

788 Appendix

789	A Preliminaries	21
790	B Deferred proofs	31
791	C Auxiliary lemmas	35
792	D Additional discussions	36
793	E Implementation detail and additional experiments	43

794 A Preliminaries

795 A.1 Riemannian geometry

796 In this appendix, we introduce key concepts in Riemannian geometry briefly discussed in Section 2.
 797 We mainly mention the known results, and omit the proof and well-definedness of definitions. For
 798 detail, interested reader can find relevant material in textbooks, e.g., [Lee12, Lee18, Bou23].

799 A Riemannian manifold is a manifold equipped with an inner product for each tangent space, called a
 800 Riemannian metric.

801 **Definition A.1** (Riemannian manifold). *A Riemannian manifold (M, g) is a real smooth manifold*
 802 *equipped with a Riemannian metric g which assigns to each $p \in M$ a positive-definite inner product*
 803 *$g_p(v, w) = \langle v, w \rangle_p$ on the tangent space $T_p M$.*

804 Often, this tangent space $T_p M$ is conveniently expressed in the form of the vector field, which takes
 805 a point in a manifold as an input and returns a tangent space vector at that point. Formally, the vector
 806 field of M is defined as follows:

807 **Definition A.2** (Vector field). *A map $X : C^\infty(M) \rightarrow C^\infty(M)$ is called a smooth vector field if it is*
 808 *a derivation, i.e., X satisfies*

$$X.(fg) = X.(f)g(\cdot) + f(\cdot)X.(g).$$

809 *Here $\cdot \in M$ is the input of the function.*

810 As the name derivation indicates, one can think of the vector field as a directional derivative along the
 811 direction of the vector field. The following familiar example may help.

812 **Example A.3** (Vector field in \mathbb{R}^d). *For $f \in C^\infty(\mathbb{R}^d)$, $p \in \mathbb{R}^d$, and $v \in \mathbb{R}^d$, think of a directional*
 813 *derivative of f at p along direction v , $d_v f(p)$. If we fix p and view f as a variable input, then $v \in T_p M$*
 814 *can be identified with the functional $f \mapsto d_v f(p)$. In other words, by defining $X_p(f) := d_v f(p)$, the*
 815 *value of vector field X_p at each point $p \in M$ can be identified as a tangent vector $v \in T_p \mathbb{R}^d$.*

816 From now on, we will write X as a vector field, and this will mean a function $X_p(f) = d_v f(p)$
 817 where $v = X_p$. For the definition of a directional derivative in general manifolds, we refer to [Lee12].
 818 We write $\mathfrak{X}(M)$ as a set of smooth vector fields on M .

819 One of the fundamental structure of a manifold is an affine connection, a concept that connects
 820 tangent spaces of different points of the manifold.

821 **Definition A.4** (Affine connection). *Let M be a manifold, and $\mathfrak{X}(M)$ be the set of all smooth vector*
 822 *fields on M . An operator $\nabla \cdot : \mathfrak{X}(M) \times \mathfrak{X}(M) \rightarrow \mathfrak{X}(M)$ is called an affine connection if for all*
 823 *$f \in C^\infty(M)$ and $X, Y \in \mathfrak{X}(M)$ it satisfies the following properties:*

824 1. $\nabla_{fX} Y = f \nabla_X Y$, i.e. linear in the first variable.

825 2. $\nabla_X(fY) = (d_X f)Y + f\nabla_X Y$, that is, ∇ satisfies the Leibniz rule in the second variable.

826 In the case of Riemannian manifolds, we have a natural connection induced from the Riemannian
827 metric, called Levi-Civita connection.

828 **Definition A.5** (Levi-Civita connection). *For a Riemannian manifold (M, g) , let $\mathfrak{X}(M)$ be a set*
829 *of smooth vector field on M . The Levi-Civita connection is the unique affine connection $\nabla \cdot : \mathfrak{X}(M) \times \mathfrak{X}(M) \rightarrow \mathfrak{X}(M)$, satisfying the following properties:*

831 1. $\nabla_X Y - \nabla_Y X = [X, Y]$, i.e. it is torsion-free. Here, $[\cdot, \cdot]$ denotes a Lie bracket.

832 2. $X(g(Y, Z)) = g(\nabla_X Y, Z) + g(Y, \nabla_X Z)$, that is, the connection is compatible with the
833 metric g .

834 The choice of the affine connection determines multiple geometric concepts. One fundamental
835 concept is geodesic curve, which is a constant speed curve on the manifold.

836 **Definition A.6** (Geodesic). *A smooth curve $\gamma : [0, 1] \rightarrow M$ is called a geodesic curve if $\nabla_{\dot{\gamma}} \dot{\gamma} = 0$.*

837 A Riemannian manifold is called *complete* if any two points are connected by some geodesic. We
838 will always assume M is a complete Riemannian manifold.

839 We say a Riemannian submanifold $\widetilde{M} \subseteq M$ is *totally geodesic* if for every $v \in T\widetilde{M}$, the geodesic
840 with respect to \widetilde{M} , γ_v , lies entirely in M .

841 Equipped with the notion of geodesic, one can define the exponential map and logarithmic map on a
842 Riemannian manifold.

843 **Definition A.7** (Exponential map, logarithmic map). *Let $p \in M$.*

844 1. *For any $v \in T_p M$, one can define a geodesic curve $\gamma_v : [0, 1] \rightarrow M$ such that $\gamma_v(0) = p$*
845 *and $\gamma'_v(0) = v$. Then, one can define a map $\exp_p(v) := \gamma_v(1)$. This map is called the*
846 *exponential map.*

847 2. *It is known that the exponential map is a local diffeomorphism on U , the open neighborhood*
848 *of $0 \in T_p M$. Therefore, one can define $\log_p q := \exp_p^{-1}(q)$ for $q \in \exp_p(U)$. This map is*
849 *called the logarithmic map.*

850 To understand the notions of the exponential map and logarithmic map, we illustrate these concepts in
851 the Euclidean case. In the Euclidean space, $\exp_p(v) = p + v$ and $\log_p q = q - p$. In other words, the
852 exponential map moves p along the tangent direction v , and the logarithmic map returns the tangent
853 direction from p to q .

