We divide the supplementary material in three sections. Appendix A contains the proofs related to the accelerated algorithm, i.e. the proofs of Theorems 2.4 and 2.5. In Appendix B we prove the results related to the reductions in Section 3. Finally, in Appendix C, we prove the geometric lemmas that take into account the geodesic map h to obtain relationships between F and f, namely Lemmas 2.1, 2.2 and 2.3.

A Acceleration. Proofs of Theorem 2.4 and Theorem 2.5

Diakonikolas & Orecchia (2019) developed the *approximate duality gap technique* which is a technique that provides a structure to design and prove first order methods and their guarantees for the optimization of convex problems. We take inspiration from this ideas to apply them to the non-convex problem we have at hand Theorem 2.4, as it was sketched in Section 2.1. We start with two basic definitions.

Definition A.1. Given two points $\tilde{x}, \tilde{y}$, we define the Bregman divergence with respect to $\psi(\cdot)$ as

$$D_\psi(\tilde{x}, \tilde{y}) \stackrel{\text{def}}{=} \psi(\tilde{x}) - \psi(\tilde{y}) - \langle \nabla\psi(\tilde{y}), \tilde{x} - \tilde{y}\rangle.$$

Definition A.2. Given a closed convex set Q and a function $\psi : Q \to \mathbb{R}$, we define the convex conjugate of ψ, also known as its Fenchel dual, as the function

$$\psi^*(\tilde{z}) = \max_{\tilde{x}\in Q}\{\langle \tilde{z}, \tilde{x}\rangle - \psi(\tilde{x})\}.$$

For simplicity, we will use $\psi(\tilde{x}) = \frac{1}{2}\|\tilde{x}\|^2$ in Algorithm 1, but any strongly convex map works. The gradient of the Fenchel dual of $\psi(\cdot)$ is $\nabla\psi^*(\tilde{z}) = \arg\min_{\tilde{z}'\in\mathcal{X}}\{\|\tilde{z}' - \tilde{z}\|\}$, that is, the Euclidean projection $\Pi_Q(\tilde{z})$ of the point $\tilde{z}$ onto Q. Note that when we apply Theorem 2.4 to Theorem 2.5 our constraint $\mathcal{X}$ will be a ball centered at 0 of radius $\tilde{R}$, so the projection of a point $\tilde{z}$ outside of $\mathcal{X}$ will be the vector normalization $\tilde{R}\tilde{z}/\|\tilde{z}\|$. Any continuously differentiable strongly convex ψ would work, provided that $\psi^*(z)$ is easily computable, preferably in closed form. Note that by the Fenchel-Moreau theorem we have for any such map that $\psi^{**} = \psi$.

We recall we assume that f satisfies

$$\begin{aligned} f(\tilde{x}) + \frac{1}{\gamma_{\mathrm{n}}}\langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x}\rangle \le f(\tilde{y}) \qquad &\text{if } \langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x}\rangle \le 0, \\ f(\tilde{x}) + \gamma_{\mathrm{p}}\langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x}\rangle \le f(\tilde{y}) \qquad &\text{if } \langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x}\rangle \ge 0. \end{aligned} \tag{8}$$

Let α_t be an increasing function of time t. We want to work with continuous and discrete approaches in a unified way so we use Lebesgue-Stieltjes integration. Thus, when α_t is a discrete measure, we have that $\alpha_t = \sum_{i=1}^{\infty} a_i \delta(t - (t_0 + i - 1))$ is a weighted sum of Dirac delta functions. We define $A_t = \int_{t_0}^{t} d\alpha_\tau = \int_{t_0}^{t} \dot{\alpha}_\tau d\tau$. In discrete time, it is $A_t = \sum_{i=1}^{t-t_0+1} a_i$. In the continuous case note that we have $\alpha_t - A_t = a_{t_0}$.

We start defining a continuous method that we discretize with an approximate implementation of the implicit Euler method. Let $\tilde{x}_t$ be the solution obtained by the algorithm at time t. We define the duality gap $G_t \stackrel{\text{def}}{=} U_t - L_t$ as the difference between a differentiable upper bound U_t on the function at the current point and a lower bound on $f(x^*)$. Since in our case f is differentiable we use $U_t \stackrel{\text{def}}{=} f(\tilde{x}_t)$. The idea is to enforce the invariant $\frac{d}{dt}(\alpha_t G_t) = 0$, so we have at any time $f(\tilde{x}_t) - f(\tilde{x}^*) \le G_t = G_{t_0}\alpha_{t_0}/\alpha_t$.

Note that for a global minimum $\tilde{x}^*$ of f and any other point $\tilde{x} \in Q$, we have $\langle \nabla f(\tilde{x}), \tilde{x}^* - \tilde{x}\rangle \le 0$. Otherwise, we would obtain a contradiction since by (8) we would have

$$f(\tilde{x}) < f(\tilde{x}) + \gamma_{\mathrm{p}}\langle \nabla f(\tilde{x}), \tilde{x}^* - \tilde{x}\rangle \le f(\tilde{x}^*).$$

Therefore, in order to define an appropriate lower bound, we will make use of the inequality $f(\tilde{x}^*) \ge f(\tilde{x}) + \frac{1}{\gamma_{\mathrm{n}}}\langle \nabla f(\tilde{x}), \tilde{x}^* - \tilde{x}\rangle$, for any $\tilde{x} \in Q$, which holds true by (8), for $\tilde{y} = \tilde{x}^*$. Combining this inequality for all the points visited by the continuous method we have

$$f(\tilde{x}^*) \ge \frac{\int_{t_0}^{t} f(\tilde{x}_\tau) d\alpha_\tau}{A_t} + \frac{\int_{t_0}^{t} \frac{1}{\gamma_{\mathrm{n}}}\langle \nabla f(\tilde{x}_\tau), \tilde{x}^* - \tilde{x}_\tau\rangle d\alpha_\tau}{A_t}.$$

We cannot compute this lower bound, since the right hand side depends on the unknown point $\tilde{x}^*$. We could compute a looser lower bound by taking the minimum over $\tilde{u} \in Q$ of this expression, substituting $\tilde{x}^*$ by $\tilde{u}$. However, this would make the lower bound non-differentiable and we could have problems at t_0. In order to solve the first problem, we first add a regularizer and then take the minimum over $\tilde{u} \in Q$.

$$f(\tilde{x}^*)+\frac{D_\psi(\tilde{x}^*,\tilde{x}_{t_0})}{A_t} \geq \frac{\int_{t_0}^t f(\tilde{x}_\tau)d\alpha_\tau}{A_t} + \frac{\min_{\tilde{u}\in Q}\left\{\int_{t_0}^t \frac{1}{\gamma_\mathrm{n}}\langle\nabla f(\tilde{x}_\tau),\tilde{u}-\tilde{x}_\tau\rangle d\alpha_\tau + D_\psi(\tilde{u},\tilde{x}_{t_0})\right\}}{A_t}$$

In order to solve the second problem, we mix this lower bound with the optimal lower bound $f(\tilde{x}^*)$ with weight $\alpha_t - A_t$ (this is only necessary in continuous time, in discrete time this term is 0). Not knowing $f(\tilde{x}^*)$ or $D_\psi(\tilde{x}^*,\tilde{x}_{t_0})$ will not be problematic. Indeed, we only need to guarantee $\frac{d}{dt}(\alpha_t G_t) = 0$, so after taking the derivative these terms will vanish. After rescaling the normalization factor, we finally obtain the lower bound

$$\begin{aligned} f(\tilde{x}^*) \geq L_t \stackrel{\text{def}}{=} \frac{\int_{t_0}^t f(\tilde{x}_\tau)d\alpha_\tau}{\alpha_t} + \frac{\min_{\tilde{u}\in Q}\left\{\int_{t_0}^t \langle\frac{1}{\gamma_\mathrm{n}}\nabla f(\tilde{x}_\tau),\tilde{u}-\tilde{x}_\tau\rangle d\alpha_\tau + D_\psi(\tilde{u},\tilde{x}_{t_0})\right\}}{\alpha_t} \\ + \frac{(\alpha_t - A_t)f(\tilde{x}^*) - D_\psi(\tilde{x}^*,\tilde{x}_{t_0})}{\alpha_t}. \end{aligned} \tag{9}$$

Let $\tilde{z}_t = \nabla\psi(\tilde{x}_{t_0}) - \int_{t_0}^t \frac{1}{\gamma_\mathrm{n}}\nabla f(\tilde{x}_\tau)d\alpha_\tau$. Then, by Fact A.7, we can compute the optimum $\tilde{u}$ above as

$$\nabla\psi^*(\tilde{z}_t) = \operatorname*{arg\,min}_{\tilde{u}\in Q}\left\{\int_{t_0}^t \langle\frac{1}{\gamma_\mathrm{n}}\nabla f(\tilde{x}_\tau),\tilde{u}-\tilde{x}_\tau\rangle d\alpha_\tau + D_\psi(\tilde{u},\tilde{x}_{t_0})\right\}. \tag{10}$$

Recalling $U_t = f(\tilde{x}_t)$ and using (9) and Danskin's theorem in order to differentiate inside the min we obtain:

$$\begin{aligned} \frac{d}{dt}(\alpha_t G_t) &= \frac{d}{dt}(\alpha_t f(\tilde{x}_t)) - \dot{\alpha}_t f(\tilde{x}_t) - \dot{\alpha}_t\frac{1}{\gamma_\mathrm{n}}\langle\nabla f(\tilde{x}_t),\nabla\psi^*(\tilde{z}_t)-\tilde{x}_t\rangle \\ &= \frac{1}{\gamma_\mathrm{n}}\langle\nabla f(\tilde{x}_t),\gamma_\mathrm{n}\alpha_t\dot{\tilde{x}} - \dot{\alpha}_t(\nabla\psi^*(\tilde{z}_t)-\tilde{x}_t)\rangle. \end{aligned}$$

Thus, to satisfy the invariant $\frac{d}{dt}(\alpha_t G_t) = 0$, it is enough to set $\gamma_\mathrm{n}\alpha_t\dot{\tilde{x}}_t = \dot{\alpha}_t(\nabla\psi^*(\tilde{z}_t)-\tilde{x}_t)$, yielding the following continuous accelerated dynamics

$$\begin{aligned} \dot{\tilde{z}}_t &= -\frac{1}{\gamma_\mathrm{n}}\dot{\alpha}_t\nabla f(\tilde{x}_t), \\ \dot{\tilde{x}}_t &= \frac{1}{\gamma_\mathrm{n}}\dot{\alpha}_t\frac{\nabla\psi^*(\tilde{z}_t)-\tilde{x}_t}{\alpha_t}, \\ \tilde{z}_{(t_0)} &= \nabla\psi(\tilde{x}_{t_0}), \\ \tilde{x}_{t_0} &\in Q \text{ is an arbitrary initial point.} \end{aligned} \tag{11}$$

Now we proceed to discretize the dynamics. Let $E_{i+1} \stackrel{\text{def}}{=} A_{i+1}G_{i+1} - A_i G_i$ be the discretization error. Then we have

$$G_k = \frac{A_1}{A_k}G_1 + \frac{\sum_{i=1}^{k-1} E_{i+1}}{A_k}.$$

Lemma A.3. *If we have*

$$f(\tilde{x}_{i+1}) - f(\tilde{x}_i) \leq \hat{\gamma}_i\langle\nabla f(\tilde{x}_{i+1}),\tilde{x}_{i+1}-\tilde{x}_i\rangle + \hat{\varepsilon}_i, \tag{12}$$

for some $\hat{\gamma}_i, \hat{\varepsilon}_i \geq 0$, then the discretization error satisfies

$$E_{i+1} \leq \langle\nabla f(\tilde{x}_{i+1}),(A_i\hat{\gamma}_i + \frac{a_{i+1}}{\gamma_\mathrm{n}})\tilde{x}_{i+1} - \hat{\gamma}_i A_i\tilde{x}_i - \frac{a_{i+1}}{\gamma_\mathrm{n}}\nabla\psi^*(\tilde{z}_{i+1}))\rangle - D_{\psi^*}(\tilde{z}_i,\tilde{z}_{i+1}) + A_i\hat{\varepsilon}_i.$$

Proof. In a similar way to Diakonikolas & Orecchia (2018), we could compute the discretization error as the difference between the gap and the gap computed allowing continuous integration rules in the integrals that it contains. However, we will directly bound E_{i+1} as $A_{i+1}G_{i+1} - A_iG_i$ instead. Using the definition of G_i, U_i, L_i we have

$$
\begin{aligned}
&A_{i+1}G_{i+1} - A_iG_i \\
&\quad \leq (A_{i+1}f(\tilde{x}_{i+1}) - A_if(\tilde{x}_i)) - A_{i+1}L_{i+1} + A_iL_i \\
&\quad \overset{①}{\leq} (A_if(\tilde{x}_{i+1}) - A_if(\tilde{x}_i) + a_{i+1}f(\tilde{x}_{i+1})) \\
&\qquad - \sum_{j=1}^{i+1} a_jf(\tilde{x}_j) - \sum_{j=1}^{i+1} \frac{a_j}{\gamma_{\mathrm{n}}}\langle \nabla f(\tilde{x}_j), \nabla\psi^*(\tilde{z}_{i+1}) - \tilde{x}_j\rangle - D_\psi(\nabla\psi^*(\tilde{z}_{i+1}), \tilde{x}_0) \\
&\qquad + \sum_{j=1}^{i} a_jf(\tilde{x}_j) + \sum_{j=1}^{i} \frac{a_j}{\gamma_{\mathrm{n}}}\langle \nabla f(\tilde{x}_j), \nabla\psi^*(\tilde{z}_i) - \tilde{x}_j\rangle + D_\psi(\nabla\psi^*(\tilde{z}_i), \tilde{x}_0) \\
&\quad \overset{②}{\leq} A_i(f(\tilde{x}_{i+1}) - f(\tilde{x}_i)) - \langle \frac{a_{i+1}}{\gamma_{\mathrm{n}}}\nabla f(\tilde{x}_{i+1}), \nabla\psi^*(\tilde{z}_{i+1}) - \tilde{x}_{i+1}\rangle \\
&\qquad + \sum_{j=1}^{i} \langle \frac{a_j}{\gamma_{\mathrm{n}}}\nabla f(\tilde{x}_j), \nabla\psi^*(\tilde{z}_i) - \nabla\psi^*(\tilde{z}_{i+1})\rangle \\
&\qquad [-\langle \nabla\psi(\tilde{x}_0), \nabla\psi^*(\tilde{z}_i) - \nabla\psi^*(\tilde{z}_{i+1})\rangle + \psi(\nabla\psi^*(\tilde{z}_i)) - \psi(\nabla\psi^*(\tilde{z}_{i+1}))] \\
&\quad \overset{③}{\leq} A_i(f(\tilde{x}_{i+1}) - f(\tilde{x}_i)) - \langle \frac{a_{i+1}}{\gamma_{\mathrm{n}}}\nabla f(\tilde{x}_{i+1}), \nabla\psi^*(\tilde{z}_{i+1}) - \tilde{x}_{i+1}\rangle - D_{\psi^*}(\tilde{z}_i, \tilde{z}_{i+1}) \\
&\quad \overset{④}{\leq} \langle \nabla f(\tilde{x}_{i+1}), (A_i\hat{\gamma}_i + \frac{a_{i+1}}{\gamma_{\mathrm{n}}})\tilde{x}_{i+1} - \hat{\gamma}_iA_i\tilde{x}_i - \frac{a_{i+1}}{\gamma_{\mathrm{n}}}\nabla\psi^*(\tilde{z}_{i+1}))\rangle - D_{\psi^*}(\tilde{z}_i, \tilde{z}_{i+1}) + A_i\hat{\varepsilon}_i.
\end{aligned}
$$

In ① we write down the definitions of L_{i+1} and L_i and split the first summand so it is clear that in ② we cancel all the $a_jf(\tilde{x}_j)$. In ② we also cancel some terms involved in the inner products, we write the definitions of the Bregman divergences and cancel some terms. We recall $\tilde{z}_i = \nabla\psi(x_0) - \sum_{j=1}^{i}\frac{a_j}{\gamma_{\mathrm{n}}}\nabla f(x_j)$ so we use this fact for the second line of ② and the first summand of the third line to obtain, along with the last two summands, the term $D_\psi(\nabla\psi^*(\tilde{z}_{i+1}), \nabla\psi^*(\tilde{z}_i))$. We use Lemma A.8 to finally obtain ③. Inequality ④ uses (12). □

We show now how to cancel out the discretization error by an approximate implementation of implicit Euler discretization of (11). Note that we need to take into account the assumptions (8) instead of the usual convexity assumption. According to the previous lemma, we can set $\tilde{x}_{i+1}$ so that the right hand side of the inner product in E_{i+1} is 0. Assume for the moment, that the $\tilde{x}_{i+1}$ we are going to compute satisfies the assumption of the previous lemma for some $\hat{\gamma}_i \in [\gamma_{\mathrm{p}}, 1/\gamma_{\mathrm{n}}]$. Thus, the implicit equation that defines the ideal method we would like to have is

$$
\tilde{x}_{i+1} = \frac{\hat{\gamma}_iA_i}{A_i\hat{\gamma}_i + a_{i+1}/\gamma_{\mathrm{n}}}\tilde{x}_i + \frac{a_{i+1}/\gamma_{\mathrm{n}}}{A_i\hat{\gamma}_i + a_{i+1}/\gamma_{\mathrm{n}}}\nabla\psi^*(\tilde{z}_i - \frac{a_{i+1}}{\gamma_{\mathrm{n}}}\nabla f(\tilde{x}_{i+1})).
$$

Note that $\tilde{x}_{i+1}$ is a convex combination of the other two points so it stays in Q. Indeed, $x_0 \in Q$ and by (10) we have that $\nabla\psi^*(\tilde{z}_j) \in Q$ for all $j \geq 0$. However this method is implicit and possibly computationally expensive to implement. Nonetheless, two steps of a fixed point iteration procedure of this equation will be enough to have discretization error that is bounded by the $A_i\hat{\varepsilon}_i$: the last term of our bound. The error in E_{i+1} that the inner product incurs is compensated by the Bregman divergence term. In such a case, the equations of this method become

$$
\begin{cases}
\tilde{\chi}_i = \frac{\hat{\gamma}_iA_i}{A_i\hat{\gamma}_i + a_{i+1}/\gamma_{\mathrm{n}}}\tilde{x}_i + \frac{a_{i+1}/\gamma_{\mathrm{n}}}{A_i\hat{\gamma}_i + a_{i+1}/\gamma_{\mathrm{n}}}\nabla\psi^*(\tilde{z}_i) \\
\tilde{\zeta}_i = \tilde{z}_i - \frac{a_{i+1}}{\gamma_{\mathrm{n}}}\nabla f(\tilde{\chi}_i) \\
\tilde{x}_{i+1} = \frac{\hat{\gamma}_iA_i}{A_i\hat{\gamma}_i + a_{i+1}/\gamma_{\mathrm{n}}}\tilde{x}_i + \frac{a_{i+1}/\gamma_{\mathrm{n}}}{A_i\hat{\gamma}_i + a_{i+1}/\gamma_{\mathrm{n}}}\nabla\psi^*(\tilde{\zeta}_i) \\
\tilde{z}_{i+1} = \tilde{z}_i - \frac{a_{i+1}}{\gamma_{\mathrm{n}}}\nabla f(\tilde{x}_{i+1})
\end{cases}
\tag{13}
$$

We prove now that this indeed leads to an accelerated algorithm. After this, we will show that we can perform a binary search at each iteration, to ensure that even if we do not know $\tilde{x}_{i+1}$ a priori, we can compute a $\hat{\gamma}_i \in [\gamma_\mathrm{p}, 1/\gamma_\mathrm{n}]$ satisfying assumption (12). This will only add a log factor to the overall complexity.

