1}^k a_j} f(\mathbf{x}_i)$$

holds for $k = 1, \ldots, N$.

For $k = 1$, this is obviously true because $\bar{\mathbf{x}}_1 = \mathbf{x}_1$. Suppose

$$f(\bar{\mathbf{x}}_k) \le \sum_{i=1}^{k} \frac{a_i}{\sum_{j=1}^k a_j} f(\mathbf{x}_i)$$

holds for some k, then by geodesic convexity,

$$\begin{aligned} f(\bar{\mathbf{x}}_{k+1}) &\le \left(1 - \frac{a_{k+1}}{\sum_{j=1}^{k+1} a_j}\right) f(\bar{\mathbf{x}}_k) + \frac{a_{k+1}}{\sum_{j=1}^{k+1} a_j} f(\mathbf{x}_{k+1}) \\ &\le \sum_{i=1}^{k} \frac{a_i}{\sum_{j=1}^{k+1} a_j} f(\mathbf{x}_i) + \frac{a_{k+1}}{\sum_{j=1}^{k+1} a_j} f(\mathbf{x}_{k+1}) \\ &= \sum_{i=1}^{k+1} \frac{a_i f(\mathbf{x}_i)}{\sum_{j=1}^{k+1} a_j}. \end{aligned} \tag{74}$$

The first inequality is due to gsc-convexity: for the geodesic determined by $\gamma(0) = \bar{\mathbf{x}}_k$ and $\gamma(1) = \mathbf{x}_{k+1}$ we have $\bar{\mathbf{x}}_{k+1} = \gamma\left(\frac{a_{k+1}}{\sum_{i=1}^{k+1} a_i}\right)$ and thus $f(\gamma(t)) \le (1-t) f(\gamma(0)) + t f(\gamma(1))$. For the second inequality, we use the induction hypothesis. Given $\sum_{i=1}^{N} a_i = 1$, the lemma is proved. ■

The computation of the geodesic averaging is summarized in Algorithm 6, which serves as a sub-routine of RADAR.

Algorithm 6: Geodesic Averaging

```
Data: $N$ points $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathcal{N}$ and $N$ real weights $w_1, \dots, w_N$.
Let $\bar{\mathbf{x}}_1 = \mathbf{x}_1$
for $k = 2, \dots, N$ do
    $\bar{\mathbf{x}}_k = \mathrm{Exp}_{\bar{\mathbf{x}}_{k-1}}\left(\frac{w_k}{\sum_{i=1}^{k} w_i} \mathrm{Exp}_{\bar{\mathbf{x}}_{k-1}}^{-1} \mathbf{x}_k\right)$
end
Return $\bar{\mathbf{x}}_N$.
```

Lemma 20 *Suppose $\mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{y}_1, \dots, \mathbf{y}_N \in \mathcal{N}$ where $\mathcal{N}$ is a gsc-convex subset of a Hadamard manifold $\mathcal{M}$ and the diameter of $\mathcal{N}$ is upper bounded by D. Let $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ be the weighted Fréchet mean with respect to coefficient vectors* $\mathbf{a}$ *and* $\mathbf{b}$ *($\mathbf{a}, \mathbf{b} \in \Delta_N$), defined as $\bar{\mathbf{x}} = \mathrm{argmin}_{\mathbf{x} \in \mathcal{N}} \sum_{i=1}^{N} a_i \cdot d(\mathbf{x}, \mathbf{x}_i)^2$ and $\bar{\mathbf{y}} = \mathrm{argmin}_{\mathbf{y} \in \mathcal{N}} \sum_{i=1}^{N} b_i \cdot d(\mathbf{y}, \mathbf{y}_i)^2$. Then we have*

$$d(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \le \sum_{i=1}^{N} a_i \cdot d(\mathbf{x}_i, \mathbf{y}_i) + D \sum_{i=1}^{N} |a_i - b_i|. \tag{75}$$

Proof Recall that on Hadamard manifolds, the following inequality (Sturm, 2003, Prop. 2.4)

$$d(\mathbf{x}, \mathbf{y})^2 + d(\mathbf{u}, \mathbf{v})^2 \le d(\mathbf{x}, \mathbf{v})^2 + d(\mathbf{y}, \mathbf{u})^2 + 2d(\mathbf{x}, \mathbf{u}) \cdot d(\mathbf{y}, \mathbf{v})$$

holds for any $\mathbf{x}, \mathbf{y}, \mathbf{u}, \mathbf{v} \in \mathcal{M}$. A direct application of the above inequality yields

$$d(\mathbf{x}_i, \bar{\mathbf{y}})^2 + d(\mathbf{y}_i, \bar{\mathbf{x}})^2 \le d(\mathbf{x}_i, \bar{\mathbf{x}})^2 + d(\mathbf{y}_i, \bar{\mathbf{y}})^2 + 2d(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \cdot d(\mathbf{x}_i, \mathbf{y}_i) \qquad \forall i \in [N]. \tag{76}$$