854 Note the logarithmic map is only defined locally. While our analysis assumed the global existence
855 of the logarithmic map over the geodesically convex subset N (Assumption 3.1), whether there is a
856 global logarithmic map is not always guaranteed.

857 Another geometric concept induced from the connection is a covariant derivative, a notion of
858 differentiation of the vector field along the curve.

859 **Definition A.8.** [Bou23][Definition 5.28][Vector field along the curve] *Let $\gamma : [0, 1] \rightarrow M$ be a*
860 *smooth curve. A map $Z : [0, 1] \rightarrow TM$ is called a vector field on γ if $Z(t) \in T_{\gamma(t)} M$ for all*
861 *$t \in [0, 1]$. We write the set of vector fields on γ as $\mathfrak{X}(\gamma)$.*

862 **Definition A.9.** [Bou23][Theorem 5.29][Covariant derivative] *Let $\gamma : [0, 1] \rightarrow M$ be a smooth*
863 *curve and ∇ be an affine connection. Then, the covariant derivative is the unique operator $D_t :$*
864 *$\mathfrak{X}(\gamma) \rightarrow \mathfrak{X}(\gamma)$ satisfying the following properties for all $Y, Z \in \mathfrak{X}(\gamma)$, $W \in \mathfrak{X}(M)$, $g \in c^\infty([0, 1])$*
865 *and $a, b \in \mathbb{R}$:*

866 1. $D_t(aY + bZ) = aD_t(Y) + bD_t(Z)$.

867 2. $D_t(gZ) = (\frac{d}{dt}g)Z + gD_t(Z)$.

868 3. $(D_t(W \circ \gamma))(t) = \nabla_{\gamma'(t)} W$ for all $t \in [0, 1]$.

869 If ∇ is the Levi-Civita connection, then the covariant derivative also satisfies

$$\frac{d}{dt} \langle Y, Z \rangle = \langle D_t Y, Z \rangle + \langle Y, D_t Z \rangle.$$

870 Parallel transport is a notion of transporting vectors between different tangent space parallelly. The
871 parallel transport is uniquely determined by the covariant derivative.

872 **Definition A.10.** [Bou23][Definition 10.33] A vector field $Z \in \mathfrak{X}(\gamma)$ is called parallel if $D_t Z = 0$.

873 **Definition A.11.** [Bou23][Definition 10.35][Parallel transport] Let $\gamma : [0, 1] \rightarrow M$ be a smooth
874 curve. The parallel transport of the tangent vector at $T_{\gamma(t_0)} M$ to the tangent vector at $T_{\gamma(t_1)} M$ along
875 the curve γ is the map

$$\Gamma(\gamma)_{t_0}^{t_1} : T_{\gamma(t_0)} M \rightarrow T_{\gamma(t_1)} M$$

876 defined by $\Gamma(\gamma)_{t_0}^{t_1}(Z(t_0)) = Z(t_1)$ for the parallel vector field $Z \in \mathfrak{X}(\gamma)$.

877 We collect some properties of the parallel transport.

878 **Proposition A.12.** [Bou23][Proposition 10.36]

- 879 1. $\Gamma(\gamma)_{t_0}^{t_1}$ is a linear map.
- 880 2. $\Gamma(\gamma)_{t_1}^{t_2} \circ \Gamma(\gamma)_{t_0}^{t_1} = \Gamma(\gamma)_{t_0}^{t_2}$.
- 881 3. $\Gamma(\gamma)_{t_0}^{t_1} \circ \Gamma(\gamma)_{t_1}^{t_0} = id$.
- 882 4. $\langle v, w \rangle_{\gamma(t_0)} = \langle \Gamma(\gamma)_{t_0}^{t_1} v, \Gamma(\gamma)_{t_0}^{t_1} w \rangle_{\gamma(t_1)}$.

883 When γ is chosen to be the geodesic curve such that $\gamma(0) = x$ and $\gamma(1) = y$, we denote the parallel
884 transport $\Gamma(\gamma)_0^1$ as Γ_x^y . When context is clear, we will denote Γ_x^y as the (geodesic) parallel transport
885 from x to y .

886 **Remark A.13** (Properties of geodesic parallel transport). By Proposition A.12, a geodesic parallel
887 transport Γ_x^y satisfies the following properties:

- 888 1. Γ_x^y is a linear map.
- 889 2. $\Gamma_x^y \circ \Gamma_y^x = id$.
- 890 3. $\langle v, w \rangle_x = \langle \Gamma_x^y v, \Gamma_x^y w \rangle_y$.

891 Note the second property is dropped, as geodesics from x to y and y to z do not necessarily be in the
892 same curve.

893 Remark A.13 is the key properties of parallel transport used in our analysis. These properties play a
894 pivotal role when we define the parallel transport in 2-Wasserstein space (Proposition A.30).

895 The last geometric concept induced from the Levi-Civita connection is curvature.

896 **Definition A.14** (Riemannian curvature). The Riemannian curvature tensor $R(\cdot, \cdot) : \mathfrak{X}(M) \times$
897 $\mathfrak{X}(M) \times \mathfrak{X}(M) \rightarrow \mathfrak{X}(M)$ is defined by the following formula:

$$R(X, Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X, Y]} Z$$

898 where $[\cdot, \cdot]$ denotes a Lie bracket.