Lemma A.4. *Consider the method given in* (13)*, starting from and arbitrary point $\tilde{x}_0 \in Q$ with $\tilde{z}_0 = \nabla\psi(\tilde{x}_0)$ and $A_0 = 0$. Assume we can compute $\hat{\gamma}_i$ such that $\tilde{x}_{i+1}$ satisfies* (12)*. Then, the error from Lemma A.3 is bounded by*

$$E_{i+1} \leq \frac{a_{i+1}}{\gamma_\mathrm{n}}\langle \nabla f(\tilde{x}_{i+1}) - \nabla f(\tilde{\chi}_i), \nabla\psi^*(\tilde{\zeta}_i) - \nabla\psi^*(\tilde{z}_{i+1})\rangle - D_{\psi^*}(\tilde{\zeta}_i, \tilde{z}_{i+1}) - D_{\psi^*}(\tilde{z}_i, \tilde{\zeta}_i) + A_i\hat{\varepsilon}_i.$$

Proof. Using Lemma A.3 and the third line of (13) we have

$$\begin{aligned} E_{i+1} - A_i\hat{\varepsilon}_i &\leq \frac{a_{i+1}}{\gamma_\mathrm{n}}\langle \nabla f(\tilde{x}_{i+1}), \nabla\psi^*(\tilde{\zeta}_i) - \nabla\psi^*(\tilde{z}_{i+1})\rangle - D_{\psi^*}(\tilde{z}_i, \tilde{z}_{i+1}) \\ &\leq \frac{a_{i+1}}{\gamma_\mathrm{n}}\langle \nabla f(\tilde{x}_{i+1}) - \nabla f(\tilde{\chi}_i) + \nabla f(\tilde{\chi}_i), \nabla\psi^*(\tilde{\zeta}_i) - \nabla\psi^*(\tilde{z}_{i+1})\rangle - D_{\psi^*}(\tilde{z}_i, \tilde{z}_{i+1}) \end{aligned}$$

By the definition of $\tilde{\zeta}_i$ we have $(a_{i+1}/\gamma_\mathrm{n})\nabla f(\tilde{\chi}_i) = \tilde{z}_i - \tilde{\zeta}_i$. Using this fact and the triangle inequality of Bregman divergences Lemma A.9, we obtain

$$\begin{aligned} \frac{a_{i+1}}{\gamma_\mathrm{n}}\langle \nabla f(\tilde{\chi}_i), \nabla\psi^*(\tilde{\zeta}_i) - \nabla\psi^*(\tilde{z}_{i+1})\rangle &= \langle \tilde{z}_i - \tilde{\zeta}_i, \nabla\psi^*(\tilde{\zeta}_i) - \nabla\psi^*(\tilde{z}_{i+1})\rangle \\ &= D_{\psi^*}(\tilde{z}_i, \tilde{z}_{i+1}) - D_{\psi^*}(\tilde{\zeta}_i, \tilde{z}_{i+1}) - D_{\psi^*}(\tilde{z}_i, \tilde{\zeta}_i). \end{aligned}$$

The lemma follows after combining these two equations. □

Theorem A.5. *Let Q be a convex set of diameter D. Let $f : Q \to \mathbb{R}$ be an $\tilde{L}$-smooth function satisfying* (8)*. Assume there is a point $\tilde{x}^* \in Q$ such that $\nabla f(x^*) = 0$. Let $\tilde{x}_i, \tilde{z}_i, \tilde{\chi}_i, \tilde{\zeta}_i$ be updated according to* (13)*, for $i \geq 0$ starting from an arbitrary initial point $\tilde{x}_0 \in Q$ with $\tilde{z}_0 = \nabla\psi(\tilde{x}_0)$ and $A_0 = 0$, assuming we can find $\hat{\gamma}_i$ satisfying* (12)*. Let $\psi : \mathcal{B} \to \mathbb{R}$ be σ-strongly convex. If $\tilde{L}a_{i+1}^2/\gamma_\mathrm{n}\sigma \leq a_{i+1} + A_i\gamma_\mathrm{n}\gamma_\mathrm{p}$, then for all $t \geq 1$ we have*

$$f(\tilde{x}_t) - f(\tilde{x}^*) \leq \frac{D_\psi(\tilde{x}^*, \tilde{\chi}_0)}{A_t} + \sum_{i=1}^{t-1}\frac{A_i\hat{\varepsilon}_i}{A_t}.$$

In particular, if $a_i = \frac{i}{2}\cdot\frac{\sigma}{\tilde{L}}\cdot\gamma_\mathrm{n}^2\gamma_\mathrm{p}$, $\psi(\tilde{x}) = \frac{\sigma}{2}\|\tilde{x}\|^2$, $\hat{\varepsilon}_i = \frac{A_t\varepsilon}{2(t-1)A_i}$ and $t = \sqrt{\frac{2\tilde{L}\|\tilde{x}_0 - \tilde{x}^*\|^2}{\gamma_\mathrm{n}^2\gamma_\mathrm{p}\varepsilon}} = O(\sqrt{\tilde{L}/(\gamma_\mathrm{n}^2\gamma_\mathrm{p}\varepsilon)})$ then

$$f(\tilde{x}_t) - f(\tilde{x}^*) \leq \frac{2\tilde{L}\|\tilde{x}_0 - x^*\|^2}{\gamma_\mathrm{n}^2\gamma_\mathrm{p}t(t+1)} + \frac{\varepsilon}{2} < \varepsilon.$$

Proof. We bound the right hand side of the discretization error given by Lemma A.4. Define $a = \|\nabla\psi^*(\tilde{\zeta}_i) - \nabla\psi^*(\tilde{z}_{i+1})\|$ and $b = \|\nabla\psi^*(\tilde{\zeta}_i) - \nabla\psi^*(\tilde{z}_i)\|$. We have

$$\begin{aligned} E_{i+1} - A_i\hat{\varepsilon}_i &\overset{①}{\leq} \frac{a_{i+1}}{\gamma_\mathrm{n}}\langle \nabla f(\tilde{x}_{i+1}) - \nabla f(\tilde{\chi}_i), \nabla\psi^*(\tilde{\zeta}_i) - \nabla\psi^*(\tilde{z}_{i+1})\rangle - D_{\psi^*}(\tilde{\zeta}_i, \tilde{z}_{i+1}) - D_{\psi^*}(\tilde{z}_i, \tilde{\zeta}_i) \\ &\overset{②}{\leq} \frac{a_{i+1}}{\gamma_\mathrm{n}}\tilde{L}\|\tilde{x}_{i+1} - \tilde{\chi}_i\|\cdot a - D_{\psi^*}(\tilde{\zeta}_i, \tilde{z}_{i+1}) - D_{\psi^*}(\tilde{z}_i, \tilde{\zeta}_i) \\ &\overset{③}{\leq} \frac{a_{i+1}}{\gamma_\mathrm{n}}\tilde{L}\|\tilde{x}_{i+1} - \tilde{\chi}_i\|\cdot a - \frac{\sigma}{2}(a^2 + b^2) \\ &\overset{④}{\leq} \frac{a_{i+1}^2/\gamma_\mathrm{n}^2}{A_i\hat{\gamma}_i + a_{i+1}/\gamma_\mathrm{n}}\tilde{L}\cdot ab - \frac{\sigma}{2}(a^2 + b^2) \\ &\overset{⑤}{\leq} ab\left(\frac{a_{i+1}^2/\gamma_\mathrm{n}^2}{A_i\hat{\gamma}_i + a_{i+1}/\gamma_\mathrm{n}}\tilde{L} - \sigma\right). \end{aligned}$$

Here ① follows from Lemma A.4, ② uses the Cauchy-Schwartz inequality and smoothness, ③ uses Lemma A.10, and ④ uses the fact that by the definition of the method (13) we have $\tilde{x}_{i+1} - \tilde{\chi}_i = \frac{a_{i+1}/\gamma_n}{A_i\hat{\gamma}_i + a_{i+1}/\gamma_n}(\nabla\psi^*(\tilde{\zeta}_i) - \nabla\psi^*(\tilde{z}_i))$. Finally ⑤ uses $-(a^2+b^2) \leq -2ab$, which comes from $(a-b)^2 \geq 0$. By the previous inequality, if we want $E_{i+1} \leq A_i\hat{\varepsilon}_i$, it is enough to guarantee the right hand side of the last expression is ≤ 0 which is implied by

$$\frac{\tilde{L}}{\sigma\gamma_n}a_{i+1}^2 \leq a_{i+1} + A_i\gamma_n\gamma_p,$$

since $\gamma_p \leq \hat{\gamma}_i$. By inspection, if we use the value in the statement of the theorem $a_i = \frac{i}{2}\cdot\frac{\sigma}{\tilde{L}}\cdot\gamma_n^2\gamma_p$ into the previous inequality and noting that $A_i = \frac{i(i+1)}{4}\cdot\frac{\sigma}{\tilde{L}}\cdot\gamma_n^2\gamma_p$ we have

$$\begin{aligned}
\frac{\tilde{L}}{\sigma\gamma_n}a_{i+1}^2 &= \frac{(i+1)^2}{4}\cdot\frac{\sigma}{\tilde{L}}\cdot\gamma_n^3\gamma_p^2 \\
&\leq \left(\frac{i+1}{2} + \frac{i(i+1)}{4}\right)\frac{\sigma}{\tilde{L}}\cdot\gamma_n^3\gamma_p^2 \\
&\leq \frac{i+1}{2}\frac{\sigma}{\tilde{L}}\cdot\gamma_n^2\gamma_p + \frac{i(i+1)}{4}\frac{\sigma}{\tilde{L}}\cdot\gamma_n^3\gamma_p^2 \\
&= a_{i+1} + A_i\gamma_n\gamma_p
\end{aligned}$$

which holds true. So this choice guarantees discretization error $E_{i+1} \leq A_i\hat{\varepsilon}_i$. By the definition of G_i and E_i we have

$$f(\tilde{x}_t) - f(\tilde{x}^*) \leq \frac{A_1G_1}{G_t} + \sum_{i=1}^{t}\frac{A_{i-1}\hat{\varepsilon}_i}{A_t}$$

So it only remains to bound the initial gap G_1. In order to do this, we note that the initial conditions and the method imply the following computation of the first points, from $\tilde{x}_0 \in Q$, which is an arbitrary initial point:

$$\begin{cases}
\tilde{z}_0 = \nabla\psi(\tilde{x}_0) \\
\tilde{\chi}_0 = \frac{\hat{\gamma}_0 A_0}{A_0\hat{\gamma}_0 + a_1/\gamma_n}\tilde{x}_0 + \frac{a_1/\gamma_n}{A_0\hat{\gamma}_0 + a_1/\gamma_n}\nabla\psi^*(\tilde{z}_0) = \nabla\psi^*(\nabla\psi(\tilde{x}_0)) = \tilde{x}_0 \\
\tilde{\zeta}_0 = \tilde{z}_0 - \frac{a_1}{\gamma_n}\nabla f(\tilde{\chi}_0) = \tilde{z}_0 - \frac{a_1}{\gamma_n}\nabla f(\tilde{x}_0) \\
\tilde{x}_1 = \frac{\hat{\gamma}_0 A_0}{A_0\hat{\gamma}_0 + a_1/\gamma_n}\tilde{x}_0 + \frac{a_1/\gamma_n}{A_0\hat{\gamma}_0 + a_1/\gamma_n}\nabla\psi^*(\tilde{\zeta}_0) = \nabla\psi^*(\tilde{\zeta}_0)
\end{cases} \tag{14}$$

We have used $A_0 = 0$. Note this first iteration does not depend on $\hat{\gamma}_0$. Recall also that, using (9), the first lower bound computed is

$$L_1 = f(\tilde{x}_1) + \frac{1}{\gamma_n}\langle\nabla f(\tilde{x}_1), \nabla\psi^*(\tilde{z}_1) - \tilde{x}_1\rangle + \frac{1}{A_1}D_\psi(\nabla\psi^*(\tilde{z}_1), \tilde{\chi}_0) - \frac{1}{A_1}D_\psi(\tilde{x}^*, \tilde{\chi}_0).$$

Using $a_1 = A_1$, $\tilde{x}_1 = \nabla\psi^*(\tilde{\zeta}_0)$, $(a_1/\gamma_n)\nabla f(\tilde{\chi}_0) = \tilde{z}_0 - \tilde{\zeta}_0$, and the triangle inequality for Bregman divergences Lemma A.9 we obtain

$$\begin{aligned}
\frac{1}{\gamma_n}\langle\nabla f(\tilde{\chi}_0), \nabla\psi^*(\tilde{z}_1) - \tilde{x}_1\rangle &= \frac{1}{A_1}\langle\tilde{z}_0 - \tilde{\zeta}_0, \nabla\psi^*(\tilde{z}_1) - \nabla\psi^*(\tilde{\zeta}_0)\rangle \\
&= \frac{1}{A_1}\left(D_{\psi^*}(\tilde{z}_0, \tilde{\zeta}_0) - D_{\psi^*}(\tilde{z}_0, \tilde{z}_1) + D_{\psi^*}(\tilde{\zeta}_0, \tilde{z}_1)\right).
\end{aligned} \tag{15}$$

On the other hand, by smoothness of f and the initial condition we have

$$\frac{1}{\gamma_n}\langle\nabla f(\tilde{x}_1) - \nabla f(\tilde{\chi}_0), \nabla\psi^*(\tilde{z}_1) - \tilde{x}_1\rangle \geq -\frac{\tilde{L}}{\gamma_n}\|\nabla\psi^*(\tilde{\zeta}_0) - \tilde{\chi}_0\|\|\nabla\psi^*(\tilde{z}_1) - \tilde{x}_1\|. \tag{16}$$

We can now finally bound G_1:

$$\begin{aligned}
G_1 &\overset{①}{\leq} \frac{\tilde{L}}{\gamma_{\mathrm{n}}}\|\nabla\psi^*(\tilde{\zeta}_0)-\tilde{\chi}_0\|\cdot\|\nabla\psi^*(\tilde{z}_1)-\tilde{x}_1\| \\
&\quad - \frac{1}{A_1}\left(D_{\psi^*}(\tilde{z}_0,\tilde{\zeta}_0)+D_{\psi^*}(\tilde{\zeta}_0,\tilde{z}_1)\right)+\frac{1}{A_1}D_{\psi}(\tilde{x}^*,\tilde{\chi}_0) \\
&\overset{②}{\leq} \frac{\tilde{L}}{\gamma_{\mathrm{n}}}\|\nabla\psi^*(\tilde{\zeta}_0)-\tilde{\chi}_0\|\cdot\|\nabla\psi^*(\tilde{z}_1)-\tilde{x}_1\| \\
&\quad - \frac{\sigma}{2A_1}\left(\|\nabla\psi^*(\tilde{\zeta}_0)-\tilde{\chi}_0\|^2+\|\nabla\psi^*(\tilde{z}_1)-\tilde{x}_1\|^2\right)+\frac{1}{A_1}D_{\psi}(\tilde{x}^*,\tilde{\chi}_0) \\
&\overset{③}{\leq} \|\nabla\psi^*(\tilde{\zeta}_0)-\tilde{\chi}_0\|\cdot\|\nabla\psi^*(\tilde{z}_1)-\tilde{x}_1\|\left(\frac{\tilde{L}}{\gamma_{\mathrm{n}}}-\frac{\sigma}{A_1}\right)+\frac{1}{A_1}D_{\psi}(\tilde{x}^*,\tilde{\chi}_0) \\
&\overset{④}{\leq} \frac{1}{A_1}D_{\psi}(\tilde{x}^*,\tilde{\chi}_0).
\end{aligned}$$

We used in ① the definition of $G_1 = U_1 - L_1 = f(\tilde{x}_1) - L_1$ and we bound the inner product in L_1 using (15), and (16). Also, since $\tilde{z}_0 = \nabla\psi(\tilde{\chi}_0)$ we have $D_{\psi^*}(\tilde{z}_0,\tilde{z}_1) = D_{\psi}(\nabla\psi^*(\tilde{z}_1),\nabla\psi^*(\tilde{z}_0)) = D_{\psi}(\nabla\psi^*(\tilde{z}_1),\tilde{\chi}_0)$, so we can cancel two of the Bregman divergences. In ②, we used Lemma A.10, $\nabla\psi^*(\tilde{z}_0) = \tilde{\chi}_0$, and $\nabla\psi^*(\tilde{\zeta}_0) = \tilde{x}_1$. In ③ we used again the inequality $-(a^2+b^2) \leq -2ab$. Finally ④ is deduced from $A_1 = a_1 \leq \sigma\gamma_{\mathrm{n}}/\tilde{L}$ which comes from the assumption $\tilde{L}a_{i+1}^2/\gamma_{\mathrm{n}}\sigma \leq a_{i+1} + A_i\gamma_{\mathrm{n}}\gamma_{\mathrm{p}}$ for $i = 0$.

The first part of the theorem follows. The second one is a straightforward application of the first one as we see below. Indeed, taking into account $A_t = \frac{t(t+1)\sigma\gamma_{\mathrm{n}}^2\gamma_{\mathrm{p}}}{4\tilde{L}}$ and the choice of t we derive the second statement.

$$f(\tilde{x}_t) - f(\tilde{x}^*) \leq \frac{A_1G_1}{A_t} + \sum_{i=1}^{t-1}\frac{A_i\hat{\varepsilon}_i}{A_t} \leq \frac{\frac{\sigma}{2}\|\tilde{x}_0-\tilde{x}^*\|^2}{A_t} + \frac{\varepsilon}{2} < \varepsilon.$$

□

We present now the final lemma, that proves that $\hat{\gamma}_i$ can be found efficiently. As we advanced in the sketch of the main paper, we use a binary search. The idea behind it is that due to (8) we satisfy the equation for $\hat{\gamma}_i = \frac{1}{\gamma_{\mathrm{n}}}$ or $\hat{\gamma}_i = \gamma_{\mathrm{p}}$, or there is $\hat{\gamma}_i \in (\gamma_{\mathrm{p}}, 1/\gamma_{\mathrm{n}})$ such that $\langle\nabla f(\tilde{x}_{i+1}), \tilde{x}_{i+1} - \tilde{x}_i\rangle = 0$. The existence of $\tilde{x}^*$ that satisfies $\nabla f(\tilde{x}^*) = 0$ along with the boundedness of Q and smoothness, imply the Lipschitzness of f. Both Lipschitzness and smoothness allow to prove that a binary search finds efficiently a suitable point.

Lemma A.6. *Let $Q \subseteq \mathbb{R}^d$ be a convex set of diameter $2\tilde{R}$. Let $f : Q \to \mathbb{R}$ be a function that satisfies 8, is $\tilde{L}$ smooth and such that there is $\tilde{x}^* \in Q$ such that $\nabla f(x^*) = 0$. Let the strongly convex parameter of $\psi(\cdot)$ be $\sigma = O(1)$. Let $i \geq 1$ be an index. Given two points $\tilde{x}_i, \tilde{z}_i \in Q$ and the method in* (6) *using the learning rates $a_i = \frac{i}{2}\cdot\frac{\sigma}{\tilde{L}}\cdot\gamma_{\mathrm{n}}^2\gamma_{\mathrm{p}}$ prescribed in Theorem A.5, we can compute $\hat{\gamma}_i$ satisfying* (12)*, i.e.*

$$f(\tilde{x}_{i+1}) - f(\tilde{x}_i) \leq \hat{\gamma}_i\langle\nabla f(\tilde{x}_{i+1}), \tilde{x}_{i+1} - \tilde{x}_i\rangle + \hat{\varepsilon}_i. \tag{17}$$

And the computation of $\hat{\gamma}_i$ requires no more than

$$O\left(\log\left(\frac{\tilde{L}\tilde{R}}{\gamma_{\mathrm{n}}\hat{\varepsilon}_i}\cdot i\right)\right)$$

queries to the gradient oracle.