By (Bacák, 2014a, Theorem 2.4):

$$\sum_{i=1}^{N} a_i \cdot d(\mathbf{x}_i, \bar{\mathbf{x}})^2 + \sum_{i=1}^{N} b_i \cdot d(\mathbf{y}_i, \bar{\mathbf{y}})^2 + 2d(\bar{\mathbf{x}}, \bar{\mathbf{y}})^2 \le \sum_{i=1}^{N} a_i \cdot d(\mathbf{x}_i, \bar{\mathbf{y}})^2 + \sum_{i=1}^{N} b_i \cdot d(\mathbf{y}_i, \bar{\mathbf{x}})^2 \tag{77}$$

Multiplying Equation (76) by a_i, summing from $i = 1$ to n and adding Equation (77), we have

$$\begin{aligned} 2d(\bar{\mathbf{x}}, \bar{\mathbf{y}})^2 \leq & 2d(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \sum_{i=1}^{N} a_i \cdot d(\mathbf{x}_i, \mathbf{y}_i) + \sum_{i=1}^{N} (a_i - b_i) d(\mathbf{y}_i, \bar{\mathbf{y}})^2 + \sum_{i=1}^{N} (b_i - a_i) d(\mathbf{y}_i, \bar{\mathbf{x}})^2 \\ = & 2d(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \sum_{i=1}^{N} a_i \cdot d(\mathbf{x}_i, \mathbf{y}_i) + \sum_{i=1}^{N} (a_i - b_i) \cdot (d(\mathbf{y}_i, \bar{\mathbf{y}}) - d(\mathbf{y}_i, \bar{\mathbf{x}})) \cdot (d(\mathbf{y}_i, \bar{\mathbf{y}}) + d(\mathbf{y}_i, \bar{\mathbf{x}})) \\ \leq & 2d(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \sum_{i=1}^{N} a_i \cdot d(\mathbf{x}_i, \mathbf{y}_i) + 2D \sum_{i=1}^{N} |a_i - b_i| d(\bar{\mathbf{x}}, \bar{\mathbf{y}}), \end{aligned} \tag{78}$$

where for the last inequality we use the triangle inequality for geodesic metric spaces and Assumption 2. Now dividing both sides by $2d(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ and we complete the proof. ■

Lemma 21 *(Zhang and Sra, 2016, Lemma 5). Let $\mathcal{M}$ be a Riemannian manifold with sectional curvature lower bounded by $\kappa \leq 0$. Consider $\mathcal{N}$, a gsc-convex subset of $\mathcal{M}$ with diameter D. For a geodesic triangle fully lies within $\mathcal{N}$ with side lengths a, b, c, we have*

$$a^2 \leq \zeta(\kappa, D) b^2 + c^2 - 2bc \cos A$$

where $\zeta(\kappa, D) := \sqrt{-\kappa} D \coth(\sqrt{-\kappa} D)$.

Lemma 22 *(Sakai, 1996, Prop. 4.5) Let $\mathcal{M}$ be a Riemannian manifold with sectional curvature upper bounded by $\kappa \leq 0$. Consider $\mathcal{N}$, a gsc-convex subset of $\mathcal{M}$ with diameter D. For a geodesic triangle fully lies within $\mathcal{N}$ with side lengths a, b, c, we have*

$$a^2 \geq b^2 + c^2 - 2bc \cos A.$$

Lemma 23 *(Bacák, 2014b, Theorem 2.1.12) Let $(\mathcal{H}, d)$ be a Hadamard space and $C \subset \mathcal{H}$ be a closed convex set. Then $\Pi_C \mathbf{x}$ is singleton and $d(\mathbf{x}, \Pi_C \mathbf{x}) \leq d(\mathbf{x}, \mathbf{z})$ for any $\mathbf{z} \in C \setminus \{\Pi_C \mathbf{x}\}$.*

Lemma 24 *(Sturm, 2003, Prop. 6.1 and Theorem 6.2) Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathcal{N}$ where $\mathcal{N}$ is a bounded gsc-convex subset of a Hadamard space. $\bar{\mathbf{x}}$ is the weighted Fréchet mean of $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathcal{N}$ w.r.t. non-negative $w_1, \ldots, w_N$ such that $\sum_{i=1} w_i = 1$ and $f : \mathcal{N} \to R$ is a gsc-convex function. Then $\bar{\mathbf{x}} \in \mathcal{N}$ and*

$$f(\bar{\mathbf{x}}) \leq \sum_{i=1}^{N} w_i f(\mathbf{x}_i).$$

Lemma 25 *Let*

$$g(x) := \frac{-a + (a^2 + bx^2)^{\frac{1}{2}}}{x},$$

where $a, b \in \mathbb{R}^+$, then $g(x)$ increases on $[0, \infty)$.