899 The key geometric quantity in our analysis is sectional curvature, which generalizes Gaussian
900 curvature in a 2-dimensional surface.

901 **Definition A.15** (Sectional curvature). Let $p \in M$, and denote Σ_p a set of two-dimensional subspaces
902 in $T_p M$. The sectional curvature $K : \Sigma_p \rightarrow \mathbb{R}$ is defined by the following formula:

$$K(\sigma_p) = \frac{\langle R(u, v)v, u \rangle_p}{\|u\|_p^2 \|v\|_p^2 - \langle u, v \rangle_p^2}$$

903 where $\{u, v\}$ is a basis of σ_p .

904 Note that we can write this sectional curvature as a function of two linearly independent vectors in
 905 $T_p M$ as well. In particular, if u, v are orthonormal, then $K(u, v) = \langle R(u, v)v, u \rangle_p$.
 906 A Riemannian manifold is called flat if for all p and σ_p sectional curvature $K(\sigma_p) = 0$, positively
 907 curved if $K(\sigma_p) > 0$, and negatively curved if $K(\sigma_p) < 0$.

908 A.1.1 Functional properties of functions on Riemannian manifolds

909 In this appendix, we introduce additional functional properties of functions on a Riemannian manifold.
 910 We begin with introducing the notion of geodesically convex set.

911 **Definition A.16.** [Bou23][Definition 11.2] Let (M, g) be a complete Riemannian manifold. $N \subseteq M$
 912 is called geodesically convex subset of M if for all $x, y \in N$, there exists a geodesic $\gamma : [0, 1] \rightarrow M$
 913 such that $\gamma(0) = x$, $\gamma(1) = y$, and $\gamma(t) \in N$ for all $t \in [0, 1]$.

914 Next, we introduce the notion of geodesic convexity and smoothness.

915 **Definition A.17** (Geodesic convexity and smoothness). Let $f : N \rightarrow \mathbb{R}$ be a differentiable function.

916 1. f is called geodesically α -strongly convex if for all $x, y \in N$

$$f(y) \geq f(x) + \langle \text{Grad } f(x), \log_x y \rangle_x + \frac{\alpha}{2} d^2(x, y).$$

917 If $\alpha = 0$, we say f is geodesically convex.

918 2. f is called geodesically L -smooth if for all $x, y \in N$

$$\|\Gamma_y^x \text{Grad } f(y) - \text{Grad } f(x)\|_x \leq L d(x, y).$$

919 Now, we show the key inequality induced from the geodesic L -smoothness. This is often called
 920 descent lemma.

921 **Lemma A.18** (Descent lemma). If f is geodesically L -smooth, then for all $x, y \in N$

$$f(y) \leq f(x) + \langle \text{Grad } f(x), \log_x y \rangle + \frac{L}{2} d^2(x, y).$$

922 *Proof.* Let $\gamma : [0, 1] \rightarrow M$ be a geodesic curve such that $\gamma(0) = x$, $\gamma(1) = y$. By the definition of
 923 the Riemannian logarithmic map, we get $\gamma'(0) = \log_x y$. By Fundamental Theorem of Calculus and
 924 properties of the parallel transport,

$$\begin{aligned} f(y) &= f(\gamma(1)) = f(\gamma(0)) + \int_0^1 \frac{d}{dt} (f \circ \gamma)(t) dt = f(x) + \int_0^1 \langle \text{Grad } f(\gamma(t)), \gamma'(t) \rangle dt \\ &= f(x) + \int_0^1 \langle \Gamma_{\gamma(t)}^{\gamma(0)} \text{Grad } f(\gamma(t)), \gamma'(0) \rangle dt = f(x) + \int_0^1 \langle \Gamma_{\gamma(t)}^x \text{Grad } f(\gamma(t)), \log_x y \rangle dt. \end{aligned}$$

925 Then, by subtracting $f(x) + \langle \text{Grad } f(x), \log_x y \rangle$ from the both hand sides,

$$\begin{aligned} f(y) - f(x) - \langle \text{Grad } f(x), \log_x y \rangle &= \int_0^1 \langle \Gamma_{\gamma(t)}^x \text{Grad } f(\gamma(t)) - \text{Grad } f(x), \log_x y \rangle dt \\ &\stackrel{(i)}{\leq} \int_0^1 \left\| \Gamma_{\gamma(t)}^x \text{Grad } f(\gamma(t)) - \text{Grad } f(x) \right\| \|\log_x y\| dt \\ &\stackrel{(ii)}{\leq} \int_0^1 L d(\gamma(t), x) d(x, y) dt \stackrel{(iii)}{=} L d^2(x, y) \int_0^1 t dt \\ &= \frac{L}{2} d^2(x, y). \end{aligned}$$

926 For (i) we used Cauchy-Schwartz inequality, and for (ii) we used L -smoothness property. For (iii)
 927 we used the fact that the geodesic curve satisfies $d(x, \gamma(t)) = td(x, y)$ due to the constant speed
 928 property. ■

929 A.1.2 Product Riemannian manifold

930 In Appendix A.2.1, we will encounter a product manifold. To that end, we present some preliminary
 931 facts here. We omit the details and simply list a few useful results. For more information on product
 932 Riemannian manifolds, we refer the reader to [Lee18].