Proof. Let $\hat{\Gamma}_i(\lambda) : [\frac{a_{i+1}}{A_{i+1}}, \frac{a_{i+1}/\gamma_{\mathrm{n}}}{A_i\gamma_{\mathrm{p}}+a_{i+1}/\gamma_{\mathrm{n}}}] \to \mathbb{R}$ be defined as

$$\hat{\Gamma}_i\left(\frac{a_{i+1}/\gamma_{\mathrm{n}}}{A_i\mathfrak{x}+a_{i+1}/\gamma_{\mathrm{n}}}\right) = \mathfrak{x}, \text{ for } \mathfrak{x} \in [\gamma_{\mathrm{p}}, \frac{1}{\gamma_{\mathrm{n}}}]. \tag{18}$$

By monotonicity, that it is well defined. Let $\tilde{x}_{i+1}^{\lambda}$ be the point computed by one iteration of (6) using the parameter $\hat{\gamma}_i = \hat{\Gamma}_i(\lambda)$. Likewise, we define the rest of the points in the iteration (6) depending on λ. We first try $\hat{\gamma}_i = 1/\gamma_{\mathrm{n}}$ and $\hat{\gamma}_i = \gamma_{\mathrm{p}}$ and use any of them if they satisfy the conditions. If neither of them do, it means that for the first choice we had $\langle \nabla f(\tilde{x}_{i+1}^{\lambda_1}), \tilde{x}_{i+1}^{\lambda_1} - \tilde{x}_i \rangle < 0$ and for the second one, it is $\langle \nabla f(\tilde{x}_{i+1}^{\lambda_2}), \tilde{x}_{i+1}^{\lambda_2} - \tilde{x}_i \rangle > 0$, for $\lambda_1 = \hat{\Gamma}_i^{-1}(1/\gamma_{\mathrm{n}})$ and $\lambda_2 = \hat{\Gamma}_i^{-1}(\gamma_{\mathrm{p}})$. Therefore, by continuity, there is $\lambda^* \in [\lambda_1, \lambda_2]$ such that $\langle \nabla f(\tilde{x}_{i+1}^{\lambda^*}), \tilde{x}_{i+1}^{\lambda^*} - \tilde{x}_i \rangle = 0$. The continuity condition is easy to prove. We omit it because it is derived from the Lipschitzness condition that we will prove below. Such a point satisfies (8) for $\hat{\varepsilon}_i = 0$. We will prove that the function $G_i : [\frac{a_{i+1}}{A_{i+1}}, \frac{a_{i+1}/\gamma_{\mathrm{n}}}{A_i\gamma_{\mathrm{p}} + a_{i+1}/\gamma_{\mathrm{n}}}] \to \mathbb{R}$, defined as

$$G_i(\lambda) \stackrel{\text{def}}{=} -\hat{\Gamma}_i(\lambda)\langle \nabla f(\tilde{x}_{i+1}^{\lambda}), \tilde{x}_{i+1}^{\lambda} - \tilde{x}_i \rangle + (f(\tilde{x}_{i+1}^{\lambda}) - f(\tilde{x}_i)), \tag{19}$$

is Lipschitz so we can guarantee that (12) holds for an interval around λ^*. Finally, we will be able to perform a binary search to efficiently find a point in such interval or another interval around another point that satisfies that the inner product is 0.

So

$$\begin{aligned} |G_i(\lambda) - G_i(\lambda')| \leq\ & |f(\tilde{x}_{i+1}^{\lambda}) - f(\tilde{x}_{i+1}^{\lambda'})| \\ & + |\hat{\Gamma}_i(\lambda')| \cdot |\langle \nabla f(\tilde{x}_{i+1}^{\lambda'}), \tilde{x}_{i+1}^{\lambda'} - \tilde{x}_i \rangle - \langle \nabla f(\tilde{x}_{i+1}^{\lambda}), \tilde{x}_{i+1}^{\lambda} - \tilde{x}_i \rangle| \\ & + |\langle \nabla f(\tilde{x}_{i+1}^{\lambda}), \tilde{x}_{i+1}^{\lambda} - \tilde{x}_i \rangle| \cdot |\hat{\Gamma}_i(\lambda') - \hat{\Gamma}_i(\lambda)| \end{aligned} \tag{20}$$

We have used used the triangular inequality and the inequality

$$|\alpha_1\beta_1 - \alpha_2\beta_2| \leq |\alpha_1||\beta_1 - \beta_2| + |\beta_2||\alpha_1 - \alpha_2|, \tag{21}$$

which is a direct consequence of the triangular inequality, after adding and subtracting $\alpha_1\beta_2$ in the $|\cdot|$ on the left hand side. We bound each of the three summands of the previous inequality separately, but first we bound the following which will be useful for our other bounds,

$$\begin{aligned} \|\tilde{x}_{i+1}^{\lambda'} - \tilde{x}_{i+1}^{\lambda}\| &\overset{①}{=} \|(\lambda'\nabla\psi^*(\tilde{\zeta}_i^{\lambda'}) + (1-\lambda')\tilde{x}_i) - (\lambda\nabla\psi^*(\tilde{\zeta}_i^{\lambda}) + (1-\lambda)\tilde{x}_i)\| \\ &\overset{②}{\leq} \|\nabla\psi^*(\tilde{\zeta}_i^{\lambda}) - \tilde{x}_i\||\lambda' - \lambda| + \|\lambda'\nabla\psi^*(\tilde{\zeta}_i^{\lambda'}) - \lambda'\nabla\psi^*(\tilde{\zeta}_i^{\lambda})\| \\ &\overset{③}{\leq} 2\tilde{R}|\lambda - \lambda'| + \|\nabla\psi^*(\tilde{\zeta}_i^{\lambda'}) - \nabla\psi^*(\tilde{\zeta}_i^{\lambda})\| \\ &\overset{④}{\leq} 2\tilde{R}|\lambda - \lambda'| + \frac{1}{\gamma_{\mathrm{n}}\sigma}\|\nabla f(\tilde{\chi}_i^{\lambda}) - \nabla f(\tilde{\chi}_i^{\lambda'})\| \\ &\overset{⑤}{\leq} 2\tilde{R}|\lambda - \lambda'| + \frac{\tilde{L}}{\gamma_{\mathrm{n}}\sigma}\|\tilde{\chi}_i^{\lambda} - \tilde{\chi}_i^{\lambda'}\| \\ &\overset{⑥}{\leq} \left(2\tilde{R} + \frac{2L\tilde{R}}{\gamma_{\mathrm{n}}\sigma}\right)|\lambda - \lambda'| \end{aligned} \tag{22}$$

Here, ① uses the definition of $\tilde{x}_{i+1}^{\lambda}$ as a convex combination of $\tilde{x}_i$ and $\nabla\psi^*(\tilde{\zeta}_i^{\lambda})$. ② adds and substracts $\lambda'\nabla\psi^*(\tilde{\zeta}_i^{\lambda})$, groups terms and uses the triangular inequality. In ③ we use the fact that the diameter of Q is $2\tilde{R}$ and bound $\lambda' \leq 1$, and $|\lambda| \leq 1$. ④ uses the $\frac{1}{\sigma}$ smoothness of $\nabla\psi^*(\cdot)$, which is a consequence of the σ-strong convexity of $\psi(\cdot)$. ⑤ uses the smoothness of f. In ⑥, from the definition of $\tilde{\chi}_i^{\lambda}$ we have that $\|\tilde{\chi}_i^{\lambda} - \tilde{\chi}_i^{\lambda'}\| \leq \|\tilde{x}_i - \tilde{z}_i\||\lambda - \lambda'|$. We bounded this further using the diameter of Q.

Note that f is Lipschitz over Q. By the existence of x^*, $\tilde{L}$-smoothness, and the diameter of Q we obtain that the Lipschitz constant L_{p} is $L_{\mathrm{p}} \leq 2R^2L$. Now we can proceed and bound the three summands of (20). The first one reduces to the inequality above after using Lipschitzness of $f(\cdot)$:

$$|f(\tilde{x}_{i+1}^{\lambda}) - f(\tilde{x}_{i+1}^{\lambda'})| \leq L_{\mathrm{p}}\|\tilde{x}_{i+1}^{\lambda'} - \tilde{x}_{i+1}^{\lambda}\|. \tag{23}$$

In order to bound the second summand, we note that

$$|(\hat{\Gamma}_i^{-1})'(\tilde{\mathbf{x}})| = \left|\frac{A_i a_{i+1}/\gamma_{\mathrm{n}}}{(A_i\tilde{\mathbf{x}} + a_{i+1}/\gamma_{\mathrm{n}})^2}\right| \geq \frac{\gamma_{\mathrm{n}}A_i a_{i+1}}{A_{i+1}^2}, \tag{24}$$

so $\hat{\Gamma}_i(\lambda')$, appearing in the first factor, is bounded by $A_{i+1}^2/(\gamma_{\mathrm{n}} A_i a_{i+1})$. We used $\tilde{\mathfrak{x}} \in [\gamma_{\mathrm{p}}, 1/\gamma_{\mathrm{n}}]$ for the bound. For the second factor, we add and subtract $\langle \nabla f(\tilde{x}_{i+1}^{\lambda}), \tilde{x}_{i+1}^{\lambda'} - \tilde{x}_i \rangle$ and use the triangular inequality and then Cauchy-Schwartz. Thus, we obtain

$$\begin{aligned} &|\langle \nabla f(\tilde{x}_{i+1}^{\lambda'}), \tilde{x}_{i+1}^{\lambda'} - \tilde{x}_i \rangle - \langle \nabla f(\tilde{x}_{i+1}^{\lambda}), \tilde{x}_{i+1}^{\lambda} - \tilde{x}_i \rangle| \\ &\leq \|\nabla f(\tilde{x}_{i+1}^{\lambda})\| \cdot \|\tilde{x}_{i+1}^{\lambda'} - \tilde{x}_{i+1}^{\lambda}\| + \|\nabla f(\tilde{x}_{i+1}^{\lambda'}) - \nabla f(\tilde{x}_{i+1}^{\lambda})\| \cdot \|\tilde{x}_{i+1}^{\lambda'} - \tilde{x}_i\| \\ &\overset{①}{\leq} (2L_{\mathrm{p}} + 2\tilde{L}\tilde{R}) \|\tilde{x}_{i+1}^{\lambda'} - \tilde{x}_{i+1}^{\lambda}\|. \end{aligned} \tag{25}$$

In ①, we used Lipschitzness to bound the first factor. We also used the diameter of Q to bound the last factor and the smoothness of $f(\cdot)$ to bound the first factor of the second summand.

For the third summand, we will bound the first factor using Cauchy-Schwartz, smoothness of $f(\cdot)$ and the diameter of Q. We just proved in (24) that $\hat{\Gamma}_i$ is Lipschitz, so use use this property for the second factor. The result is the following

$$|\langle \nabla f(\tilde{x}_{i+1}^{\lambda}), \tilde{x}_{i+1}^{\lambda} - \tilde{x}_i \rangle| \cdot |\hat{\Gamma}_i(\lambda') - \hat{\Gamma}_i(\lambda)| \leq 4\tilde{L}\tilde{R}^2 \frac{A_{i+1}^2}{\gamma_{\mathrm{n}} A_i a_{i+1}} |\lambda' - \lambda|. \tag{26}$$

Applying the bounds of the three summands (23), (24), (25), (26) into (20) we obtain the inequality $|G_i(\lambda') - G_i(\lambda)| \leq \hat{L}|\lambda' - \lambda|$ for

$$\hat{L} = \left(2\tilde{R} + \frac{2\tilde{L}\tilde{R}}{\gamma_{\mathrm{n}}\sigma}\right)\left(L_{\mathrm{p}} + (2L_{\mathrm{p}} + 2\tilde{L}\tilde{R})\frac{A_{i+1}^2}{\gamma_{\mathrm{n}} A_i a_{i+1}}\right) + 4\tilde{L}\tilde{R}^2 \frac{A_{i+1}^2}{\gamma_{\mathrm{n}} A_i a_{i+1}}.$$

We will use the following to bound $\hat{L}$. If we use the learning rates prescribed in Theorem A.5, namely $a_i = \frac{i\sigma\gamma_{\mathrm{n}}^2\gamma_{\mathrm{p}}}{2L}$ and thus $A_i = \frac{i(i+1)\sigma\gamma_{\mathrm{n}}^2\gamma_{\mathrm{p}}}{4L}$ we can bound $A_{i+1}^2/(A_i a_{i+1}) \leq 3(i+2)$, using that $i \geq 1$. In our setting, by smoothness and the existence of $\tilde{x}^* \in Q$ such that $\nabla f(\tilde{x}^*) = 0$, we have that $L_{\mathrm{p}} \leq 2\tilde{R}\tilde{L}$. Recall we assume $\sigma = O(1)$. In Algorithm 1 we use $\sigma = 1$.

Recall we are denoting by λ^* a value such that $\langle \nabla f(\tilde{x}_{i+1}^{\lambda^*}), \tilde{x}_{i+1}^{\lambda^*} - \tilde{x}_i \rangle = 0$ so $G_i(\lambda^*) \leq 0$. Lipschitzness of G implies that if $G_i(\lambda^*) \leq 0$ then $G_i(\lambda) \leq \hat{\varepsilon}_i$ for $\lambda \in [\lambda^* - \frac{\hat{\varepsilon}_i}{\hat{L}}, \lambda^* + \frac{\hat{\varepsilon}_i}{\hat{L}}] \cap [\Gamma_i^{-1}(\gamma_{\mathrm{n}}), \Gamma_i^{-1}(\gamma_{\mathrm{p}})]$. If the extremal points, $\Gamma_i^{-1}(\gamma_{\mathrm{n}}), \Gamma_i^{-1}(\gamma_{\mathrm{p}})$ did not satisfy (17), then this interval is of length $\frac{2\hat{\varepsilon}_i}{\hat{L}}$ and a point in such interval or another interval that is around another point $\bar{\lambda}^*$ that satisfies $\langle \nabla f(\tilde{x}_{i+1}^{\bar{\lambda}^*}), \tilde{x}_{i+1}^{\bar{\lambda}^*} - \tilde{x}_i \rangle = 0$ can be found with a binary search in at most

$$O\left(\log\left(\frac{\hat{L}}{\hat{\varepsilon}_i}\right)\right) \overset{①}{=} O\left(\log\left(\frac{\tilde{L}\tilde{R}}{\gamma_{\mathrm{n}}\hat{\varepsilon}_i} \cdot i\right)\right)$$

iterations, provided that at each step we can ensure we halve the size of the search interval. The bounds of the previous paragraph are applied in ①.The binary search can be done easily: we start with $[\Gamma_i^{-1}(\gamma_{\mathrm{n}}), \Gamma_i^{-1}(\gamma_{\mathrm{p}})]$ and assume the extremes do not satisfy (17), so the sign of $\langle \nabla f(\tilde{x}_{i+1}^{\lambda}), \tilde{x}_{i+1}^{\lambda} - \tilde{x}_i \rangle$ is different for each extreme. Each iteration of the binary search queries the midpoint of the current working interval and if (17) is not satisfied, we keep the half of the interval such that the extremes keep having the sign of $\langle \nabla f(\tilde{x}_{i+1}^{\lambda}), \tilde{x}_{i+1}^{\lambda} - \tilde{x}_i \rangle$ different from each other, ensuring that there is a point in which this expression evaluates to 0 and thus keeping the invariant. We include the pseudocode of this binary search in Algorithm 2. □

We proceed to prove Theorem 2.4, which is an immediate consequence of the previous results.

Proof of Theorem 2.4. The proof follows from Theorem A.5, provided that we can find $\hat{\gamma}_i$ satisfying (12). Lemma A.6 shows that this is possible after performing a logarithmic number of queries to the gradient oracle. Note that given our choice of $\hat{\varepsilon}_i$, t and a_i, the number of queries to the gradient oracle Lemma A.6 requires is no more than $O(\log(\tilde{L}R/\gamma_{\mathrm{n}}\varepsilon))$ for any $i \leq t$. So we find an ε-minimizer of f after $\widetilde{O}(\sqrt{\tilde{L}/(\gamma^2\gamma_{\mathrm{p}}\varepsilon)})$ queries to the gradient oracle. □

Proof of Theorem 2.5. Given the function to optimize $F : \mathcal{M} \to \mathbb{R}$ and the geodesic map h, we define $f = F \circ h^{-1}$. Using Lemma 2.3 we know that f is $\tilde{L}$-smooth, with $\tilde{L} = O(L)$. Lemma 2.2 proves that f satisfies (8) for constants γ_{n} and γ_{p} depending on R. So Theorem 2.4 applies and the total number of queries to the oracle needed to obtain an ε-minimizer of f is $\widetilde{O}(\sqrt{\tilde{L}/\gamma_{\mathrm{n}}^2\gamma_{\mathrm{p}}\varepsilon}) = \widetilde{O}(\sqrt{L/\varepsilon})$. The result follows, since $f(\tilde{x}_t) - f(\tilde{x}^*) = F(x_t) - F(x^*)$. □

We recall a few concepts that were assumed during Section 2 to better interpret Theorem 2.5. We work in the hyperbolic space or an open hemisphere. The aim is to minimize a smooth and g-convex function defined on any of these manifolds, or a subset of them. The existence of a point x^* that satisfies $\nabla F(x^*) = 0$ is assumed. Starting from an arbitrary point x_0, we let R be a bound of the distance between x_0 and x^*, that is, $R \geq d(x_0, x^*)$. We let $\mathcal{M} = \mathrm{Exp}_x(\bar{B}(0, R))$ so that $x^* \in \mathcal{M}$. We assume $F : \mathcal{M}' \to \mathbb{R}$ is a differentiable function, where $\mathcal{M}' = \mathrm{Exp}_x B(0, R')$ and $R' > R$. We define F on $\mathcal{M}'$ only for simplicity, to avoid the use of subdifferentials. $\mathcal{M}$ has constant sectional curvature K. If K is positive, we restrict $R < \pi/2\sqrt{K}$ so $\mathcal{M}$ is contained in an open hemisphere and it is uniquely geodesic. We define a geodesic map h from the hyperbolic plane or a open hemisphere onto a subset of $\mathbb{R}^d$ and define the function $f : h(\mathcal{M}) \to \mathbb{R}$ as $f = F \circ h^{-1}$. We optimize this function in an accelerated way up to constants and log factors, where the constants appear as an effect of the deformation of the geometry and depend on R and K only. Note the assumption of the existence of x^* such that $\nabla F(x^*) = 0$ is not necessary, since $\arg\min_{x \in \mathrm{Exp}_{x_0}(\bar{B}(0,R))}\{F(x)\}$ also satisfies the first inequality in (8) so the lower bounds L_i can be defined in the same way as we did. In that case, if we want to perform constrained optimization, one needs to use the Lipschitz constant of F, when restricted to $\mathrm{Exp}_x(\bar{B}(0, R))$, for the analysis of the binary search.

Algorithm 2 BinaryLineSearch$(\tilde{x}_i, \tilde{z}_i, f, \mathcal{X}, a_{i+1}, A_i, \varepsilon, \tilde{L}, \gamma_{\mathrm{n}}, \gamma_{\mathrm{p}})$

Input: Points $\tilde{x}_i$, $\tilde{z}_i$, function f, domain $\mathcal{X}$, learning rate a_{i+1}, accumulated learning rate A_i, final target accuracy ε, final number of iterations t, smoothness constant $\tilde{L}$, constants $\gamma_{\mathrm{n}}, \gamma_{\mathrm{p}}$. Define $\hat{\varepsilon}_i \leftarrow (A_t\varepsilon)/(2(t-1)A_i)$ as in Theorem A.5, i.e. with $A_t = t(t+1)\gamma_n^2\gamma_p/4\tilde{L}$. $\hat{\Gamma}_i$ defined as in (18) and G_i defined as in (19) i.e.

$$G_i(\lambda) \stackrel{\mathrm{def}}{=} -\hat{\Gamma}_i(\lambda)\langle \nabla f(\tilde{x}_{i+1}^{\lambda}), \tilde{x}_{i+1}^{\lambda} - \tilde{x}_i\rangle + (f(\tilde{x}_{i+1}^{\lambda}) - f(\tilde{x}_i)),$$

for x_{i+1}^{λ} being the result of method (13) when $\hat{\gamma}_i = \hat{\Gamma}_i(\lambda)$.