Proof Taking the derivative w.r.t. x, we have

$$\begin{aligned}g'(x) =&\frac{\frac{1}{2}\cdot 2bx(a^2+bx^2)^{-\frac{1}{2}}\cdot x-(-a+(a^2+bx^2)^{\frac{1}{2}})\cdot 1}{x^2}\\ =&\frac{bx^2-(-a\sqrt{a^2+bx^2}+a^2+bx^2)}{x^2\sqrt{a^2+bx^2}}\\ =&\frac{a\sqrt{a^2+bx^2}-a^2}{x^2\sqrt{a^2+bx^2}}\geq 0\end{aligned}$$

holds for any $x>0$. By L'Hôpital's rule, $g(0)=0$ and $g'(0)=\frac{b}{2a}$. Thus we know $g(x)$ increases on $[0,\infty)$. ■

Lemma 26 *(Zhou and Huang, 2019, Theorem 3.1) On a Hadamard manifold $\mathcal{M}$, a subset C gsc-convex, iff it contains the geodesic convex combinations of any countable points in C.*

Lemma 27 *(Ahn and Sra, 2020, Prop. H.1) Let $\mathcal{M}$ be a Riemannian manifold with sectional curvatures lower bounded by $-\kappa<0$ and the distance function $d(\mathbf{x})=\frac{1}{2}d(\mathbf{x},\mathbf{p})^2$ where $\mathbf{p}\in\mathcal{M}$. For $D\geq 0$, $d(\cdot)$ is gsc-smooth within the domain $\{\mathbf{u}\in\mathcal{M}:d(\mathbf{u},\mathbf{p})\leq D\}$.*

Lemma 28 *(Ballmann, 2012, Corollary 5.6) Let $\mathcal{M}$ be a Hadamard space and $C\subset\mathcal{M}$ a convex subset. Then $d(\mathbf{z},C)$ is gsc-convex for $\mathbf{z}\in\mathcal{M}$.*

Lemma 29 *(Bacák, 2014b, Section 2.1) Let $\mathcal{H}$ be a Hadamard manifold, $f:\mathcal{H}\to(-\infty,\infty)$ be a convex lower semicontinuous function. Then any β-sublevel set of f:*

$$\{\mathbf{x}\in\mathcal{H}:f(\mathbf{x})\leq\beta\}$$

is a closed convex set.

Lemma 30 *(Sun et al., 2019, Lemma 4) Let the sectional curvature of $\mathcal{M}$ is in $[-K,K]$ and $\mathbf{x},\mathbf{y},\mathbf{z}\in\mathcal{M}$ with pairwise distance upper bounded by R. Then*

$$\|\mathrm{Exp}_{\mathbf{x}}^{-1}\mathbf{y}-\mathrm{Exp}_{\mathbf{x}}^{-1}\mathbf{z}\|\leq(1+c(K)R^2)d(\mathbf{y},\mathbf{z}).$$

Lemma 31 *(Duchi et al., 2012, Lemma 2) Let*

$$\Pi_{\mathcal{X}}^{\psi}(\mathbf{z},\alpha)=\operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}}\langle\mathbf{z},\mathbf{x}\rangle+\frac{\psi(\mathbf{x})}{\alpha}$$

where $\psi(\cdot)$ is 1-strongly convex w.r.t. $\|\cdot\|$, then

$$\|\Pi_{\mathcal{X}}^{\psi}(\mathbf{u},\alpha)-\Pi_{\mathcal{X}}^{\psi}(\mathbf{v},\alpha)\|\leq\alpha\|\mathbf{u}-\mathbf{v}\|_{\star}.$$

Lemma 32 *(Cesa-Bianchi and Lugosi, 2006, Prop. 3.1 and Theorem 3.2) Suppose the loss function ℓ_t is exp-concave for $\eta>0$, then the regret of Hedge is $\frac{\ln N}{\eta}$, where N is the number of experts.*