933 **Definition A.19** (Product Riemannian manifold). *A product Riemannian manifold is a manifold*
 934 *$M = M_1 \times M_2$ such that each (M_1, g_1) and (M_2, g_2) are Riemannian manifolds, and the Riemannian*
 935 *metric g is defined by the product metric:*

$$g((X_1, X_2), (Y_1, Y_2)) = g_1(X_1, Y_1) + g_2(X_2, Y_2).$$

936 Product Riemannians manifold have useful properties that make the computation easier.

937 **Theorem A.20** (Levi-Civita connection of a product Riemannian manifold). *The Levi-Civita con-*
 938 *nection of a product Riemannian manifold $(M, g) = (M_1, g_1) \times (M_2, g_2)$ satisfies the following*
 939 *property:*

$$\nabla_{(X_1, X_2)}(Y_1, Y_2) = \nabla_{1, X_1} Y_1 \oplus \nabla_{2, X_2} Y_2.$$

940 The following corollary is a direct consequence of the definition of Riemannian curvature, Lie bracket,
 941 and Theorem A.20.

Corollary A.21 (Riemannian curvature of a product Riemannian manifold).

$$R((X_1, X_2), (Y_1, Y_2))(Z_1, Z_2) = R_1(X_1, Y_1)Z_1 \oplus R_2(X_2, Y_2)Z_2.$$

942 Lastly, we obtain the following collorary, which will play an important role in our later section.

943 **Corollary A.22** (Sectional curvature of product Riemannian manifold). *Let $(u_1, u_2), (v_1, v_2)$ be*
 944 *orthonormal vectors in $T_p M$. Write $A_i := \|u_i\|^2 \|v_i\|^2 - g_i(u_i, v_i)^2$. Then,*

$$K((u_1, u_2), (v_1, v_2)) = A_1 K_1(u_1, v_1) + A_2 K_2(u_2, v_2).$$

945 *Proof.* From Definition A.15, Definition A.19, and Corollary A.21, we have

$$\begin{aligned} K((u_1, u_2), (v_1, v_2)) &= g(R((u_1, u_2), (v_1, v_2))(v_1, v_2), (u_1, u_2)) \\ &= g((R_1(u_1, v_1)v_1, R_2(u_2, v_2)v_2), (u_1, u_2)) \\ &= g_1(R_1(u_1, v_1)v_1, u_1) + g_2(R_2(u_2, v_2)v_2, u_2) \\ &= A_1 K_1(u_1, v_1) + A_2 K_2(u_2, v_2). \end{aligned}$$

946

■

947 In particular, if $K_1 = 0$, i.e., one of the spaces is flat, the the curvature behavior of the product
 948 manifold is entirely determined by K_2 . This will be the case in Appendix A.2.1.

949 A.2 Wasserstein geometry

950 In this appendix, we introduce the core concept of Wasserstein geometry, which is one of our key
 951 application. We write the space of probability measures with a finite p th moment on \mathbb{R}^d by $\mathcal{P}_p(\mathbb{R}^d)$.
 952 Again, we mainly introduce the known results without proofs. For interested readers, we refer to
 953 [Vil08, AGS08, San14, Che24].

954 For $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, let $\Gamma(\mu, \nu)$ be a set of couplings of μ and ν . Wasserstein distance between μ and
 955 ν are defined as follows.

956 **Definition A.23** (Wasserstein metric). *Let $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$. Denote $\Gamma(\mu, \nu)$ to be a set of coupling*
 957 *measures of μ and ν . p -Wasserstein distance between μ and ν is defined as follows:*

$$W_p^p(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu, \nu)} \mathbb{E}_{(x, y) \sim \gamma} [\|x - y\|^p].$$

958 This is known to be a well-defined metric. A metric space $(\mathcal{P}_p(\mathbb{R}^d), W_p)$ is called p -Wasserstein
 959 space.

960 2-Wasserstein space is typically a more interesting space compared to other p -Wasserstein spaces due
 961 to its geometric properties. [Bre91, JKO98, Ott01] found out that if we restrict our attention to the
 962 probability measures which are absolutely continuous with respect to Lebesgue measure and have
 963 a finite second moment, denoted by $\mathcal{P}_{2,ac}(\mathbb{R}^d)$, then $(\mathcal{P}_{2,ac}(\mathbb{R}^d), W_2)$ endows a richer geometric
 964 properties. Specifically, while $(\mathcal{P}_{2,ac}(\mathbb{R}^d), W_2)$ is not precisely a Riemannian manifold, its geometry
 965 is almost same to the non-negatively curved Riemannian manifold.

966 The reason $(\mathcal{P}_{2,ac}(\mathbb{R}^d), W_2)$ endows a Riemannian structure is rooted from the following theorem
 967 [Bre91]:

968 **Theorem A.24** (Brenier Theorem). *If $\mu, \nu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$, then*

$$W_2^2(\mu, \nu) = \min_{T \in L^2(\mu) \text{ s.t. } T_{\#}\mu = \nu} \mathbb{E}_{x \sim \mu} [\|T(x) - x\|^2] = \min_{T \in L^2(\mu) \text{ s.t. } T_{\#}\mu = \nu} \|T - id\|_{L^2(\mu; \mathbb{R}^d)}^2.$$

969 Denote the minima as $T_{\mu, \nu}$. Then $T_{\mu, \nu}$ is a gradient of some convex function ϕ on \mathbb{R}^d μ -a.e.
 970 Furthermore, $T_{\mu, \nu} \circ T_{\nu, \mu} = id$. The minima $T_{\mu, \nu}$ is called the optimal transport map from μ to ν .