Output: $\lambda = \frac{a_{i+1}/\gamma_{\mathrm{n}}}{A_i\hat{\gamma}_i + a_{i+1}/\gamma_{\mathrm{n}}}$ for $\hat{\gamma}_i$ such that $G_i(\hat{\Gamma}_i^{-1}(\hat{\gamma}_i)) \leq \hat{\varepsilon}_i$.

if $G_i(\hat{\Gamma}_i^{-1}(1/\gamma_{\mathrm{n}})) \leq \hat{\varepsilon}_i$ **then** $\lambda = \hat{\Gamma}_i^{-1}(1/\gamma_{\mathrm{n}})$
else if $G_i(\hat{\Gamma}_i^{-1}(\gamma_{\mathrm{p}})) \leq \hat{\varepsilon}_i$ **then** $\lambda = \hat{\Gamma}_i^{-1}(\gamma_{\mathrm{p}})$
else
  left $\leftarrow \hat{\Gamma}_i^{-1}(1/\gamma_{\mathrm{n}})$
  right $\leftarrow \hat{\Gamma}_i^{-1}(\gamma_{\mathrm{p}})$
  $\lambda \leftarrow (\mathrm{left} + \mathrm{right})/2$
  while $G_i(\lambda) > \hat{\varepsilon}_i$ **do**
    if $\langle \nabla f(\tilde{x}_{i+1}^{\lambda}), \tilde{x}_{i+1}^{\lambda} - \tilde{x}_i\rangle < 0$ **then** right $\leftarrow \lambda$
    else left $\leftarrow \lambda$
    end if
    $\lambda \leftarrow (\mathrm{left} + \mathrm{right})/2$
  end while
end if
return λ

A.1 AUXILIARY LEMMAS

The following are classical lemmas of convex optimization that we used in this section and that we add for completeness.

Fact A.7. *Let $\psi : Q \to \mathbb{R}$ be a differentiable strongly-convex function. Then*

$$\nabla\psi^*(\tilde{z}) = \arg\max_{\tilde{x} \in Q}\{\langle \tilde{z}, \tilde{x}\rangle - \psi(\tilde{x})\}.$$

Lemma A.8 (Duality of Bregman Div.). $D_\psi(\nabla\psi^*(\tilde{z}), \tilde{x}) = D_{\psi^*}(\nabla\psi(\tilde{x}), \tilde{z})$ *for all* $\tilde{z}, \tilde{x}$.

Proof. From the definition of the Fenchel dual (A.2) and (A.7) we have

$$\psi^*(\tilde{z}) = \langle \nabla\psi^*(\tilde{z}), \tilde{z}\rangle - \psi(\nabla\psi^*(\tilde{z})) \text{ for all } \tilde{z}.$$

Since by the Fenchel-Moreau Theorem we have $\psi^{**} = \psi$, it holds

$$\psi(\tilde{x}) = \langle \nabla\psi(\tilde{x}), \tilde{x}\rangle - \psi^*(\nabla\psi(\tilde{x})), \text{ for all } \tilde{x}.$$

Using the definition of Bregman divergence (A.1) and (A.7):

$$\begin{aligned} D_\psi(\nabla\psi^*(\tilde{z}), \tilde{x}) &= \psi(\nabla\psi^*(\tilde{z})) - \psi(\tilde{x}) - \langle \nabla\psi(\tilde{x}), \nabla\psi^*(\tilde{z}) - \tilde{x}\rangle \\ &= \psi(\nabla\psi^*(\tilde{z})) + \psi^*(\nabla\psi(\tilde{x})) - \langle \nabla\psi(\tilde{x}), \nabla\psi^*(\tilde{z})\rangle \\ &= \psi^*(\nabla\psi(\tilde{x})) - \psi^*(\tilde{z}) - \langle \nabla\psi^*(\tilde{z}), \nabla\psi(\tilde{x}) - \tilde{z}\rangle \\ &= D_{\psi^*}(\nabla\psi(\tilde{x}), \tilde{z}). \end{aligned}$$

□

Lemma A.9 (Triangle inequality of Bregman Divergences). *For all* $\tilde{x}, \tilde{y}, \tilde{z} \in Q$ *we have*

$$D_{\psi^*}(\tilde{x}, \tilde{y}) = D_{\psi^*}(\tilde{z}, \tilde{y}) + D_{\psi^*}(\tilde{x}, \tilde{z}) + \langle \nabla\psi^*(\tilde{z}) - \nabla\psi^*(\tilde{y}), \tilde{x} - \tilde{z}\rangle.$$

Proof.

$$\begin{aligned} &D_{\psi^*}(\tilde{z}, \tilde{y}) + D_{\psi^*}(\tilde{x}, \tilde{z}) + \langle \nabla\psi^*(\tilde{z}) - \nabla\psi^*(\tilde{y}), \tilde{x} - \tilde{z}\rangle \\ &\quad = (\psi^*(\tilde{z}) - \psi^*(\tilde{y}) - \langle \nabla\psi^*(\tilde{y}), \tilde{z} - \tilde{y}\rangle) \\ &\qquad + (\psi^*(\tilde{x}) - \psi^*(\tilde{z}) - \langle \nabla\psi^*(\tilde{z}), \tilde{x} - \tilde{z}\rangle) \\ &\qquad + \langle \nabla\psi^*(\tilde{z}) - \nabla\psi^*(\tilde{y}), \tilde{x} - \tilde{z}\rangle \\ &\quad = \psi^*(\tilde{x}) - \psi^*(\tilde{y}) - \langle \nabla\psi^*(\tilde{y}), \tilde{z} - \tilde{y}\rangle + \langle -\nabla\psi^*(\tilde{y}), \tilde{x} - \tilde{z}\rangle \\ &\quad = D_{\psi^*}(\tilde{x}, \tilde{y}). \end{aligned}$$

□

Lemma A.10. *Given a* σ*-strongly convex function* $\psi(\cdot)$ *the following holds:*

$$D_{\psi^*}(\tilde{z}_1, \tilde{z}_2) \geq \frac{\sigma}{2}\|\nabla\psi^*(\tilde{z}_1) - \nabla\psi^*(\tilde{z}_2)\|^2.$$

Proof. Using the first order optimality condition of the Fenchel dual and (A.7) we obtain

$$\langle \nabla\psi(\nabla\psi^*(\tilde{z}_1)) - \tilde{z}_1 l, \nabla\psi^*(\tilde{z}_2) - \nabla\psi^*(\tilde{z}_1)\rangle \geq 0$$

Using σ-strong convexity of ψ and the previous inequality we have

$$\begin{aligned} D_{\psi^*}(\tilde{z}_1, \tilde{z}_2) &= \psi(\nabla\psi^*(\tilde{z}_2)) - \psi(\nabla\psi^*(\tilde{z}_1)) - \langle \tilde{z}_1, \nabla\psi^*(\tilde{z}_2) - \nabla\psi^*(\tilde{z}_1)\rangle \\ &\geq \frac{\sigma}{2}\|\nabla\psi^*(\tilde{z}_1) - \nabla\psi^*(\tilde{z}_2)\|^2 + \langle \nabla\psi(\nabla\psi^*(\tilde{z}_1)) - \tilde{z}_1, \nabla\psi^*(\tilde{z}_2) - \nabla\psi^*(\tilde{z}_1)\rangle \\ &\geq \frac{\sigma}{2}\|\nabla\psi^*(\tilde{z}_1) - \nabla\psi^*(\tilde{z}_2)\|^2. \end{aligned}$$

□

B REDUCTIONS. PROOFS OF RESULTS IN SECTION 3.

Proof of Theorem 3.1. Let $\mathcal{A}_{\text{ns}}$ be the algorithm in the statement of the theorem. By strong g-convexity of F and the assumptions on $\mathcal{A}_{\text{ns}}$ we have that $\hat{x}_T$ satisfies

$$\frac{\mu}{2}d(\hat{x}_T, x^*)^2 \leq F(\hat{x}_T) - F(x^*) \leq \frac{\mu}{2}\frac{d(x_0, x^*)^2}{2},$$

after $T = \mathrm{Time}_{\mathrm{ns}}(L, \mu, R)$ queries to the gradient oracle. This implies $d(\hat{x}_T, x^*)^2 \leq d(x_0, x^*)^2/2$. Then, by repeating this process $r \stackrel{\mathrm{def}}{=} \lceil \log(\mu \cdot d(x_0, x^*)^2/\varepsilon) - 1 \rceil$ times, using the previous output as input for the next round, we obtain a point $\hat{x}_T^r$ that satisfies

$$F(\hat{x}_T^r) - F(x^*) \leq \frac{\mu \cdot d(\hat{x}_T^{r-1}, x^*)^2}{4} \leq \cdots \leq \frac{\mu \cdot d(x_0, x^*)^2}{4 \cdot 2^{r-1}} \leq \varepsilon.$$

And the total running time is $\mathrm{Time}_{\mathrm{ns}}(L, \mu, R) \cdot r = O(\mathrm{Time}_{\mathrm{ns}}(L, \mu, R) \log(\mu \cdot d(x_0, x^*)^2/\varepsilon)) = O(\mathrm{Time}_{\mathrm{ns}}(L, \mu, R) \log(\mu/\varepsilon))$. □

Proof of Corollary 3.2. Let R be an upper bound on the distance between the initial point x_0 and an optimum x^*, i.e. $d(x_0, x^*) \leq R$. Note that $\|\tilde{x}_0 - \tilde{x}^*\|/R$ is bounded by a constant depending on R by Lemma 2.1.a). Note that γ_{n} and γ_{p} are constants depending on R by Lemma 2.2. As any g-strongly convex function is g-convex, by using Theorem A.5 and Lemma A.6 with $\varepsilon = \mu \frac{R^2}{4}$ we obtain that Algorithm 1 obtains a $\mu \frac{R^2}{4}$-minimizer in at most

$$T = O\left(\frac{\|\tilde{x}_0 - \tilde{x}^*\|}{R} \sqrt{\frac{L}{\mu \gamma_{\mathrm{n}}^2 \gamma_{\mathrm{p}}}} \log\left(\frac{\|\tilde{x}_0 - \tilde{x}^*\|}{R} \sqrt{\frac{L}{\mu \gamma_{\mathrm{n}}^2 \gamma_{\mathrm{p}}}} \right) \right) = O\left(\sqrt{L/\mu} \log(L/\mu) \right)$$

queries to the gradient oracle. Subsequent stages, i.e. calls to Algorithm 1, need a time at most equal to this. The analysis is the same, but we start at the previous output point and take into account that the initial distance to the optimum has decreased. Using the reduction Theorem 3.1 we conclude that given $\varepsilon > 0$ and running Algorithm 1 in stages, we obtain an ε-minimizer of F in

$$O(\sqrt{L/\mu} \log(L/\mu) \log(\mu \cdot d(x_0, x^*)^2/\varepsilon)) = O^*(\sqrt{L/\mu} \log(\mu/\varepsilon)),$$

queries to the gradient oracle.

As advanced in the main paper, each of the stages of the algorithm resulting from combining Theorem 3.1 and Corollary 3.2 reduces the distance to x^* by a factor of $1/\sqrt{2}$. This means that subsequent stages can be run using a geodesic map centered at the new starting point, and with the new parameter R being the previous one reduced by a factor of $1/\sqrt{2}$. This reduces the constants incurred by the deformation of the geometry which ultimately reduces the overall constant in the rate. Note that in order to perform the method with the recentering steps, we need the function F to be defined over at least $\mathrm{Exp}_{x_0}(\bar{B}(0, R \cdot (1 + 2^{-1/2})))$, since subsequent centers are only guaranteed to be $\leq R/\sqrt{2}$ close to x^*, and they could get slightly farther from x_0. □

B.1 Proof of Theorem 3.3

The algorithm is the following. We successively regularize the function with strongly g-convex regularizers in this way $F^{(\mu_i)}(x) \stackrel{\mathrm{def}}{=} F(x) + \frac{\mu_i}{2} d(x, x_0)^2$ for $i \geq 0$. For each $i \geq 0$, we use the algorithm $\mathcal{A}$ on the function $F^{(\mu_i)}$ for the time in the statement of the theorem and obtain a point $\hat{x}_{i+1}$, starting from point $\hat{x}_i$, where $\hat{x}_0 = x_0$. The regularizers are decreased exponentially $\mu_{i+1} = \mu_i/2$ until we reach roughly $\mu_T = \varepsilon/R^2$, see below for the precise value. Let's see how this algorithm works. We first state the following fact, that says that indeed $\frac{\mu_i}{2} d(x, x_0)^2$ is a strongly g-convex regularizer. Let D be the diameter of $\mathcal{M}$. We define the following quantities

$$\mathcal{K}_R^+ \stackrel{\mathrm{def}}{=} \begin{cases} 1 & \text{if } K_{\max} \leq 0 \\ \sqrt{K_{\max}} D \cot(\sqrt{K_{\max}} D) & \text{if } K_{\max} > 0 \end{cases}$$

$$\mathcal{K}_R^- \stackrel{\mathrm{def}}{=} \begin{cases} \sqrt{-K_{\min}} D \coth(\sqrt{-K_{\min}} D) & \text{if } K_{\min} < 0 \\ 1 & \text{if } K_{\min} \geq 0 \end{cases}$$

Here $K_{\max}$ and $K_{\min}$ are the upper and lower bounds on the sectional curvature of the manifold $\mathcal{M}$. In Theorem 3.3, it is $D = 2R$.

Fact B.1. *Let $\mathcal{M} = \mathrm{Exp}_{x_0}(\bar{B}(0, R))$ be a manifold with sectional curvature bounded below and above by $K_{\min}$ and $K_{\max}$, respectively, where $x_0 \in \mathcal{M}$ is a point. The function $f : \mathcal{M} \to \mathbb{R}$ defined as $f(x) = \frac{1}{2} d(x, x_0)^2$ is $\mathcal{K}_R^+$-g-strongly convex and $\mathcal{K}_R^-$-smooth.*

The result regarding strong convexity can be found, for instance, in Alimisis et al. (2019) and it is a direct consequence of the following inequality, which can also be found in Alimisis et al. (2019):

$$d(y,x_0)^2 \geq d(x,x_0)^2 - 2\langle \operatorname{Exp}_x^{-1}(x_0), y-x\rangle + \mathcal{K}_R^+ d(x,y)^2,$$

along with the fact that $\operatorname{grad} f(x) = -\operatorname{Exp}_x^{-1}(x_0)$. The result regarding smoothness is, similarly, obtained from the following inequality:

$$d(y,x_0)^2 \leq d(x,x_0)^2 - 2\langle \operatorname{Exp}_x^{-1}(x_0), y-x\rangle + \mathcal{K}_R^- d(x,y)^2,$$

which can be found in Zhang & Sra (2016) (Lemma 6). Alternatively, one can derive these inequalities from upper and lower bounds on the Hessian of $f(x) = \frac{1}{2}d(x,x_0)$, cf. Jost & Jost (2008), Theorem 4.6.1, as it was done in Lezcano-Casado (2020).

We prove now that the regularization makes the minimum to be closer to x_0, so the assumption of the Theorem on $\hat{F}$ holds for the functions we use. Define x_{i+1} as the minimizer of $F^{(\mu_i)}$.

Lemma B.2. *We have* $d(x_{i+1},x_0) \leq d(x^*,x_0)$.

Proof. By the fact that x_{i+1} is the minimizer of $F^{(\mu_i)}$ we have $F^{(\mu_i)}(x_{i+1}) - F^{(\mu_i)}(x^*) \leq 0$. Note that by g-strong convexity, equality only holds if $x_{i+1} = x^*$ which only happens if $x_0 = x_{i+1} = x^*$. By using the definition of $F^{(\mu_i)}(x) = F(x) + \frac{\mu_i}{2}d(x,x_0)^2$ we have:

$$\begin{aligned} &F(x_{i+1}) + \frac{\mu_i}{2}d(x_{i+1},x_0)^2 - F(x^*) - \frac{\mu_i}{2}d(x^*,x_0)^2 \leq 0 \\ &\quad \Rightarrow \;\; d(x_{i+1},x_0) \leq d(x^*,x_0), \end{aligned}$$

where in the last step we used the fact $F(x_{i+1}) - F(x^*) \geq 0$ that holds because x^* is the minimizer of F. □

We note that previous techniques proved and used the fact that $d(x_{i+1},x^*) \leq d(x_0,x^*)$ instead Allen Zhu & Hazan (2016). But crucially, we need our former lemma in order to prove the bound for our non-Euclidean case. Our technique could be applied to Allen Zhu & Hazan (2016) to decrease the constants of the Euclidean reduction. Now we are ready to prove the theorem.

Proof of Theorem 3.3. We recall the definitions above. $F^{(\mu_i)}(x) = F(x) + \frac{\mu_i}{2}d(x,x_0)^2$. We start with $\hat{x}_0 = x_0$ and compute $\hat{x}_{i+1}$ using algorithm $\mathcal{A}$ with starting point $\hat{x}_i$ and function $F^{(\mu_i)}$ for time $\operatorname{Time}(L^{(i)},\mu^{(i)},\mathcal{M},R)$, where $L^{(i)}$ and $\mu^{(i)}$ are the smoothness and strong g-convexity parameters of $F^{(\mu_i)}$. We denote by x_{i+1} the minimizer of $F^{(\mu_i)}$. We pick $\mu_i = \mu_{i-1}/2$ and we will choose later the value of μ_0 and the total number of stages. By the assumption of the theorem on $\mathcal{A}$, we have that

$$F^{(\mu_i)}(\hat{x}_{i+1}) - \min_{x\in\mathcal{M}} F^{(\mu_i)}(x) = F^{(\mu_i)}(\hat{x}_{i+1}) - F^{(\mu_i)}(x_{i+1}) \leq \frac{F^{(\mu_i)}(\hat{x}_i) - F^{(\mu_i)}(x_{i+1})}{4}. \quad (27)$$

Define $D_i \overset{\text{def}}{=} F^{(\mu_i)}(\hat{x}_i) - F^{(\mu_i)}(x_{i+1})$ to be the initial objective distance to the minimum on function $F^{(\mu_i)}$ before we call $\mathcal{A}$ for the $(i+1)$-th time. At the beginning, we have the upper bound $D_0 = F^{(\mu_0)}(\hat{x}_0) - \min_x F^{(\mu_0)}(x) \leq F(x_0) - F(x^*)$. For each stage $i \geq 1$, we compute that

$$\begin{aligned} D_i &= F^{(\mu_i)}(\hat{x}_i) - F^{(\mu_i)}(x_{i+1}) \\ &\overset{①}{=} F^{(\mu_{i-1})}(\hat{x}_i) - \frac{\mu_{i-1}-\mu_i}{2}d(x_0,\hat{x}_i)^2 - F^{(\mu_{i-1})}(x_{i+1}) + \frac{\mu_{i-1}-\mu_i}{2}d(x_0,x_{i+1})^2 \\ &\overset{②}{\leq} F^{(\mu_{i-1})}(\hat{x}_i) - F^{(\mu_{i-1})}(x_i) + \frac{\mu_{i-1}-\mu_i}{2}d(x_0,x_{i+1})^2 \\ &\overset{③}{\leq} \frac{D_{i-1}}{4} + \frac{\mu_i}{2}d(x_0,x_{i+1})^2 \\ &\overset{④}{\leq} \frac{D_{i-1}}{4} + \frac{\mu_i}{2}d(x_0,x^*)^2. \end{aligned}$$