971 Theorem A.24 gives a notion of tangent direction at μ .

972 **Definition A.25** (Riemannian metric in 2-Wasserstein space). *For $\mu \in W_2(\mathbb{R}^d)$, a tangent space
 973 of μ is $T_\mu \mathcal{P}_{2,ac}(\mathbb{R}^d) = \overline{\{\nabla \psi \mid \psi \in C_c^\infty(\mathbb{R}^d)\}}^{\mathcal{L}^2(\mu)} \subset \mathcal{L}^2(\mu)$. Here, $C_c^\infty(\mathbb{R}^d)$ is a set of compactly
 974 supported smooth functions on \mathbb{R}^d . The Riemannian metric is defined as a $\mathcal{L}^2(\mu)$ -inner product. In
 975 other words, $\langle v, w \rangle_\mu = \mathbb{E}_{x \sim \mu} [\langle v(x), w(x) \rangle]$.*

976 **Remark A.26** (Interpretation of the tangent space). *By Brenier theorem, $T_{\mu, \nu} = \nabla \phi$. For arbitrary
 977 $\lambda > 0$, it follows that $\lambda(T_{\mu, \nu} - id) = \nabla(\lambda\phi - \lambda \frac{\|\cdot\|^2}{2}) \in T_\mu \mathcal{P}_{2,ac}(\mathbb{R}^d)$. This implies that the tangent
 978 space $T_\mu \mathcal{P}_{2,ac}(\mathbb{R}^d)$ can be interpreted as the set of scaled displacement fields $\lambda(T_{\mu, \nu} - id)$. If $X \sim \mu$
 979 and $Y \sim \nu$, then $\lambda(T_{\mu, \nu} - id)(X) = \lambda(Y - X)$, which corresponds to directions in the usual
 980 Euclidean sense. From this perspective, the tangent space is naturally constructed to represent
 981 Euclidean directions at the level of individual particles.*

982 One can naturally define a geodesic curve in $(\mathcal{P}_{2,ac}(\mathbb{R}^d), W_2)$, by pushforwarding the interpolation
 983 between particles to the measure space.

984 **Definition A.27** (Geodesic in Wasserstein space). *A geodesic curve $\gamma : [0, 1] \rightarrow \mathcal{P}_{2,ac}(\mathbb{R}^d)$ such that
 985 $\gamma(0) = \mu$ and $\gamma(1) = \nu$ can be defined as follows:*

$$\gamma(t) = ((1-t)id + tT_{\mu, \nu})_{\#}\mu.$$

986 The exponential map and logarithmic map are then defined accordingly.

987 **Definition A.28** (Exponential map and Logarithmic map in Wasserstein space). *For $\mu, \nu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$
 988 and $v \in \mathcal{L}^2(\mu)$, exponential map and logarithmic map of $(\mathcal{P}_{2,ac}(\mathbb{R}^d), W_2)$ are defined as follows:*

$$\begin{aligned} \exp_\mu(v) &= (v + id)_{\#}\mu, \\ \log_\mu(\nu) &= T_{\mu, \nu} - id. \end{aligned}$$

989 A favorable property of 2-Wasserstein space is that the exponential map (and accordingly logarithmic
 990 map) is globally well-defined on $\mathcal{L}^2(\mu)$, i.e., 2-Wasserstein space satisfies Assumption 3.1.

991 This Riemannian structure induces 2-Wasserstein metric. Observe the Riemannian distance in-
 992 duced from the above structure coincides with the Wasserstein distance; $d(\mu, \nu)^2 = \|\log_\mu \nu\|^2 =$
 993 $\|T_{\mu, \nu} - id\|^2 = W_2^2(\mu, \nu)$.

994 One can define a geodesic parallel transport as well.

995 **Definition A.29.** [AG08][Parallel transport] *For $\mu, \nu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ and $v \in T_\mu \mathcal{P}_{2,ac}(\mathbb{R}^d)$,*

$$\Gamma_\mu^\nu v := \Pi_\nu(v \circ T_{\nu, \mu}).$$

996 *Here, Π_\cdot is a projection operator $\mathcal{L}^2(\cdot) \rightarrow T_\cdot \mathcal{P}_{2,ac}(\mathbb{R}^d)$.*

997 This definition of parallel transport is not entirely satisfactory, as it involves the operator Π , which
 998 lacks an explicit form. However, recall our analysis only requires the properties of parallel transport
 999 in Remark A.13. It turns out that even if we drop Π , and consider $\Gamma_\mu^\nu v = v \circ T_{\nu,\mu}$ as a parallel
 1000 transport onto $\mathcal{L}^2(\mu)$, the corresponding parallel transport still has properties in Remark A.13, which
 1001 are sufficient for our analyses.

1002 **Proposition A.30** (Transfer lemma). *For $\mu, \nu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ and $v \in \mathcal{L}^2(\mu)$, define $\Gamma_\mu^\nu v := v \circ T_{\nu,\mu}$.
 1003 Then,*

1004 1. Γ_μ^ν is linear operator on $\mathcal{L}^2(\mu)$.

1005 2. $\Gamma_\mu^\nu \circ \Gamma_\nu^\mu = id$.

1006 3. $\langle v, w \rangle_\mu = \langle \Gamma_\mu^\nu v, \Gamma_\mu^\nu w \rangle_\nu$.

1007 *Proof.* Property 1 is direct: for $v, w \in \mathcal{L}^2(\mu)$ and $a, b \in \mathbb{R}$, $\Gamma_\mu^\nu(av + bw) = av \circ T_{\mu,\nu} + bw \circ T_{\mu,\nu} =$
 1008 $a\Gamma_\mu^\nu v + b\Gamma_\mu^\nu w$.

1009 Property 2 is from Theorem A.24.

1010 Property 3 is a direct consequence of the change of the measure formula:

$$\langle v, w \rangle_\mu = \int \langle v(x), w(x) \rangle d(T_{\nu,\mu})_{\#}\nu(x) = \int \langle v \circ T_{\nu,\mu}(x), w \circ T_{\nu,\mu}(x) \rangle d\nu(x) = \langle \Gamma_\mu^\nu v, \Gamma_\mu^\nu w \rangle_\nu.$$

1011 ■

1012 Therefore, by Proposition A.30, we can use the *un-projected* parallel transport $\cdot \circ T_{\nu,\mu}$ as a parallel
 1013 transport $\Gamma_\mu^\nu \cdot$ and $\mathcal{L}^2(\mu)$ as the tangent space for our analysis. In fact, such parallel transport and
 1014 tangent space are sufficient for other first-order Wasserstein gradient flow analyses as well (e.g.,
 1015 [AGS08, SKL20]).

1016 Now, we introduce a sectional curvature in 2-Wasserstein space. Note that in our analysis, the use of
 1017 non-negative curvature is solely through Lemma C.1. In the 2-Wasserstein space, an analogous result
 1018 follows solely from the transport map's optimality, without invoking the concept of sectional curvature
 1019 of Wasserstein space (Lemma C.3). Nevertheless, for the sake of completeness, we present the result
 1020 that $(\mathcal{P}_{2,ac}(\mathbb{R}^d), W_2)$ is indeed a non-negatively curved space. To establish this, we introduce the
 1021 continuity equation and the notion of covariant derivative in the 2-Wasserstein space.

1022 **Definition A.31** (Continuity equation). *Let μ_t be a flow in $\mathcal{P}_{2,ac}(\mathbb{R}^d)$. For given μ_t , there exists a
 1023 vector field $v_t \in \mathcal{L}^2(\mu_t)$ such that*

$$\partial_t \mu_t = -\text{div}(\mu_t v_t).$$

1024 *Such v_t is called a (velocity) vector field of the flow μ_t .*

1025 One can think of v_t as a velocity at μ_t , and plays a similar role as $\gamma'(t)$ in Riemannian manifolds.

1026 **Definition A.32** (Covariant derivative). *A covariant derivative of $w_t \in T_{\mu_t} \mathcal{P}_{2,ac}(\mathbb{R}^d)$ along a curve
 1027 μ_t is defined by the following formula:*

$$\nabla_{v_t} w_t = \Pi_{\mu_t} \left(\lim_{h \rightarrow 0} \frac{\Gamma_{\mu_t}^{\mu_{t+h}} w_{t+h} - w_t}{h} \right).$$

1028 *Here, Γ is a parallel transport defined in Definition A.29, and v_t is a vector field of the flow μ_t .*

1029 We are ready to introduce the result that 2-Wasserstein space is non-negatively curved.

1030 **Lemma A.33.** *Let v_t, w_t be orthonormal elements in $T_{\mu_t} \mathcal{P}_{2,ac}(\mathbb{R}^d)$. Then, the sectional curvature of
 1031 the subspace spanned by these two tangent vectors is as follows:*

$$K_{\mu_t}(v_t, w_t) = 3 \|\nabla v_t \cdot w_t - \nabla_{w_t} v_t\|_{\mathcal{L}^2(\mu_t)}^2$$

1032 *where the first ∇ is Euclidean gradient, and the second $\nabla_{w_t} v_t$ is a covariant derivative.*

1033 We refer to [AG08][Proposition 7.2] or [Lot07][Corollary 5.13] for the derivation.

1034 The last ingredients we need for the analysis of the Wasserstein space are notions of gradient,
1035 convexity, and smoothness. These concepts are defined in an analogous manner to the Riemannian
1036 case. Again, we omit the detail and just present the result.

1037 Wasserstein gradient is defined analogously to the formula $d_v f(x) = \langle \text{Grad } f(x), v \rangle_x$ in Riemannian
1038 manifold.

1039 **Definition A.34** (Wasserstein gradient). *For a functional $\mathcal{F} : \mathcal{P}_{2,ac}(\mathbb{R}^d) \rightarrow \mathbb{R}$, the Wasserstein*
1040 *gradient of \mathcal{F} at μ_0 is an element of $\mathcal{L}^2(\mu_0)$ satisfying the following equation:*

$$\partial_t \mathcal{F}(\mu_t) \big|_{t=0} = \langle \text{Grad}_{W_2} \mathcal{F}(\mu_0), v_0 \rangle_{\mu_0}.$$

1041 Here v_t is a vector field of the flow μ_t .

1042 One has the following explicit formula:

$$\text{Grad}_{W_2} \mathcal{F}(\mu) = \nabla \frac{\delta \mathcal{F}(\mu)}{\delta \mu}.$$

1043 Here, ∇ is Euclidean gradient and $\frac{\delta \mathcal{F}(\mu)}{\delta \mu}$ is the first variation.

1044 Here, the role of $\gamma'(0)$ is changed to v_0 . For the derivation we refer to [Che24][Theorem 1.4.1].

1045 Now equipped with the Wasserstein gradient, we can define a generalized geodesic convexity and
1046 smoothness. Motivated by Proposition A.30, we use *un-projected* parallel transport instead of the
1047 true parallel transport for the entire constructions. The construction of generalized geodesic convexity
1048 using *un-projected* parallel transport in Wasserstein space was already introduced in various literature
1049 of optimal transport [AGS08, San14, SKL20, DBCS23].