Above, ① follows from the definition of $F^{(\mu_i)}(\cdot)$ and $F^{(\mu_{i-1})}(\cdot)$; ② follows from the fact that x_i is the minimizer of $F^{(\mu_{i-1})}(\cdot)$. We also drop the negative term $-(\mu_{i-1}-\mu_i)d(x_0,\hat{x}_i)/2$. ③ follows from the definition of D_{i-1}, the assumption on $\mathcal{A}$, and the choice $\mu_i = \mu_{i-1}/2$ for $i \geq 1$; and ④ follows from Lemma B.2. Now applying the above inequality recursively, we have

$$D_T \leq \frac{D_0}{4^T} + d(x_0,x^*)^2 \cdot (\frac{\mu_T}{2} + \frac{\mu_{T-1}}{8} + \cdots) \leq \frac{F(x_0)-F(x^*)}{4^T} + \mu_T \cdot d(x_0,x^*)^2. \tag{28}$$

We have used the choice $\mu_i = \mu_{i-1}/2$ for the second inequality. Lastly, we can prove that $\hat{x}_T$, the last point computed, satisfies

$$\begin{aligned}
F(\hat{x}_T) - F(x^*) &\overset{①}{\leq} F^{(\mu_T)}(\hat{x}_T) - F^{(\mu_T)}(x^*) + \frac{\mu_T}{2} d(x_0,x^*)^2 \\
&\overset{②}{\leq} F^{(\mu_T)}(\hat{x}_T) - F^{(\mu_T)}(x_{T+1}) + \frac{\mu_T}{2} d(x_0,x^*)^2 \\
&\overset{③}{=} D_T + \frac{\mu_T}{2} d(x_0,x^*)^2 \\
&\overset{④}{\leq} \frac{F(x_0)-F(x^*)}{4^T} + \frac{3\mu_T}{2} d(x_0,x^*)^2.
\end{aligned}$$

We use the definition of $F^{(\mu_T)}$ in ① and drop $-\frac{\mu_T}{2}d(x_0,\hat{x}_T)^2$. In ② we use the fact that x_{T+1} is the minimizer of $F^{(\mu_T)}$. The definition of D_T is used in ③. We use inequality (28) for step ④. Finally, by choosing $T = \lceil \log_2(\Delta/\varepsilon)/2 \rceil + 1$ and $\mu_0 = \Delta/R^2$ we obtain that the point $\hat{x}_T$ satisfies

$$F(\hat{x}_T) - F(x^*) \leq \frac{F(x_0)-F(x^*)}{4\Delta/\varepsilon} + \frac{3\mu_0}{8\Delta/\varepsilon} d(x_0,x^*)^2 \leq \frac{\varepsilon}{4} + \frac{3\varepsilon}{8} < \varepsilon,$$

and can be computed in time $\sum_{t=0}^{T-1} \mathrm{Time}(L + 2^{-t}\mu_0\mathcal{K}_R^-, 2^{-t}\mu_0\mathcal{K}_R^+, \mathcal{M}, R)$, since by Fact B.1 the function $F^{(\mu_t)}$ is $L + 2^{-t}\mu_0\mathcal{K}_R^-$ smooth and $2^{-t}\mu_0\mathcal{K}_R^+$ g-strongly convex. □

B.2 Example 3.4

We use the algorithm in Corollary 3.2 as the algorithm $\mathcal{A}$ of the reduction of Theorem 3.3. Given $\mathcal{M} = \mathcal{H}$ or $\mathcal{M} = \mathcal{S}$, the assumption on $\mathcal{A}$ is satisfied for $\mathrm{Time}(L,\mu,\mathcal{M},R) = O(\sqrt{L/\mu}\log(L/\mu))$. Indeed, if Δ is a bound on the gap $\hat{F}(x_0) - \hat{F}(x^*) = \hat{F}(x_0) - \min_{x\in\mathcal{M}} \hat{F}(x) = \hat{F}(x_0) - \min_{x\in \mathrm{Exp}_{x_0}(\bar{B}(0,R))} \hat{F}(x)$ for some μ strongly g-convex $\hat{F}$, then we know that $d(x_0,x^*)^2 \leq \frac{2\Delta}{\mu}$. By calling the algorithm in Corollary 3.2 with $\varepsilon = \frac{\Delta}{4}$ we require a time that is

$$\begin{aligned}
O(\sqrt{L/\mu}\log(L/\mu)\log(\mu \cdot d(x_0,x^*)^2/(\Delta/4))) &= O(\sqrt{L/\mu}\log(L/\mu)\log(\mu\cdot(2\Delta/\mu)/(\Delta/4))) \\
&= O(\sqrt{L/\mu}\log(L/\mu)).
\end{aligned}$$

Let $T = \lceil \log_2(\Delta/\varepsilon)/2 \rceil + 1$. The reduction of Theorem 3.3 gives an algorithm with rates

$$\begin{aligned}
&\sum_{t=0}^{T-1} \mathrm{Time}(L + 2^{-t}\mu_0\mathcal{K}_R^-, 2^{-t}\mu_0\mathcal{K}_R^+, \mathcal{M}, R) \\
&\quad = O\left(\sum_{t=0}^{T-1} \sqrt{\frac{L}{2^{-t}\mathcal{K}_R^+\Delta/R^2} + \frac{\mathcal{K}_R^-}{\mathcal{K}_R^+}} \cdot \log\left(\frac{L}{2^{-t}\mathcal{K}_R^+\Delta/R^2} + \frac{\mathcal{K}_R^-}{\mathcal{K}_R^+}\right)\right) \\
&\quad \overset{①}{=} O\left(\left(\sqrt{\frac{L}{\mathcal{K}_R^+\varepsilon}} + \sqrt{\frac{\mathcal{K}_R^-}{\mathcal{K}_R^+}}\log(\Delta/\varepsilon)\right)\log\left(\frac{L}{\mathcal{K}_R^+\varepsilon} + \frac{\mathcal{K}_R^-}{\mathcal{K}_R^+}\right)\right) \\
&\quad = \widetilde{O}(\sqrt{L/\varepsilon})
\end{aligned}$$

In ① we have used Minkowski's inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. We used the value $\mu_0 = \Delta/R^2$. In order to group the sum of the first summands, we used the fact that $\sqrt{1/\varepsilon} + \sqrt{1/2\varepsilon} + \cdots = O(\sqrt{1/\varepsilon})$,

along with the fact $2^{-(T-1)}\mu_0 \geq \log(\varepsilon/\Delta)$. We added up the group of second summands. For the log factor, we upper bounded $L/(2^{-t}\mathcal{K}_R^+\Delta/R^2) = O(L/\mathcal{K}_R^+\varepsilon)$, for $t < T$. Note that by L-smoothness and the diameter being $2R$, we have $\Delta \leq 2LR^2$ so $\sqrt{\mathcal{K}_R^-/\mathcal{K}_R^+}\log(\Delta/\varepsilon) = \widetilde{O}(1)$.

C GEOMETRIC RESULTS. PROOFS OF LEMMAS 2.1, 2.2 AND 2.3

In this section we prove the lemmas that take into account the deformations of the geometry and the geodesic map h to obtain relationships between F and f. Namely Lemma 2.1, Lemma 2.2 and Lemma 2.3. First, we recall the characterizations of the geodesic map and some consequences. Then in Appendix C.2, Appendix C.3 and Appendix C.4 we prove the results related to distances angles and gradient deformations, respectively. That is, each of the three parts of Lemma 2.1. In Appendix C.4 we also prove Lemma 2.3, which comes naturally after the proof of Lemma 2.1.c). Finally, in Appendix C.5 we prove Lemma 2.2. Before this, we note that we can assume without loss of generality that the curvature of our manifolds of interest can be taken to be $K \in \{1, -1\}$. One can see that the final rates depend on K through R, L and μ.

Remark C.1. For a function $F : \mathcal{M} \to \mathbb{R}$ where $\mathcal{M} = \mathcal{H}$ or $\mathcal{M} = \mathcal{S}$ is a manifold of constant sectional curvature $K \notin \{1, -1, 0\}$, we can apply a rescaling to the Gnomonic or Beltrami-Klein projection to define a function on a manifold of constant sectional curvature $K \in \{1, -1\}$. Namely, we can map $\mathcal{M}$ to $\mathcal{B}$ via h, then we can rescale $\mathcal{B}$ by multiplying each vector in $\mathcal{B}$ by the factor $\sqrt{|K|}$ and then we can apply the inverse geodesic map for the manifold of curvature $K \in \{1, -1\}$. If R is the original bound of the initial distance to an optimum, and F is L-smooth and μ-strongly g-convex (possibly with $\mu = 0$) then the initial distance bound becomes $\sqrt{|K|}R$ and the induced function becomes $L/|K|$-smooth and $\mu/|K|$-strongly g-convex. This is a simple consequence of the transformation rescaling distances by a factor of $\sqrt{|K|}$, i.e. if $r : \mathcal{M}_K \to \mathcal{M}_{K/|K|}$ is the rescaling function, then $d_K(x, y)\sqrt{|K|} = d_{K/|K|}(r(x), r(y))$, where $d_c(\cdot, \cdot)$ denotes the distance on the manifold of constant sectional curvature c.

C.1 PRELIMINARIES

We recall our characterization of the geodesic map. Given two points $\tilde{x}, \tilde{y} \in \mathcal{B}$, we have that $d(x, y)$, the distance between x and y with the metric of $\mathcal{M}$, satisfies

$$C_K(d(x, y)) = \frac{1 + K\langle \tilde{x}, \tilde{y}\rangle}{\sqrt{1 + K\|\tilde{x}\|^2} \cdot \sqrt{1 + K\|\tilde{y}\|^2}}. \tag{29}$$

And since the expression is symmetric with respect to rotations, $\mathcal{X} = h(\mathcal{M})$ is a closed ball of radius $\tilde{R}$, with $C_K(R) = (1 + K\tilde{R}^2)^{-1/2}$. Equivalently,

$$\begin{aligned} \tilde{R} &= \tan(R) \quad \text{if } K = 1, \\ \tilde{R} &= \tanh(R) \quad \text{if } K = -1. \end{aligned} \tag{30}$$

Similarly, we can write the distances as

$$\begin{aligned} d(x, y) &= \arccos\left(\frac{1 + \langle \tilde{x}, \tilde{y}\rangle}{\sqrt{1 + \|\tilde{x}^2\|}\sqrt{1 + \|\tilde{y}^2\|}}\right) \quad \text{if } K = 1, \\ d(x, y) &= \operatorname{arccosh}\left(\frac{1 - \langle \tilde{x}, \tilde{y}\rangle}{\sqrt{1 - \|\tilde{x}^2\|}\sqrt{1 - \|\tilde{y}^2\|}}\right) \quad \text{if } K = -1, \end{aligned} \tag{31}$$

Alternatively, we have the following expression for the distance $d(x, y)$ when $K = -1$. Let $\tilde{a}, \tilde{b}$ be the two points of intersection of the ball $\mathcal{B} = B(0, 1)$ with the line joining $\tilde{x}, \tilde{y}$, so the order of the points in the line is $\tilde{a}, \tilde{x}, \tilde{y}, \tilde{b}$. Then

$$d(x, y) = \frac{1}{2}\log\left(\frac{\|\tilde{a} - \tilde{y}\|\|\tilde{x} - \tilde{b}\|}{\|\tilde{a} - \tilde{x}\|\|\tilde{b} - \tilde{y}\|}\right) \text{ if } K = -1. \tag{32}$$

We will use this expression when working with the hyperbolic space. A simple elementary proof of the equivalence of the expressions in (31) and (32) is the following. We can assume without loss of

generality that we work with the hyperbolic plane, i.e. $d = 2$. By rotational symmetry, we can also assume that $\tilde{x} = (x_1, x_2)$ and $\tilde{y} = (y_1, y_2)$, for $x_1 = y_1$. In fact, it is enough to prove it in the case $x_2 = 0$ because we can split a general segment into two, each with one endpoint at $(x_1, 0)$, and then add their lengths up. So according to (31) and (32), respectively, we have

$$\frac{1}{\cosh^2(d(x,y))} = \frac{(1-x_1^2)(1-y_1^2-y_2^2)}{(1-x_1^2)^2} = \frac{(1-x_1^2-y_2^2)}{1-x_1^2}.$$

$$\begin{aligned} d(x,y) &= \frac{1}{2}\log\left(\frac{(\sqrt{1-y_1^2}+y_2)(\sqrt{1-x_1^2})}{(\sqrt{1-x_1^2})(\sqrt{1-y_1^2}-y_2)}\right) = \frac{1}{2}\log\left(\frac{1+y_2/\sqrt{1-x_1^2}}{1-y_2/\sqrt{1-x_1^2}}\right) \\ &= \operatorname{arctanh}\left(\frac{y_2}{\sqrt{1-x_1^2}}\right) \end{aligned}$$

where we have used the equality $\tanh(t) = \frac{1}{2}\log(\frac{1+t}{1-t})$. Now, using the trigonometric identity $\frac{1}{\cosh^2(t)} = 1 - \tanh^2(t)$, for $t = d(x, y)$, we obtain that the two expressions above are equal. See Theorem 7.4 in (Greenberg, 1993) (p. 268) for more details about the distance formula under this geodesic map.

The spherical case is of a remarkable simplicity. If we have a d-sphere of radius 1 centered at 0, we can see the transformation as the projection onto the plane $x_d = 1$. Given two points $\mathbf{x} = (\tilde{x}, 1)$, $\mathbf{y} = (\tilde{y}, 1)$ then the angle between these two vectors is the distance of the projected points on the sphere so we have $\cos(d(x, y)) = \langle \mathbf{x}, \mathbf{y}\rangle / \|\mathbf{x}\|\|\mathbf{y}\|$ which is equivalent to the corresponding formula in 31.

C.2 Distance deformation

Lemma C.2. *Let $\mathcal{H} = \operatorname{Exp}_{x_0}(\bar{B}(0, R))$ be a subset of the hyperbolic space with constant sectional curvature $K = -1$. Let $x, y \in \mathcal{H}$ be two different points. Then, we have*

$$1 \leq \frac{d(x,y)}{\|\tilde{x} - \tilde{y}\|} \leq \cosh^2(R).$$

Proof. We can assume without loss of generality that the dimension is $d = 2$. As in (30), let $\tilde{R} = \tanh(R)$, so any point $\tilde{x} \in \mathcal{X}$ satisfies $\|\tilde{x}\| \leq \tilde{R}$, or equivalently $d(x, x_0) \leq R$. Recall $\tilde{x}_0 = h(x_0) = 0$. Without loss of generality, we parametrize an arbitrary segment of length ℓ in $\mathcal{X}$ by two endpoints $\tilde{x}, \tilde{y}$ with coordinates $\tilde{x} = (x_1, x_2)$ and $\tilde{y} = (x_1 - \ell, x_2)$, for $0 \leq x_2 \leq \tilde{R}$, $0 \leq x_1 \leq \sqrt{\tilde{R}^2 - x_2^2}$ and $0 < \ell \leq x_1 + \sqrt{\tilde{R}^2 - x_2^2}$. Let $\mathfrak{d}(x_1, x_2, \ell) \stackrel{\text{def}}{=} \frac{d(x,y)}{\ell}$, the quantity we aim to bound. We will prove the upper bound on $\mathfrak{d}(x_1, x_2, \ell)$ in three steps.

1. If $x_1 = \ell$ then $\mathfrak{d}(\cdot)$ is larger the larger x_1 is. This allows to prove that it is enough to consider points with the extra constraint $\ell \leq x_1$.
2. The partial derivative of $\mathfrak{d}(\cdot)$ with respect to x_1, whenever $\ell \leq x_1$, is non-negative. So we can just look at the points for which $x_1 = \sqrt{\tilde{R}^2 - x_2^2}$.
3. With the constraints above, $\mathfrak{d}(\cdot)$ is larger the smaller ℓ is. So we have $\mathfrak{d}(x_1, x_2, \ell) \leq \lim_{\ell \to 0} \mathfrak{d}(\sqrt{\tilde{R}^2 - x_2^2}, x_2, \ell) = \sqrt{1 - x_2^2}/(1 - \tilde{R}^2)$. This expression is maximized at $x_2 = 0$ and evaluates to $1/(1 - \tanh^2(R)) = \cosh^2(R)$.

We proceed now to prove the steps above. For the first step, we note

$$\mathfrak{d}(x_1, x_2, x_1) = \frac{1}{2x_1}\log\left(\frac{\sqrt{1-x_2^2}(\sqrt{1-x_2^2}+x_1)}{\sqrt{1-x_2^2}(\sqrt{1-x_2^2}-x_1)}\right) = \frac{1}{2x_1}\log\left(1 + \frac{2x_1}{\sqrt{1-x_2^2}-x_1}\right).$$

We prove that the inverse of this expression is not increasing with respect to x_1. By taking a partial derivative:

$$\frac{\partial(1/\mathfrak{d}(x_1,x_2,x_1))}{\partial x_1} = 2\frac{\frac{-2x_1\sqrt{1-x_2^2}}{1-x_2^2-x_1^2} + \log(1+2x_1/(\sqrt{1-x_2^2}-x_1))}{\log(1+2x_1/(\sqrt{1-x_2^2}-x_1))^2} \overset{?}{\leq} 0$$
$$\iff \frac{2x_1\sqrt{1-x_2^2}}{1-x_2^2-x_1^2} - \log(1+(2x_1\sqrt{1-x_2^2}+2x_1^2)/(1-x_2^2-x_1^2)) \overset{?}{\geq} 0.$$

In order to see that the last inequality is true, note that the expression on the left hand side is 0 when $x_1 = x_2 = 0$. And the partial derivatives of this with respect to x_1 and x_2, respectively, are:

$$\frac{4\sqrt{1-x_2^2}x_1^2}{(1-x_2^2-x_1^2)^2} \text{ and } \frac{4x_2x_1^3}{\sqrt{1-x_2^2}(1-x_2^2-x_1^2)^2}.$$

Both are greater than 0 in the interior of the domain $0 \leq x_2 \leq \tilde{R}$, $0 \leq x_1 \leq \sqrt{\tilde{R}^2-x_2^2}$ and at least 0 in the border. Now we use this monotonicity to prove that we can consider $\ell \leq x_1$ only. Suppose $\ell > x_1$. The segment ℓ is divided into two parts by the e_2 axis and we can assume without loss of generality that the negative part is no greater than the other, i.e. $x_1 \geq \ell - x_1$. Otherwise, we can perform the computations after a symmetry over the e_2 axis. Let $\tilde{r}$ be the point $(0, x_2)$. We want to see that the segment from $\tilde{x}$ to $\tilde{r}$ gives a greater value of $\mathfrak{d}(\cdot)$:

$$\frac{d(x,r)}{x_1} \geq \frac{d(x,y)}{\ell} \iff d(x,r)(x_1+(\ell-x_1)) \geq x_1(d(x,r)+d(r,y))$$
$$\iff d(x,r)/x_1 \geq d(r,y)/(\ell-x_1),$$

and the last inequality holds true by the monotonicity we just proved.

In order to prove the second step, we take the partial derivative of $\mathfrak{d}(x_1,x_2,\ell)$ with respect to x_1. We have

$$\mathfrak{d}(x_1,x_2,\ell) = \frac{1}{2\ell}\log\left(\frac{1+\ell/(\sqrt{1-x_2^2}-x_1)}{1-\ell/\sqrt{1-x_2^2}+x_1}\right),$$
$$\frac{\partial\mathfrak{d}(x_1,x_2,\ell)}{\partial x_1} = \frac{\sqrt{1-d^2}(2x_1-\ell)}{2(1-x_2^2-x_1^2)(1-x_2^2-(x_1-\ell)^2)}.$$

And the derivative is positive in the domain we are considering.