1050 **Definition A.35** (Generalized geodesic convexity and geodesic smoothness in Wasserstein space).
1051 *Let $\mathcal{F} : \mathcal{P}_{2,ac}(\mathbb{R}^d) \rightarrow \mathbb{R}$ be a differentiable functional.*

1052 1. *\mathcal{F} is called generalized geodesically α -strongly convex with base $\pi \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ if for all*
1053 *$\mu, \nu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$*

$$\mathcal{F}(\nu) \geq \mathcal{F}(\mu) + \langle \text{Grad}_{W_2} \mathcal{F}(\mu) \circ T_{\pi,\mu}, T_{\pi,\nu} - T_{\pi,\mu} \rangle_{\pi} + \frac{\alpha}{2} \|T_{\pi,\nu} - T_{\pi,\mu}\|_{\pi}^2.$$

1054 *If $\alpha = 0$, we say it is generalized geodesically convex with base π . If for given μ, ν , \mathcal{F}*
1055 *is generalized geodesically α -strongly convex with base $\pi = \mu$, it is called geodesically*
1056 *α -strongly convex. If \mathcal{F} is generalized geodesically α -strongly convex with base π for all*
1057 *$\pi \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$, then it is called generalized geodesically α -strongly convex.*

1058 2. *\mathcal{F} is called generalized geodesically L -smooth with base $\pi \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ if for all $\mu, \nu \in$*
1059 *$\mathcal{P}_{2,ac}(\mathbb{R}^d)$*

$$\|\text{Grad}_{W_2} \mathcal{F}(\nu) \circ T_{\pi,\nu} - \text{Grad}_{W_2} \mathcal{F}(\mu) \circ T_{\pi,\mu}\|_{\pi} \leq L \|T_{\pi,\nu} - T_{\pi,\mu}\|_{\pi}.$$

1060 *Again, geodesic L -smoothness and generalized geodesic L -smoothness are defined in*
1061 *analogous way.*

1062 By the same reasoning as in Lemma A.18, geodesically L -smooth functional in Wasserstein space
1063 also satisfies the descent lemma in Wasserstein sense, i.e.,

$$\mathcal{F}(\nu) \leq \mathcal{F}(\mu) + \langle \text{Grad}_{W_2} \mathcal{F}(\mu), T_{\mu,\nu} - id \rangle_{\mu} + \frac{L}{2} W_2^2(\mu, \nu).$$

1064 Finally, we present a complete proof of Proposition 6.2.

1065 *Proof of Proposition 6.2.* Since the argument is identical for both the 2-Wasserstein and Bu-
1066 res–Wasserstein geometries, we only present the proof in the 2-Wasserstein case. First, we show
1067 whenever V is α -strongly convex then \mathcal{V} is generalized geodesically α -strongly convex. For arbitrary

1068 $\mu, \nu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ and arbitrary base $\pi \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$, let $T_{\pi,\mu}, T_{\pi,\nu}$ be the optimal transport maps.
 1069 Then, from the strong convexity of V , for any $z \sim \pi$,

$$V(T_{\pi,\nu}(z)) \geq V(T_{\pi,\mu}(z)) + \langle \nabla V(T_{\pi,\mu}(z)), T_{\pi,\nu}(z) - T_{\pi,\mu}(z) \rangle + \frac{\alpha}{2} \|T_{\pi,\nu}(z) - T_{\pi,\mu}(z)\|^2.$$

1070 Take an expectation over $z \sim \pi$ on both sides. The result follows from the fact $\text{Grad}_{W_2} \mathcal{V}(\mu)(\cdot) =$
 1071 $\nabla V(\cdot)$, which is from [San14][Remark 7.13] and Definition A.34.

1072 Now, we show the generalized geodesic L -smoothness. Again, for any $z \sim \pi$, by the L -smoothness
 1073 of V ,

$$\|\nabla V(T_{\pi,\nu}(z)) - \nabla V(T_{\pi,\mu}(z))\| \leq L \|T_{\pi,\nu}(z) - T_{\pi,\mu}(z)\|.$$

1074 Again, taking the expectation over $z \sim \pi$ on both sides yields the desired result. ■

1075 A.2.1 Bures-Wasserstein geometry

1076 In this appendix, we briefly introduce Bures-Wasserstein space $BW(\mathbb{R}^d)$, a space of Gaussian
 1077 measures equipped with W_2 metric. Main takeaways of this appendix are as follow:

- 1078 1. $BW(\mathbb{R}^d)$ is a product Riemannian manifold with non-negative sectional curvature.
- 1079 2. $BW(\mathbb{R}^d)$ is a geodesically convex subset of $(\mathcal{P}_{2,ac}(\mathbb{R}^d), W_2)$ and totally geodesic submani-
 1080 fold. In this regard, we can take $N = BW(\mathbb{R}^d)$ for our algorithm.
- 1081 3. This example shows how one can parameterize the transport map to make the algorithm
 1082 implementable as in Equation (6.3), (6.4).
- 1083 4. This example confirms that $BW(\mathbb{R}^d)$, and therefore the 2-Wasserstein space, do not admit the
 1084 curvature upper bound. Consequently, existing acceleration methods requiring the curvature
 1085 upper bound are not well-suited for solving the optimization problems in Wasserstein space.

1086 Again, we briefly list the results. For detail, we refer to [Tak09, BJL19, ACGS21, LCB⁺22,
 1087 DBCS23].

1088 **Definition A.36** (Optimal transport map between Gaussian). *The optimal transport map between*
 1089 $\mu_0 = N(m_0, \Sigma_0)$ *and* $\mu_1 = N(m_1, \Sigma_1)$ *is defined as follows:*

$$T_{\mu_0, \mu_1}(x) = m_1 + \Sigma_0^{-1/2} (\Sigma_0^{1/2} \Sigma_1 \Sigma_0^{1/2})^{1/2} \Sigma_0^{-1/2} (x - m_0).$$

1090 Definition A.36 says the optimal transport map between Gaussians is an affine map. This fact
 1091 provides two favorable results.

1092 First, since affine transform of the Gaussian is also a Gaussian, from Definition A.27 every geodesic
 1093 interpolation between two Gaussians is also Gaussian. This shows $BW(\mathbb{R}^d)$ is a geodesically convex
 1094 subset of 2-Wasserstein space. In addition it implies $BW(\mathbb{R}^d)$ is totally geodesic submanifold of
 1095 2-Wasserstein space [Lee18][Exercise 8.4].