We now prove step 3. We want to show that $\mathfrak{d}(\sqrt{\tilde{R}^2-x_2^2}, x_2, \ell\cdot)$ decreases with ℓ, within our constraints $\ell \leq x_1 = \sqrt{\tilde{R}^2-x_2^2}$, $0 \leq x_2 \leq \tilde{R}$. If we split the segment joining $\tilde{x}$ and $\tilde{y}$ in half with, respect to the metric in $\mathcal{B}$, we see that due to the monotonicity proved in step 1, the segment that is farther to the origin is longer in $\mathcal{M}$ than the other one and so $\mathfrak{d}(\cdot)$ is greater for this half of the segment than for the original one. In symbols, let $\tilde{r}$ be the middle point of the segment joining $\tilde{x}$ and $\tilde{y}$. We have by monotonicity that $\mathfrak{d}(x_1,x_2,\ell/2) \geq \mathfrak{d}(x_1,x_2-\ell/2,\ell/2)$. So $\mathfrak{d}(x_1,x_2,\ell/2) = \frac{d(\tilde{x},\tilde{r})}{\ell/2} \geq \frac{d(\tilde{x},\tilde{r})+d(\tilde{r},\tilde{y})}{\ell} = \mathfrak{d}(x_1,x_2,\ell)$. Thus,

$$\begin{aligned}
\mathfrak{d}(x_1,x_2,\ell) &\leq \lim_{\ell\to 0}\mathfrak{d}(\sqrt{\tilde{R}^2-x_2^2},x_2,\ell) \\
&= \lim_{\ell\to 0}\frac{1}{2\ell}\log\left(\frac{1+\ell/\left(\sqrt{1-x_2^2}-\sqrt{\tilde{R}^2-x_2^2}\right)}{1-\ell/\left(\sqrt{1-x_2^2}+\sqrt{\tilde{R}^2-x_2^2}\right)}\right) \\
&\overset{①}{=} \lim_{\ell\to 0}\frac{\sqrt{1-x_2^2}}{1-\tilde{R}^2-2\ell\sqrt{\tilde{R}^2-x_2^2}+\ell^2} \\
&= \frac{\sqrt{1-x_2^2}}{1-\tilde{R}^2}.
\end{aligned}$$

We used L'Hôpital's rule for ①. We can maximize the last the result of the limit by setting $x_2 = 0$ and obtain that for any two different $\tilde{x}, \tilde{y} \in \mathcal{X}$

$$\frac{d(x,y)}{\|\tilde{x}-\tilde{y}\|} \leq \frac{1}{1-\tilde{R}^2} = \frac{1}{1-\tanh^2(R)} = \cosh^2(R).$$

The lower bound is similar, assume that $\ell > x_1$ and define $\tilde{r}$ as above. We assume again without loss of generality that $x_1 \geq \ell - x_1$. Then

$$\frac{d(x,r)+d(r,y)}{\ell} \geq \frac{d(x,r)}{\ell - x_1} \iff \frac{d(r,y)}{x_1} \geq \frac{d(x,r)}{\ell - x_1}$$

and the latter is true by the monotonicity proved in step 1. This means that we can also consider $\ell \leq x_1$. But this time, according to step 2, we want x_1 to be the lowest possible, so it is enough to consider $x_1 = \ell$. Using step 1 again, we obtain that the lowest value of $\mathfrak{d}(\cdot)$ can be bounded by the limit $\lim_{\ell \to 0} \mathfrak{d}(\ell, x_2, \ell)$ which using L'Hôpital's rule in ① is

$$\begin{aligned}
\mathfrak{d}(x_1, x_2, \ell) &\geq \lim_{\ell \to 0} \mathfrak{d}(\ell, x_2, \ell) \\
&= \lim_{\ell \to 0} \frac{1}{2\ell} \log\left(1 + \frac{2\ell}{\sqrt{1-x_2^2} - \ell}\right) \\
&\overset{①}{=} \lim_{\ell \to 0} \frac{\frac{2(\sqrt{1-x_2^2}-\ell)+2\ell}{(\sqrt{1-x_2^2}-\ell)^2}}{2(1+2\ell/(\sqrt{1-x_2^2}-\ell))} \\
&= \frac{1}{\sqrt{1-x_2^2}}.
\end{aligned}$$

The expression is minimized at $x_2 = 0$ and evaluates to 1. □

The proof of the corresponding lemma for the sphere is analogous, we add it for completeness.

Lemma C.3. *Let $\mathcal{S} = \mathrm{Exp}_{x_0}(\bar{B}(0,R))$ be a subset of the sphere with constant sectional curvature $K = 1$ and $R < \frac{\pi}{2}$. Let $x, y \in \mathcal{S}$ be two different points. Then, we have*

$$\cos^2(R) \leq \frac{d(x,y)}{\|\tilde{x}-\tilde{y}\|} \leq 1.$$

Proof. We proceed in a similar way than with the hyperbolic case. We can also work with $d = 2$ only, since $\tilde{x}, \tilde{y}$ and $\tilde{x}_0$ lie on a plane. We parametrize a general pair of points as $\tilde{x} = (x_1, x_2) \in \mathcal{X}$ and $y = (x_1 - \ell, x_2) \in \mathcal{X}$, so $x_1^2 + x_2^2 \leq \tilde{R}^2$, for $\tilde{R} = \tan(R)$ and by definition $\ell = \|\tilde{x} - \tilde{y}\|$.

Let $\mathfrak{d}(x_1, x_2, \ell) \overset{\text{def}}{=} d(x,y)/\|\tilde{x} - \tilde{y}\|$. We proceed to prove the result in three steps, similarly to the hyperbolic case.

1. If $x_1 = \ell$ then $\mathfrak{d}(x_1, x_2, \ell)$ decreases whenever x_1 increases. This allows to prove that it is enough to consider points in which $\ell \leq x_1$.
2. $\frac{\partial \mathfrak{d}(\cdot)}{\partial x_1} \leq 0$, whenever $\ell \leq x_1$. So we can consider $x_1 = \sqrt{\tilde{R}^2 - x_2^2}$ only.
3. With the constraints above, $\mathfrak{d}(\cdot)$ increases with ℓ, so in order to lower bound $\mathfrak{d}(\cdot)$ we can consider $\lim_{\ell \to 0} \mathfrak{d}(\sqrt{\tilde{R} - x_2}, x_2, \ell) = \sqrt{1+x_2^2}/(1+\tilde{R}^2)$. This is minimized at $x_2 = 0$ and evaluates to $1/(1+\tilde{R}^2)$.

For the first step, we compute the partial derivative:

$$\frac{\partial \mathfrak{d}(x_1, x_2, x_1)}{\partial x_1} = \frac{x_1\sqrt{1+x_2^2}/(1+x_1^2+x_2^2) - \arccos\left(\sqrt{(1+x_2^2)/(1+x_1^2+x_2^2)}\right)}{x_1^2}. \tag{33}$$

In order to see that it is non-positive, we compute the partial derivative of the denominator with respect to x_2 and obtain

$$\frac{2x_1^3 x_2}{\sqrt{1+x_2^2}(1+x_1^2+x_2^2)} \geq 0.$$

so in order to maximize (33) we set $x_2 = \sqrt{\tilde{R} - x_1^2}$. In that case, the numerator is

$$\frac{x_1\sqrt{1+R^2-x_1^2}}{1+R^2} - \arccos\left(\sqrt{\frac{1+R^2-x_1^2}{1+R^2}}\right), \tag{34}$$

and its derivative with respect to x_1 is

$$-\frac{2x_1^2}{(1+R^2)\sqrt{1+R^2-x_1^2}} \leq 0.$$

and given that (34) with $x_1 = 0$ evaluates to 0 we conclude that (33) is non-positive. Similarly to Lemma C.2, suppose the horizontal segment that joins $\tilde{x}$ and $\tilde{y}$ passes through $\tilde{r} \stackrel{\text{def}}{=} (0, x_2)$. And suppose without loss of generality that $d(x,r) \geq d(r,y)$, i.e. $x_1 \geq \ell - x_1$. Then by the monotonicity we just proved, we have

$$\frac{d(x,r)}{\|\tilde{x}-\tilde{r}\|} = \mathfrak{d}(x_1, x_2, x_1) \leq \mathfrak{d}(\ell - x_1, x_2, \ell - x_1) = \frac{d(r,y)}{\|\tilde{r}-\tilde{y}\|}. \tag{35}$$

And this implies $\mathfrak{d}(x_1, x_2, x_1) \leq \mathfrak{d}(x_1, x_2, \ell)$. Indeed, that is equivalent to show

$$\frac{d(x,r)}{\|\tilde{x}-\tilde{r}\|} \leq \frac{d(x,y)}{\|\tilde{x}-\tilde{y}\|} = \frac{d(x,r)+d(r+y)}{\|\tilde{x}-\tilde{r}\|+\|\tilde{r}-\tilde{y}\|}.$$

Which is true, since after simplifying we arrive to (35). So in order to lower bound $\mathfrak{d}(\cdot)$, it is enough to consider $\ell \leq x_1$.

We focus on step 2 now. We have

$$\frac{\partial \mathfrak{d}(x_1, x_2, \ell)}{\partial x_1} = \frac{\sqrt{1+x_2^2}(\ell - 2x_1)}{(1+x_2^2+(\ell-x_1)^2)(1+x_2^2+x_1^2)}.$$

which is non-positive given the restrictions we imposed after step 1. So in order to lower bound $\mathfrak{d}(\cdot)$ we can consider $x_1 = \sqrt{\tilde{R} - x_2^2}$ only.

Finally, in order to complete step 3 we compute

$$\begin{aligned}\frac{\partial \mathfrak{d}(\sqrt{\tilde{R}-x_2^2}, x_2, \ell)}{\partial \ell} = &\frac{\sqrt{1+x_2^2}}{\ell(1+\tilde{R}^2) + \ell^3 - 2\ell^2\sqrt{\tilde{R}^2 - x_2^2}} \\ &- \frac{1}{\ell^2}\arccos\left(\frac{1+\tilde{R}^2 - \ell\sqrt{\tilde{R}^2-x_2^2}}{\sqrt{(1+\tilde{R}^2)(1+\tilde{R}^2+\ell^2-2\ell\sqrt{\tilde{R}^2-x_2^2})}}\right)\end{aligned}$$

And in order to prove that this is non-negative, we will prove that the same expression is non-negative, when multiplied by ℓ^2. We compute the partial derivative of the aforementioned expression with respect to ℓ:

$$\frac{\partial}{\partial \ell}\left(\frac{\partial \mathfrak{d}(\sqrt{\tilde{R}-x_2^2}, x_2, \ell)}{\partial \ell}\ell^2\right) = \frac{2\ell\sqrt{1+x_2^2}(\sqrt{\tilde{R}^2-x_2^2} - \ell)}{(1+\tilde{R}^2+\ell^2-2\ell\sqrt{\tilde{R}^2-x_2^2})^2} \geq 0.$$

And $\ell^2(\partial \mathfrak{d}(\sqrt{\tilde{R}-x_2^2}, x_2, \ell)/\partial \ell)$ evaluated at 0 is 0 for all choices of parameters R and x_2 in the domain. So we conclude that $\partial \mathfrak{d}(\sqrt{\tilde{R}-x_2^2}, x_2, \ell)/\partial \ell \geq 0$.

Thus, we can consider the limit when $\ell \to 0$ in order to lower bound $\mathfrak{d}(\cdot)$. In the defined domain, we have

$$\begin{aligned}\lim_{\ell\to 0} \mathfrak{d}(\sqrt{\tilde{R} - x_2}, x_2, \ell) &= \lim_{\ell\to 0} \frac{1}{\ell} \arccos\left(\frac{1 + \tilde{R}^2 - x\sqrt{\tilde{R}^2 - x_2^2}}{\sqrt{1+\tilde{R}^2}\sqrt{1 + x_2^2 + (\ell - \sqrt{\tilde{R}^2 - x_2^2})^2}}\right) \\ &\overset{①}{=} \lim_{\ell\to 0} \frac{\sqrt{1+x_2^2}}{1 + \tilde{R}^2 + \ell^2 - 2\ell\sqrt{\tilde{R}^2 - x_2^2}} \\ &= \frac{\sqrt{1+x_2^2}}{1+\tilde{R}^2}.\end{aligned}$$

We used L'Hôpital's rule for ①. Now, the right hand side of the previous expression is minimized at $x_2 = 0$ so we conclude that we have

$$\cos^2(R) = \frac{1}{1+\tan^2(R)} = \frac{1}{1+\tilde{R}^2} \leq \mathfrak{d}(x_1, x_2, \ell) = \frac{d(p,q)}{\|\tilde{p} - \tilde{q}\|}.$$

The upper bound uses again a similar argument. Assume that $\ell > x_1$ and define $\tilde{r}$ as above. We assume again without loss of generality that $x_1 \geq \ell - x_1$. Then

$$\frac{d(x,r) + d(r,y)}{\ell} \leq \frac{d(x,r)}{\ell - x_1} \iff \frac{d(r,y)}{x_1} \leq \frac{d(x,r)}{\ell - x_1}$$

and the latter is true by the monotonicity proved in step 1. Consequently we can just consider the points that satisfy $\ell \leq x_1$. By step 2, $\mathfrak{d}(\cdot)$ is maximal whenever x_1 is the lowest possible, so it is enough to consider $x_1 = \ell$. Using step 1 again, we obtain that the greatest value of $\mathfrak{d}(\cdot)$ can be bounded by the limit $\lim_{\ell\to 0} \mathfrak{d}(\ell, x_2, \ell)$ which using L'Hôpital's rule in ① and simplifying is

$$\begin{aligned}\mathfrak{d}(x_1, x_2, \ell) &\leq \lim_{\ell\to 0} \mathfrak{d}(\ell, x_2, \ell) \\ &= \lim_{\ell\to 0} \frac{1}{\ell} \arccos\left(\sqrt{\frac{1+x_2^2}{1+\ell^2+x_2^2}}\right) \\ &\overset{①}{=} \frac{1}{\sqrt{1+x_2^2}}.\end{aligned}$$

The expression is maximized at $x_2 = 0$ and evaluates to 1. □

C.3 Angle deformation

Lemma C.4. *Let $\mathcal{M} = \mathcal{H}$ or $\mathcal{M} = \mathcal{S}$ and $K \in \{1, -1\}$. Let $x, y \in \mathcal{M}$ be two different points and different from x_0. Let $\tilde{\alpha}$ be the angle $\angle x_0 x y$, formed by the vectors $x_0 - x$ and $y - x$. Let α be the corresponding angle between the vectors $\operatorname{Exp}_x^{-1}(x_0)$ and $\operatorname{Exp}_x^{-1}(y)$. The following holds:*

$$\sin(\alpha) = \sin(\tilde{\alpha})\sqrt{\frac{1 + K\|\tilde{x}\|^2}{1 + K\|\tilde{x}\|^2 \sin^2(\tilde{\alpha})}}, \qquad \cos(\alpha) = \cos(\tilde{\alpha})\sqrt{\frac{1}{1 + K\|\tilde{x}\|^2 \sin^2(\tilde{\alpha})}}.$$

Proof. Note that we can restrict ourselves to $\alpha \in [0, \pi]$ because we have $\widetilde{(-w)} = -\tilde{w}$ (recall our notation about vectors with tilde). This means that the result for the range $\alpha \in [-\pi, 0]$ can be deduced from the result for $-\alpha$.

We start with the case $K = -1$. We can assume without loss of generality that the dimension is $d = 2$, and that the coordinates of $\tilde{x}$ are $(0, x_2)$, for $x_2 \leq \tanh(R)$ that $\tilde{y} = (y_1, y_2)$, for some $y_1 \leq 0$ and $\tilde{\delta} \overset{\text{def}}{=} \angle \tilde{y}\tilde{x}_0\tilde{x} \in [0, \pi/2]$, since we can make the distance $\|\tilde{x} - \tilde{y}\|$ as small as we want. Recall $\tilde{x}_0 = \mathbf{0}$. We recall that $d(x, x_0) = \operatorname{arctanh}(\|\tilde{x}\|)$ and we note that $\sinh(\operatorname{arctanh}(t)) = \frac{t}{1-t^2}$, so that $\sinh(d(x, x_0)) = \|\tilde{x}\|/\sqrt{1 - \|\tilde{x}\|^2}$, for any $\tilde{x} \in \mathcal{B}$. We will apply the hyperbolic and

Euclidean law of sines Fact C.5 in order to compute the value of $\sin(\alpha)$ with respect to $\tilde{\alpha}$. Let $\tilde{a}$ and $\tilde{b}$ be points in the border of $\mathcal{B}$ such that the segment joining $\tilde{a}$ and $\tilde{b}$ is a chord that contains $\tilde{x}$ and $\tilde{y}$ and $\|\tilde{a}-\tilde{x}\| \leq \|\tilde{b}-\tilde{y}\|$. So $\|\tilde{a}-\tilde{x}\|$ and $\|\tilde{b}-\tilde{y}\|$ are $\sqrt{1-\|\tilde{x}\|^2\sin(\tilde{\alpha})} - d\cos(\tilde{\alpha})$ and $\sqrt{1-\|\tilde{x}\|^2\sin(\tilde{\alpha})} + d\cos(\tilde{\alpha})$, respectively. We have

$$
\begin{aligned}
\sin(\alpha) &\overset{①}{=} \frac{\sinh(d(x_0,y))\sin(\tilde{\delta})}{\sinh(d(x,y))} \\
&\overset{②}{=} \frac{\|\tilde{x}_0-\tilde{y}\|}{\sqrt{1-\|\tilde{x}_0-\tilde{y}\|^2}} \cdot \frac{\|\tilde{x}-\tilde{y}\|\sin(\tilde{\alpha})}{\|\tilde{x}_0-\tilde{y}\|} \cdot \frac{1}{\sinh(d(x,y))} \\
&\overset{③}{=} \frac{\sin(\tilde{\alpha})}{\sqrt{1-\|\tilde{x}\|^2+\|\tilde{x}-\tilde{y}\|(-2\|\tilde{x}\|\cos(\tilde{\alpha})+\|\tilde{x}-\tilde{y}\|)}} \cdot \frac{\|\tilde{x}-\tilde{y}\|}{\sinh(d(x,y))} \\
&\overset{④}{=} \frac{\sin(\tilde{\alpha})}{\sqrt{1-\|\tilde{x}\|^2}} \lim_{d(x,y)\to 0} \|\tilde{x}-\tilde{y}\| \frac{1}{\sinh(d(x,y))} \\
&\overset{⑤}{=} \frac{\sin(\tilde{\alpha})}{\sqrt{1-\|\tilde{x}\|^2}} \lim_{d(x,y)\to 0} \frac{(e^{2d(x,y)}-1)(\|\tilde{a}-\tilde{x}\|\cdot\|\tilde{b}-\tilde{x}\|)}{e^{2d(x,y)}\|\tilde{a}-\tilde{x}\|+\|\tilde{b}-\tilde{x}\|} \cdot \frac{2e^{d(x,y)}}{e^{2d(x,y)}-1} \\
&= \frac{\sin(\tilde{\alpha})}{\sqrt{1-\|\tilde{x}\|^2}} \cdot \frac{2\|\tilde{a}-\tilde{x}\|\cdot\|\tilde{b}-\tilde{x}\|}{\|\tilde{a}-\tilde{x}\|+\|\tilde{b}-\tilde{x}\|} \\
&\overset{⑥}{=} \frac{\sin(\tilde{\alpha})}{\sqrt{1-\|\tilde{x}\|^2}} \cdot \frac{2(1-\|\tilde{x}\|^2\sin^2(\tilde{\alpha})-\|\tilde{x}\|^2\cos^2(\tilde{\alpha}))}{2\sqrt{1-\|\tilde{x}\|^2\sin^2(\tilde{\alpha})}} \\
&= \sin(\tilde{\alpha})\sqrt{\frac{1-\|\tilde{x}\|^2}{1-\|\tilde{x}\|^2\sin^2(\tilde{\alpha})}}.
\end{aligned}
$$

In ① we used the hyperbolic sine theorem. In ② we used the expression above regarding segments that pass through the origin, and the Euclidean sine theorem. In ③, we simplify and use that the coordinates of $\tilde{y}$ are $(-\sin(\tilde{\alpha})\|\tilde{x}-\tilde{y}\|, \|\tilde{x}\|-\cos(\tilde{\alpha})\|\tilde{x}-\tilde{y}\|)$. Then, in ④, since $\sin(\alpha)$ does not depend on $\|\tilde{x}-\tilde{y}\|$, we can take the limit when $d(x,y)\to 0$, by which we mean we take the limit $\tilde{y}\to\tilde{x}$ by keeping the angle $\tilde{\alpha}$ constant. Since a posteriori the limit of each fraction exists, we compute them one at a time. ⑤ uses (32) and the definition of $\sinh(d(x,y))$. In ⑥ we substitute $\|\tilde{a}-\tilde{x}\|$ and $\|\tilde{b}-\tilde{x}\|$ by their values.

The spherical case is similar to the hyperbolic case. We also assume without loss of generality that the dimension is $d=2$. Define $\tilde{y}$ as a point such that $\angle\tilde{x_0}\tilde{x}\tilde{y}=\tilde{\alpha}$. We can assume without loss of generality that the coordinates of $\tilde{x}$ are $(0,x_2)$, that $\tilde{y}=(y_1,y_2)$, for $y_1\leq 0$, and $\tilde{\delta} \overset{\text{def}}{=} \angle\tilde{y}\tilde{x_0}\tilde{x}\in[0,\pi/2]$, since we can make the distance $\|\tilde{x}-\tilde{y}\|$ as small as we want. We recall that by (30) we have $d(x_0,x)=\arctan(\|\tilde{x}_0-\tilde{x}\|)$ and we note that $\sin(\arctan(t))=\frac{t}{1+t^2}$, so that $\sin(d(x_0,x))=\|\tilde{x}_0-\tilde{x}\|/\sqrt{1+\|\tilde{x}_0-\tilde{x}\|^2}$, for any $\tilde{x}\in\mathcal{B}$. We will apply the spherical and

Euclidean law of sines Fact C.5 in order to compute the value of $\sin(\alpha)$ with respect to $\tilde{\alpha}$. We have

$$\begin{aligned}
\sin(\alpha) &\overset{①}{=} \frac{\sin(d(x_0,y))\sin(\tilde{\delta})}{\sin(d(x,y))} \\
&\overset{②}{=} \frac{\|\tilde{x}_0-\tilde{y}\|}{\sqrt{1+\|\tilde{x}_0-\tilde{y}\|^2}} \cdot \frac{\|\tilde{x}-\tilde{y}\|\sin(\tilde{\alpha})}{\|\tilde{x}_0-\tilde{y}\|} \frac{1}{\sin(d(x,y))} \\
&\overset{③}{=} \frac{\sin(\tilde{\alpha})\|\tilde{x}-\tilde{y}\|}{\sqrt{1+\|\tilde{x}_0-\tilde{y}\|^2}\sqrt{1-\frac{(1-\|x\|\cos(\tilde{\alpha})\|\tilde{x}-\tilde{y}\|+\|\tilde{x}\|^2)^2}{(1+\|\tilde{x}\|^2)(1+\|\tilde{x}_0-\tilde{y}\|^2)}}} \\
&\overset{④}{=} \frac{\sin(\tilde{\alpha})\|\tilde{x}-\tilde{y}\|}{\sqrt{\|\tilde{x}-\tilde{y}\|^2(1+\|\tilde{x}\|^2-\|\tilde{x}\|^2\cos(\tilde{\alpha}))/(1+\|\tilde{x}\|^2)}} \\
&\overset{⑤}{=} \sin(\tilde{\alpha})\sqrt{\frac{1+\|\tilde{x}\|^2}{1+\|\tilde{x}\|^2\sin^2(\tilde{\alpha})}}.
\end{aligned}$$

In ① we used the spherical sine theorem. In ② we used the expression above regarding segments that pass through the origin, and the Euclidean sine theorem. In ③, we use the fact that the coordinates of $\tilde{y}$ are $(-\sin(\tilde{\alpha})\|\tilde{x}-\tilde{y}\|, d-\cos(\tilde{\alpha})\|\tilde{x}-\tilde{y}\|)$, use the distance formula (31) and the trigonometric inequality $\sin(\arccos(x)) = \sqrt{1-x^2}$. Then, in ④ and ⑤, we multiply and simplify.

Finally, in both cases, the cosine formula is derived from the identity $\sin^2(\alpha)+\cos^2(\alpha)=1$ after noticing that the sign of $\cos(\alpha)$ and the sign of $\cos(\tilde{\alpha})$ are the same. The latter fact can be seen to hold true by noticing that α is monotonous with respect to $\tilde{\alpha}$ and the fact that $\tilde{\alpha}=\pi/2$ implies $\sin(\alpha)=0$. □

Fact C.5 (Constant Curvature non-Euclidean Law of Sines). *Let $S_k(\cdot)$ denote the special sine, defined as $S_K(t)=\sin(\sqrt{K}t)$ if $K>0$, $S_K(t)=\sinh(\sqrt{-K}t)$ if $K<0$ and $S_k(t)=t$ if $K=0$. Let a,b,c be the lengths of the sides of a geodesic triangle defined in a manifold of constant sectional curvature. Let α,β,γ be the angles of the geodesic triangle, that are opposite to the sides a,b,c. The following holds:*

$$\frac{\sin(\alpha)}{S_K(a)} = \frac{\sin(\beta)}{S_K(b)} = \frac{\sin(\gamma)}{S_K(c)}.$$

We refer to Greenberg (1993) for a proof of this classical theorem.

C.4 Gradient deformation and smoothness of f

Lemma C.4, with $\tilde{\alpha}=\pi/2$, shows that $e_1 \perp e_j$, for $j\neq 1$. The rotational symmetry implies $e_i \perp e_j$ for $i\neq j$ and $i,j>1$. As in Lemma 2.1, let $x\in\mathcal{M}$ be a point and assume without loss of generality that $\tilde{x}\in\operatorname{span}\{\tilde{e}_1\}$ and $\nabla f(\tilde{x})\in\operatorname{span}\{\tilde{e}_1,\tilde{e}_2\}$. It can be assumed without loss of generality because of the symmetries. So we can assume the dimension is $d=2$. Using Lemma 2.1 we obtain that $\tilde{\alpha}=0$ implies $\alpha=0$. Also $\tilde{\alpha}=\pi/2$ implies $\alpha=\pi/2$, so the adjoint of the differential of h^{-1} at x, $(\mathrm{d}h^{-1})^*_x$ diagonalizes and has e_1 and e_2 as eigenvectors. We only need to compute the eigenvalues. The computation of the first one uses that the geodesic passing from x_0 and x can be parametrized as $h^{-1}(\tilde{x}_0+\arctan(\tilde{\lambda}\tilde{e}_1))$ if $K=1$ and $h^{-1}(\tilde{x}_0+\operatorname{arctanh}(\tilde{\lambda}\tilde{e}_1))$ if $K=-1$, by (29). The derivative of $\arctan(\cdot)$ or $\operatorname{arctanh}(\cdot)$ reveals that the first eigenvector, the one corresponding to e_1, is $1/(1+K\|\tilde{x}^2\|)$, i.e. $\nabla f(\tilde{x})_1 = \nabla F(x)_1/(1+K\|\tilde{x}^2\|)$. For the second one, let $x=(x_1,0)$ and $y=(y_1,y_2)$, with $y_1=x_1$ the second eigenvector results from the computation, for $K=-1$:

$$\begin{aligned}
\lim_{y_2\to 0}\frac{d(x,y)}{y_2} &= \lim_{y_2\to 0}\frac{1}{2y_2}\log\left(1+\frac{2y_2}{\sqrt{1-x_1^2}-y_2}\right) \\
&\overset{①}{=} \lim_{y_2\to 0}\frac{\frac{2}{\sqrt{1-x_1^2}-y_2}+\frac{2y_2}{(\sqrt{1-x_1^2}-y_2)^2}}{2+\frac{4y_2}{\sqrt{1-x_1^2}-y_2}} \\
&= \frac{1}{\sqrt{1-x_1^2}},
\end{aligned}$$

and for $K = 1$:

$$\lim_{y_2 \to 0} \frac{d(x,y)}{y_2} = \lim_{y_2 \to 0} \frac{1}{y_2} \arccos\left(\frac{\sqrt{1+x_1^2}}{\sqrt{1+x_1^2+y_2^2}}\right)$$
$$\overset{②}{=} \lim_{y_2 \to 0} \frac{\sqrt{1+x_1^2}}{1+x_1^2+y_2^2}$$
$$= \frac{1}{\sqrt{1+x_1^2}}.$$

So, since $x_1 = \|\tilde{x}\|$, we have $\nabla f(\tilde{x})_2 = \nabla F(x)_2 / \sqrt{1+K\|\tilde{x}\|^2}$ for $K \in \{1, -1\}$. We used L'Hôpital's rule in ① and ②.

Also note that if $v \in T_x\mathcal{M}$ is a vector normal to $\nabla F(x)$, then $\tilde{v}$ is normal to $\nabla f(x)$. It is easy to see this geometrically: Indeed, no matter how h changes the geometry, since it is a geodesic map, a geodesic in the direction of first-order constant increase of F is mapped via h to a geodesic in the direction of first-order constant increase of f. And the respective gradients must be perpendicular to all these directions. Alternatively, this can be seen algebraically. Suppose first $d = 2$, then v is proportional to $(\nabla F(x)_2, -\nabla F(x)_1) = (\sqrt{1+K\|\tilde{x}\|^2}\nabla f(\tilde{x})_2, -(1+K\|\tilde{x}\|^2)\nabla f(\tilde{x}_1))$. And a vector $\tilde{v}'$ normal to $\nabla f(x)$ must be proportional to $(-\nabla f(x)_2, \nabla f(x)_1)$. Let α be the angle formed by v and $-e_1$, $\tilde{\alpha}$ the corresponding angle formed between $\tilde{v}$ and $-\tilde{e}_1$, and $\tilde{\alpha}'$ the angle formed by $\tilde{v}'$ and $-\tilde{e}_1$. Then we have, using the expression for the vectors proportional to v and $\tilde{v}'$:

$$\sin(\alpha) = \frac{-f(x)_2}{\sqrt{\nabla f(x)_2^2 + (1+\|x\|^2)\nabla f(x)_1^2}} \text{ and } \sin(\tilde{\alpha}') = \frac{-f(x)_2}{\sqrt{\nabla f(x)_2^2 + \nabla f(x)_1^2}}$$

and an easy computation yields $\sin(\alpha) = \sin(\tilde{\alpha}')\sqrt{(1+K\|\tilde{x}^2\|)/(1+K\|\tilde{x}^2\|\sin^2(\tilde{\alpha}'))}$, which after applying Lemma C.4 we obtain $\sin(\tilde{\alpha}') = \sin(\tilde{\alpha})$ from which we conclude that $\tilde{\alpha}' = \tilde{\alpha}$ given that the angles are in the same quadrant. So $\tilde{v} \perp \nabla f(x)$. In order to prove this for $d \geq 3$ one can apply the reduction (42) to the case $d = 2$ that we obtain in the next section.

Combining the results obtained so far in Appendix C, we can prove Lemma 2.1. We continue by proving Lemma 2.3, which will generalize the computations we just performed, in order to analyze the Hessian of f and provide smoothness. Then, in the next section, we combine the results in Lemma 2.1 to prove Lemma 2.2.

Proof of Lemma 2.1. The lemma follows from Lemmas C.2, C.3, C.4 and the previous reasoning in this Section C.4. □

Proof of Lemma 2.3. We will compute the Hessian of $f = F \circ h^{-1}$ and we will bound its spectral norm for any point $\tilde{x} \in \mathcal{B}$. We can assume without loss of generality that $d = 2$ and $\tilde{x} = (\tilde{\ell}, 0)$, for $\tilde{\ell} > 0$ (the case $\tilde{\ell} = 0$ is trivial), since there is a rotational symmetry with e_1 as axis. This means that by rotating we could align the top eigenvector of the Hessian at a point so that it is in $\operatorname{span}\{e_1, e_2\}$. Let $\tilde{y} = (y_1, y_2) \in \mathcal{B}$ be another point, with $y_1 = \tilde{\ell}$. We can also assume that $y_2 > 0$ without loss of generality, because of our symmetry. Our approach will be the following. We know by Lemma C.4 and by the beginning of this section C.4 that the adjoint of the differential of h^{-1} at y, $(\mathrm{d}h^{-1})_y^*$ has $\operatorname{Exp}_y^{-1}(x_0)$ and a normal vector to it as eigenvectors. Their corresponding eigenvalues are $1/(1+K\|\tilde{y}\|^2)$ and $1/\sqrt{1+K\|\tilde{y}\|^2}$, respectively. Consider the basis of $T_x\mathcal{M}$ $\{e_1, e_2\}$ as defined at the beginning of this section, i.e. where e_1 is a unit vector proportional to $-\operatorname{Exp}_x^{-1}(x_0)$ and e_2 is the normal vector to e_1 that makes the basis orthonormal. Consider this basis being transported to y using parallel transport and denote the result $\{v_y, v_y^\top\}$. Assume we have the gradient $\nabla F(y)$ written in this basis. Then we can compute the gradient of f at y by applying $(\mathrm{d}h^{-1})_y^*$. In order to do that, we compose the change of basis from $\{v_y, v_y^\top\}$ to the basis of eigenvectors of $(\mathrm{d}h^{-1})_y^*$, then we apply a diagonal transformation given by the eigenvalues and finally we change the basis to $\{\tilde{e}_1, \tilde{e}_2\}$. Once this is done, we can differentiate with respect to y_2 in order to compute a column of the Hessian. Let $\tilde{\alpha}$ be the angle formed by the vectors $\tilde{y}$ and $\tilde{x}$. Note that $\tilde{\alpha} = \arctan(y_2/y_1)$. Let $\tilde{\gamma}$ be the angle formed by the vectors $(\tilde{y} - \tilde{x})$ and $-\tilde{y}$. That is, the angle $\tilde{\gamma} = \pi - \angle\tilde{x}\tilde{y}\tilde{x}_0$. Since $v_y^\top$

is the parallel transport of $e_2^\top$, the angle between $v_y^\top$ and the vector $\mathrm{Exp}_y^{-1}(x_0)$ is γ. Note we use the same convention as before for the angles, i.e. γ is the corresponding angle to $\tilde{\gamma}$, meaning that if γ is the angle between two intersecting geodesics in $\mathcal{M}$, then $\tilde{\gamma}$ is the angle between the respective corresponding geodesics in $\mathcal{B}$. Note the first change of basis is a rotation and that the angle of rotation is $\gamma - \pi/2$. The last change of basis is a rotation with angle equal to the angle formed by a vector $\tilde{v}$ normal to $-\tilde{y}$ ($\tilde{v}$ is the one such that $-\tilde{y} \times \tilde{v} > 0$) and the vector $\tilde{e}_2$. It is easy to see that this vector is equal to $\tilde{\alpha}$. So we have

$$\nabla f(y) = \begin{pmatrix} \cos(\tilde{\alpha}) & -\sin(\tilde{\alpha}) \\ \sin(\tilde{\alpha}) & \cos(\tilde{\alpha}) \end{pmatrix} \begin{pmatrix} \frac{1}{1+K(y_1^2+y_2^2)} & 0 \\ 0 & \frac{1}{\sqrt{1+K(y_1^2+y_2^2)}} \end{pmatrix} \begin{pmatrix} \sin(\gamma) & -\cos(\gamma) \\ \cos(\gamma) & \sin(\gamma) \end{pmatrix} \nabla F(y) \tag{36}$$

We want to take the derivative of this expression with respect to y_2 and we want to evaluate it at $y_2 = 0$. Let the matrices above be A, B and C so that $\nabla f(y) = ABC\nabla F(y)$. Using Lemma C.4 we have

$$\begin{aligned} \sin(\gamma) &= \sin(\tilde{\gamma})\sqrt{\frac{1+K(y_1^2+y_2^2)}{1+K(y_1^2+y_2^2)\sin^2(\tilde{\gamma})}} \overset{①}{=} \cos(\tilde{\alpha})\sqrt{\frac{1+K(y_1^2+y_2^2)}{1+K(y_1^2+y_2^2)\cos^2(\tilde{\alpha})}}, \\ \cos(\gamma) &= -\sin(\tilde{\alpha})\sqrt{\frac{1}{1+K(y_1^2+y_2^2)\cos^2(\tilde{\alpha})}}, \end{aligned} \tag{37}$$

where ① follows from $\sin(\tilde{\gamma}) = \sin(\tilde{\alpha} + \pi/2) = \cos(\tilde{\alpha})$. Now we can easily compute some quantities

$$\begin{aligned} A|_{y_2=0} = I,\ B|_{y_2=0} &= \begin{pmatrix} \frac{1}{1+Ky_1^2} & 0 \\ 0 & \frac{1}{\sqrt{1+Ky_1^2}} \end{pmatrix},\ C|_{y_2=0} = I, \\ \left.\frac{\partial A}{\partial y_2}\right|_{y_2=0} &= \left.\frac{\partial \tilde{\alpha}}{\partial y_2}\right|_{y_2=0} \cdot \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \overset{①}{=} \begin{pmatrix} 0 & \frac{-1}{y_1} \\ \frac{1}{y_1} & 0 \end{pmatrix}, \\ \left.\frac{\partial B}{\partial y_2}\right|_{y_2=0} &= \left.\begin{pmatrix} \frac{2Ky_2}{(1+K(y_1^2+y_2^2))^2} & 0 \\ 0 & \frac{2Ky_2}{2(1+K(y_1^2+y_2^2))^{3/2}} \end{pmatrix}\right|_{y_2=0} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \\ \left.\frac{\partial C}{\partial y_2}\right|_{y_2=0} &\overset{②}{=} \begin{pmatrix} 0 & \frac{1}{y_1\sqrt{1+Ky_1^2}} \\ \frac{-1}{y_1\sqrt{1+Ky_1^2}} & 0 \end{pmatrix}. \end{aligned}$$

Equalities ① and ② follow after using (37), $\tilde{\alpha} = \arctan(\frac{y_2}{y_1})$ and taking derivatives. Now we differentiate (36) with respect to y_2 and evaluate to $y_2 = 0$ using the chain rule. The result is

$$\begin{aligned} \begin{pmatrix} \nabla^2 f(\tilde{x})_{12} \\ \nabla^2 f(\tilde{x})_{22} \end{pmatrix} &= \left(\frac{\partial A}{\partial y_2}BC\nabla F(x) + A\frac{\partial B}{\partial y_2}C\nabla F(x) + AB\frac{\partial C}{\partial y_2}\nabla F(x) + ABC\frac{\partial \nabla F(x)}{\partial y_2}\right)\Bigg|_{y_2=0} \\ &= \begin{pmatrix} \frac{-\nabla f(\tilde{x})_2}{y_1\sqrt{1+Ky_1^2}} \\ \frac{\nabla f(\tilde{x})_1}{y_1(1+Ky_1^2)} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \frac{\nabla f(\tilde{x})_2}{y_1(1+Ky_1^2)^{3/2}} \\ \frac{-\nabla f(\tilde{x})_1}{y_1(1+Ky_1^2)} \end{pmatrix} + \begin{pmatrix} \frac{\nabla^2 F(x)_{12}}{(1+Ky_1^2)^{3/2}} \\ \frac{\nabla^2 F(x)_{22}}{1+Ky_1^2} \end{pmatrix} \end{aligned}$$

Computing the other column of the Hessian is easier. We can just consider (36) with $\tilde{\alpha} = 0$ and $\gamma = \pi/2$ and vary y_1. Taking derivatives with respect to y_1 gives

$$\begin{pmatrix} \nabla^2 f(\tilde{x})_{11} \\ \nabla^2 f(\tilde{x})_{21} \end{pmatrix} = \begin{pmatrix} \frac{-2Ky_1\nabla f(\tilde{x})_1}{(1+Ky_1^2)^2} \\ \frac{-Ky_1\nabla f(\tilde{x})_2}{(1+Ky_1^2)^{3/2}} \end{pmatrix} + \begin{pmatrix} \frac{\nabla^2 F(x)_{11}}{(1+Ky_1^2)^2} \\ \frac{\nabla^2 F(x)_{21}}{(1+Ky_1^2)^{3/2}} \end{pmatrix}.$$

Note in the computations of both of the columns of the Hessian we have used

$$\frac{\partial \nabla F(y)_i}{\partial y_1} = \nabla F(x)_{i1} \cdot \frac{1}{1+Ky_1^2} \quad \text{and} \quad \left.\frac{\partial \nabla F(y)_i}{\partial y_2}\right|_{y_2=0} = \nabla F(x)_{i2} \cdot \frac{1}{\sqrt{1+Ky_1^2}},$$

for $i = 1, 2$. The eigenvalues of the adjoint of the differential of h^{-1} appear as a factor because we are differentiating with respect to the geodesic in $\mathcal{B}$ which moves at a different speed than the

corresponding geodesic in $\mathcal{M}$. Note as well, as a sanity check, that the cross derivatives are equal, since

$$-\frac{1}{y_1\sqrt{1+Ky_1^2}}+\frac{1}{y_1(1+Ky_1^2)^{3/2}}=\frac{1}{y_1\sqrt{1+Ky_1^2}}\left(-1+\frac{1}{1+Ky_1^2}\right)=\frac{-Ky_1}{(1-y_1^2)^{3/2}}.$$

Finally, we bound the new smoothness constant $\tilde{L}$ by bounding the spectral norm of this Hessian. First note that using $y_1=\tilde{\ell}$ we have that $\frac{1}{\sqrt{1+K\tilde{\ell}^2}}=C_K(\ell)$ and then for $K=-1$ it is $\tilde{\ell}=\tanh(\ell)$ and for $K=1$ it is $\tilde{\ell}=\tan(\ell)$, where $\ell=d(x,x_0)<R$. We have that since there is a point $x^*\in\mathcal{M}$ such that $\nabla F(x^*)=0$ and F is L-smooth, then it is $\|\nabla F(x)\|\leq 2LR$. Similarly, by L-smoothness $|\nabla^2 F(x)_{ij}|\leq L$. We are now ready prove smoothness:

$$\begin{aligned}\tilde{L}^2&\leq\|\nabla^2 f(\tilde{x})\|_2^2\\&\leq\|\nabla^2 f(\tilde{x})\|_F^2=\nabla^2 f(\tilde{x})_{11}+2\nabla^2 f(\tilde{x})_{12}+\nabla^2 f(\tilde{x})_{22}\\&\leq L^2([C_K^4(R)+4RS_K(R)C_K^3(R)]^2+2[C_K^3(R)+2RS_K(R)C_K^2(R)]^2+C_K^4(R))\end{aligned}$$

and this can be bounded by $44L^2\max\{1,R^2\}$ if $K=1$ and $44L^2\max\{1,R^2\}C_K^8(R)$ if $K=-1$. In any case, it is $O(L^2)$. □

C.5 Proof of Lemma 2.2

Proof of Lemma 2.2. Assume for the moment the dimension is $d=2$. We can assume without loss of generality that $\tilde{x}=(\tilde{\ell},0)$. We are given two vectors, that are the gradients $\nabla F(x)$, $\nabla f(\tilde{x})$ and a vector $w\in T_x\mathcal{M}$. Let $\tilde{\delta}$ be the angle between $\tilde{w}$ and $-\tilde{x}$. Let δ be the corresponding angle, i.e. the angle between w and $u\overset{\text{def}}{=}\operatorname{Exp}_x^{-1}(x_0)$. Let α be the angle in between $\nabla F(x)$ and u. Let $\tilde{\beta}$ be the angle in between $\nabla f(\tilde{x})$ and $-x$. $\tilde{\alpha}$ and β are defined similarly. We claim

$$\frac{\langle\frac{\nabla F(x)}{\|\nabla F(x)\|},\frac{w}{\|w\|}\rangle}{\langle\frac{\nabla f(\tilde{x})}{\|\nabla f(\tilde{x})\|},\frac{\tilde{w}}{\|\tilde{w}\|}\rangle}=\sqrt{\frac{1+K\tilde{\ell}^2}{(1+K\tilde{\ell}^2\sin^2(\tilde{\delta}))(1+K\tilde{\ell}^2\cos^2(\tilde{\beta}))}}.\tag{38}$$

Let's see how to arrive to this expression. By Lemma 2.1.c) we have

$$\tan(\alpha)=\frac{\tan(\tilde{\beta})}{\sqrt{1+K\tilde{\ell}^2}}.\tag{39}$$

From this relationship we can deduce

$$\cos(\alpha)=\cos(\tilde{\beta})\sqrt{\frac{1+K\tilde{\ell}^2}{1+K\tilde{\ell}^2\cos^2(\tilde{\beta})}}.\tag{40}$$

This comes from squaring (39), reorganizing terms and noting that $\operatorname{sign}(\cos(\alpha))=\operatorname{sign}(\cos(\tilde{\beta}))$ which is implied by Lemma 2.1.c). We are now ready to prove the claim (38) (for $d=2$). We have

$$\begin{aligned}\frac{\langle\frac{\nabla F(x)}{\|\nabla F(x)\|},\frac{w}{\|w\|}\rangle}{\langle\frac{\nabla f(\tilde{x})}{\|\nabla f(\tilde{x})\|},\frac{\tilde{w}}{\|\tilde{w}\|}\rangle}&=\frac{\cos(\alpha-\delta)}{\cos(\tilde{\beta}-\tilde{\delta})}\\&\overset{②}{=}\frac{\cos(\delta)+\tan(\alpha)\sin(\delta)}{\cos(\tilde{\beta})\cos(\tilde{\delta})+\sin(\tilde{\beta})\sin(\tilde{\delta})}\cos(\alpha)\\&\overset{③}{=}\frac{\frac{\cos(\tilde{\delta})}{\sqrt{1+K\tilde{\ell}^2\sin^2(\tilde{\delta})}}+\frac{\tan(\tilde{\beta})}{\sqrt{1+K\tilde{\ell}^2}}\frac{\sin(\tilde{\delta})\sqrt{1+K\tilde{\ell}^2}}{\sqrt{1+K\tilde{\ell}^2\sin^2(\tilde{\delta})}}}{\cos(\tilde{\beta})\cos(\tilde{\delta})+\sin(\tilde{\beta})\sin(\tilde{\delta})}\cos(\tilde{\beta})\sqrt{\frac{1+K\tilde{\ell}^2}{1+K\tilde{\ell}^2\cos^2(\tilde{\beta})}}\\&\overset{④}{=}\sqrt{\frac{1+K\tilde{\ell}^2}{(1+K\tilde{\ell}^2\sin^2(\tilde{\delta}))(1+K\tilde{\ell}^2\cos^2(\tilde{\beta}))}}.\end{aligned}$$

Equality ① follows by the definition of $\alpha, \delta, \tilde{\delta}$ and $\tilde{\beta}$. In ②, we used trigonometric identities. In ③ we used Lemma C.4, (39) and (40). By reordering the expression, the denominator cancels out with a factor of the numerator in ④.

In order to work with arbitrary dimension, we note it is enough to prove it for $d = 3$, since in order to bound

$$\frac{\langle \frac{\nabla F(x)}{\|\nabla F(x)\|}, \frac{v}{\|v\|} \rangle}{\langle \frac{\nabla f(\tilde{x})}{\|\nabla f(\tilde{x})\|}, \frac{\tilde{v}}{\|\tilde{v}\|} \rangle},$$

it is enough to consider the following submanifold

$$\mathcal{M}' \stackrel{\text{def}}{=} \operatorname{Exp}_x(\operatorname{span}\{v, \operatorname{Exp}_x^{-1}(x_0), \nabla F(x)\}).$$

for an arbitrary vector $v \in T_x\mathcal{M}$ and a point x defined as above. The case $d = 3$ can be further reduced to the case $d = 2$ in the following way. Suppose $\mathcal{M}'$ is a three dimensional manifold (if it is one or two dimensional there is nothing to do). Define the following orthonormal basis of $T_x\mathcal{M}$, $\{e_1, e_2, e_3\}$ where $e_1 = -\operatorname{Exp}_x^{-1}(x_0)/\|\operatorname{Exp}_x^{-1}(x_0)\|$, e_2 is a unit vector, normal to e_1 such that $e_2 \in \operatorname{span}\{e_1, \nabla F(x)\}$. And e_3 is a vector that completes the orthonormal basis. In this basis, let v be parametrized by $\|v\|(\sin(\delta), \cos(\nu)\cos(\delta), \sin(\nu)\cos(\delta))$, where δ can be thought as the angle between the vector v and its projection onto the plane $\operatorname{span}\{e_2, e_3\}$ and ν can be thought as the angle between this projection and its projection onto e_2. Similarly we parametrize $\tilde{v}$ by $\|\tilde{v}\|(\sin(\tilde{\delta}), \cos(\tilde{\nu})\cos(\tilde{\delta}), \sin(\tilde{\nu})\cos(\tilde{\delta}))$, where the base used is the analogous base to the previous one, i.e. The vectors $\{\tilde{e}_1, \tilde{e}_2, \tilde{e}_3\}$. Taking into account that $e_2 \perp e_1$, $e_3 \perp e_1$, $\tilde{e}_2 \perp \tilde{e}_1$, $\tilde{e}_3 \perp \tilde{e}_1$, and the fact that e_1 is parallel to $-\operatorname{Exp}_x(x_0)$, by the radial symmetry of the geodesic map we have that $\nu = \tilde{\nu}$. Also, by looking at the submanifold $\operatorname{Exp}_x(\operatorname{span}\{e_1, v\})$ and using Lemma C.4 we have

$$\sin(\delta) = \sin(\tilde{\delta})\sqrt{\frac{1 + K\tilde{\ell}^2}{1 + K\tilde{\ell}^2 \sin(\tilde{\delta})}}.$$

If we want to compare $\langle \nabla F(x), v \rangle$ with $\langle \nabla f(\tilde{x}), \tilde{v} \rangle$ we should be able to just zero out the third components of v and $\tilde{v}$ and work in $d = 2$. But in order to completely obtain a reduction to the two-dimensional case we studied a few paragraphs above, we would need to prove that if we call $w \stackrel{\text{def}}{=} (\sin(\delta), \cos(\nu)\cos(\delta), 0)$ the vector v with the third component made 0, then $\tilde{w}$ is in the same direction of the vector $\tilde{v}$, when the third component is made 0. The norm of these two vectors will not be the same, however. Let $\tilde{w}' = (\sin(\tilde{\delta}), \cos(\nu)\cos(\tilde{\delta}), 0)$ be the vector $\tilde{v}$ when the third component is made 0. Then

$$\|w\| = \|v\|\sqrt{\sin^2(\delta) + \cos^2(\delta)\cos^2(\nu)} \text{ and } \|\tilde{w}'\| = \|\tilde{v}\|\sqrt{\sin^2(\tilde{\delta}) + \cos^2(\tilde{\delta})\cos^2(\nu)}. \tag{41}$$

But indeed, we claim

$$\tilde{w} \text{ and } \tilde{w}' \text{ have the same direction.} \tag{42}$$

This is easy to see geometrically: since we are working with a geodesic map, the submanifolds $\operatorname{Exp}_x(\operatorname{span}\{v, e_3\})$ and $\operatorname{Exp}_x(\operatorname{span}\{e_1, e_2\})$ contain w. Similarly the submanifolds $x + \operatorname{span}\{\tilde{v}, \tilde{e}_3\}$ and $x + \operatorname{span}\{\tilde{e}_1, \tilde{e}_2\}$ contain $\tilde{w}'$. If the intersections of each of these pair of manifolds is a geodesic then the geodesic map must map one intersection to the other one, implying $\tilde{w}$ is proportional to $\tilde{w}'$. If the intersections are degenerate the case is trivial. Alternatively, one can prove this fact algebraically after some computations. If we call δ^* and $\tilde{\delta}'$ the angles formed by, respectively, the vectors e_2 and w, and the vectors $\tilde{e}_2$ and $\tilde{w}'$, then we have $\tilde{w}'$ is proportional to $\tilde{w}$ iff $\tilde{\delta}' = \tilde{\delta}^*$, or equivalently $\delta' = \delta^*$. Using the definitions of w and $\tilde{w}'$ we have

$$\begin{aligned}\sin(\delta^*) = \sin\left(\arctan\left(\frac{\sin(\delta)}{\cos(\nu)\cos(\delta)}\right)\right) &= \frac{\tan(\delta)/\cos(\nu)}{(\tan(\delta)/\cos(\nu))^2 + 1} \\ &= \frac{\sin(\delta)}{\sqrt{\sin^2(\delta) + \cos^2(\nu)\cos^2(\delta)}},\end{aligned}$$

and analogously

$$\begin{aligned}\sin(\tilde{\delta}') = \sin\left(\arctan\left(\frac{\sin(\tilde{\delta})}{\cos(\nu)\cos(\tilde{\delta})}\right)\right) &= \frac{\tan(\tilde{\delta})/\cos(\nu)}{(\tan(\tilde{\delta})/\cos(\nu))^2+1} \\ &= \frac{\sin(\tilde{\delta})}{\sqrt{\sin^2(\tilde{\delta})+\cos^2(\nu)\cos^2(\tilde{\delta})}}.\end{aligned} \tag{43}$$

Using Lemma C.4 for the pairs δ', $\tilde{\delta}'$ and δ^*, $\tilde{\delta}^*$, and the equations above we obtain

$$\sin(\delta^*) = \frac{\sin(\tilde{\delta})\sqrt{\frac{1+K\tilde{\ell}^2}{1+K\tilde{\ell}^2\sin^2(\tilde{\delta})}}}{\sqrt{\sin^2(\tilde{\delta})\frac{1+K\tilde{\ell}^2}{1+K\tilde{\ell}^2\sin^2(\tilde{\delta})}+\cos^2(\nu)\frac{\cos^2(\tilde{\delta})}{1+K\tilde{\ell}^2\sin^2(\tilde{\delta})}}} = \frac{\sin(\tilde{\delta})\sqrt{1+K\tilde{\ell}^2}}{\sqrt{\sin^2(\tilde{\delta})(1+K\tilde{\ell}^2)+\cos^2(\nu)\cos^2(\tilde{\delta})}},$$

and

$$\sin(\delta') = \frac{\sin(\tilde{\delta})}{\sqrt{\sin^2(\tilde{\delta})+\cos^2(\nu)\cos^2(\tilde{\delta})}}\sqrt{\frac{1+K\tilde{\ell}^2}{1+K\tilde{\ell}^2\left(\frac{\sin^2(\tilde{\delta})}{\sin^2(\tilde{\delta})+\cos^2(\nu)\cos^2(\tilde{\delta})}\right)}},$$

The two expressions on the right hand side are equal. This implies $\sin(\delta') = \sin(\delta^*)$. Since the angles were in the same quadrant we have $\delta' = \delta^*$ by checking in which sectors the angles must be.

We can now come back to the study of $\frac{\langle \nabla F(x), v\rangle}{\langle \nabla f(\tilde{x}), \tilde{v}\rangle}$. By (41) we have

$$\frac{\langle \nabla F(x), v\rangle}{\langle \nabla f(\tilde{x}), \tilde{v}\rangle} = \frac{\|\nabla F(x)\|}{\|\nabla f(\tilde{x})\|}\frac{\|v\|}{\|\tilde{v}\|}\frac{\langle \frac{\nabla F(x)}{\|\nabla F(x)\|}, \frac{w}{\|w\|}\rangle}{\langle \frac{\nabla f(\tilde{x})}{\|\nabla f(\tilde{x})\|}, \frac{\tilde{w}}{\|\tilde{w}\|}\rangle}\frac{\sqrt{\sin^2(\delta)+\cos^2(\delta)\cos^2(\nu)}}{\sqrt{\sin^2(\tilde{\delta})+\cos^2(\tilde{\delta})\cos^2(\nu)}}$$

The last two factors can be simplified. Using (38) and (41) we get that this product is equal to

$$\sqrt{\frac{1+K\tilde{\ell}^2}{(1+K\tilde{\ell}^2\sin^2(\tilde{\delta}^*))(1+K\tilde{\ell}^2\cos^2(\tilde{\beta}))}}\frac{\sqrt{\sin^2(\tilde{\delta})\frac{1+K\tilde{\ell}^2}{(1+K\tilde{\ell}^2\sin^2(\tilde{\delta}))}+\cos^2(\nu)\frac{\cos^2(\tilde{\delta})}{1+K\tilde{\ell}^2\sin(\tilde{\delta})}}}{\sin^2(\tilde{\delta})+\cos^2(\tilde{\delta})\cos^2(\nu)}$$

which after using (43) (recall $\tilde{\delta}^* = \tilde{\delta}'$), and simplifying it yields

$$\sqrt{\frac{1+K\tilde{\ell}^2}{(1+K\tilde{\ell}^2\sin^2(\tilde{\delta}))(1+K\tilde{\ell}^2\cos^2(\tilde{\beta}))}}.$$

So finally we have

$$\frac{\langle \nabla F(x), v\rangle}{\langle \nabla f(\tilde{x}), \tilde{v}\rangle} = \frac{\|\nabla F(x)\|}{\|\nabla f(\tilde{x})\|}\frac{\|v\|}{\|\tilde{v}\|}\sqrt{\frac{1+K\tilde{\ell}^2}{(1+K\tilde{\ell}^2\sin^2(\tilde{\delta}))(1+K\tilde{\ell}^2\cos^2(\tilde{\beta}))}}.$$

We use now Lemma 2.1.a) and Lemma 2.1.c), and bound $\sin^2(\tilde{\delta})$ and $\cos^2(\tilde{\beta})$ in order to bound the previous expression. Recall that, by (30) we have $1/\sqrt{1+K\tilde{\ell}^2} = C_K(\ell)$, for $\ell = d(x, x_0) \leq R$. Let's proceed. We obtain, for $K = -1$

$$\cosh^{-3}(R) \leq \frac{1}{\cosh^2(\ell)}\cdot 1 \cdot \frac{1}{\cosh(\ell)} \leq \frac{\langle \nabla F(x), v\rangle}{\langle \nabla f(\tilde{x}), \tilde{v}\rangle} \leq \frac{1}{\cosh(\ell)}\cdot\cosh^2(\ell)\cdot\cosh(\ell) \leq \cosh^2(R).$$

and for $K = 1$ it is

$$\cos^2(R) \leq \frac{1}{\cos(\ell)}\cdot\cos^2(\ell)\cdot\cos(\ell) \leq \frac{\langle \nabla F(x), v\rangle}{\langle \nabla f(\tilde{x}), \tilde{v}\rangle} \leq \frac{1}{\cos^2(\ell)}\cdot 1\cdot\frac{1}{\cos(\ell)} \leq \cos^{-3}(R).$$

The first part of Lemma 2.2 follows, for $\gamma_{\mathrm{p}} = \cosh^{-3}(R)$ and $\gamma_{\mathrm{n}} = \cosh^{-2}(R)$ when $K = -1$, and $\gamma_{\mathrm{p}} = \cos^2(R)$ and $\gamma_{\mathrm{n}} = \cos^3(R)$ when $K = 1$.

The second part of Lemma 2.2 follows readily from the first one and g-convexity of F, as in the following. It holds

$$f(\tilde{x}) + \frac{1}{\gamma_{\mathrm{n}}}\langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x}\rangle \overset{①}{\leq} F(x) + \langle \nabla F(x), y - x\rangle \overset{②}{\leq} F(y) = f(\tilde{y}),$$

and

$$f(\tilde{x}) + \gamma_{\mathrm{p}}\langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x}\rangle \overset{③}{\leq} F(x) + \langle \nabla F(x), y - x\rangle \overset{④}{\leq} F(y) = f(\tilde{y}),$$

where ① and ③ hold if $\langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x}\rangle \leq 0$ and $\langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x}\rangle \geq 0$, respectively. Inequalities ② and ④ hold by g-convexity of F.

□