1096 Second, we can identify $\mu = N(m, \Sigma) \cong (m, \Sigma) \in \mathbb{R}^d \times \text{SPD}(d)$ and $T_\mu BW(\mathbb{R}^d) \cong (a, S) \in$
 1097 $\mathbb{R}^d \times \text{Sym}(d)$. Here, $\text{SPD}(d)$ is the space of $\mathbb{R}^{d \times d}$ symmetric positive definite matrices, and $\text{Sym}(d)$
 1098 is the space of $\mathbb{R}^{d \times d}$ symmetric matrices. By writing an affine map as $T(x) = a + S(x - m)$ for
 1099 fixed m (which is the mean of μ), any affine map starting at $\mu = N(m, \Sigma)$ can be parameterized
 1100 by (a, S) . Under this identification, we can view $BW(\mathbb{R}^d)$ space as a product Riemannian manifold
 1101 of $\mathbb{R}^d \times \text{SPD}(d)$ (Appendix A.1.2). Then one can parameterize every quantity in Appendix A.2 by
 1102 this product manifold sense. For instance, the vector corresponding to the optimal transport map is
 1103 $(m_1, \Sigma_0^{-1/2} (\Sigma_0^{1/2} \Sigma_1 \Sigma_0^{1/2})^{1/2} \Sigma_0^{-1/2})$.

1104 Then, we can define Riemannian metric, exponential map, logarithmic map, and Bures-Wasserstein
 1105 gradient in terms of parameters as well.

1106 **Definition A.37** (Riemannian metric of Bures-Wasserstein space). *Let* $\mu = N(m, \Sigma)$. *The Riemannian*
 1107 *metric of* $BW(\mathbb{R}^d)$ *is define by*

$$\langle (a_0, S_0), (a_1, S_1) \rangle_\mu = \langle a_0, a_1 \rangle_{\mathbb{R}^d} + \text{tr}(S_0 \Sigma S_1).$$

1108 **Definition A.38.** [LCB⁺22][Appendix B.3] Let $\mu_i = N(m_i, \Sigma_i)$. The exponential map and a
 1109 logarithm map in $BW(\mathbb{R}^d)$ are defined by

$$\begin{aligned}\exp_{\mu_0}((a, S)) &= N(a + m_0, (S + I)\Sigma_0(S + I)), \\ \log_{\mu_0}(\mu_1) &= (m_1 - m_0, \Sigma_0^{-1/2}(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2}\Sigma_0^{-1/2} - I).\end{aligned}$$

1110 **Definition A.39.** [LCB⁺22][Appendix B.3] Bures-Wasserstein metric of the functional \mathcal{F} can be
 1111 written as a function on $\mathbb{R}^d \times SPD(d)$, the space of the mean and covariance. Then, for $m \in \mathbb{R}$ and
 1112 $\Sigma \in SPD(d)$,

$$\text{Grad}_{BW} \mathcal{F}(m, \Sigma) = (\nabla_m \mathcal{F}(m, \Sigma), 2\nabla_\Sigma \mathcal{F}(m, \Sigma)).$$

1113 See [LCB⁺22, DBCS23] for further discussion.

1114 Using the isometry between the function representation and the vector-matrix representation of
 1115 $T_p BW(\mathbb{R}^d)$, we can define the following operation, which can be used to construct the (un-projected)
 1116 parallel transport.

1117 **Definition A.40.** For $(a, S) \in T_{\mu_1} BW(\mathbb{R}^d)$ and $(b, R) \in T_{\mu_0} BW(\mathbb{R}^d)$, we have the following
 1118 operation.

$$(a, S) \circ (b, R) = (a + Sb - Sm_1, SR).$$

1119 In particular,

$$\Gamma_{(m_0, \Sigma_0)}^{(m_1, \Sigma_1)}(a, S) = (a, S\Sigma_0^{-1/2}(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2}\Sigma_0^{-1/2}).$$

1120 Some works adopt an alternative definition of the Bures-Wasserstein metric; we make a remark that
 1121 this definition is equivalent to the one we present here. This remark plays a pivotal role when we
 1122 conduct actual calculation in $BW(\mathbb{R}^d)$ space (Appendix D.1.1).

1123 **Remark A.41** (Equivalent formulation of Bures-Wasserstein metric). In some works (e.g.,
 1124 [HJM21]), $BW(\mathbb{R}^d)$ metric is defined as $\langle (a, S), (b, R) \rangle_\mu = \langle a, b \rangle_{\mathbb{R}^d} + \frac{1}{2} \text{tr}(L_\Sigma(S)R)$, where
 1125 $L_\Sigma(S)$ is the Lyapunov operator defined via the solution of $L_\Sigma(S)\Sigma + \Sigma L_\Sigma(S) = S$. While it has
 1126 the different form with what we introduced earlier, these two formulations turned out to be equivalent:
 1127 our formulation is from Wasserstein perspective, and the other formulation is from Riemannian
 1128 perspective. In our setup, we define the tangent vector to directly parameterize the optimal transport
 1129 map. That said, this does not directly fit with the Riemannian framework. For instance, if we con-
 1130 sider the curve $\gamma(t) = \exp_\mu(t(a, S))$ defined by our exponential map, then the velocity at $t = 0$ is
 1131 $\dot{\gamma}(0) = (a, S\Sigma + \Sigma S)$, which does not coincide with the tangent vector (a, S) . By contrast, under
 1132 the Lyapunov operator based definition, the initial velocity is exactly $\dot{\gamma}(0) = (a, S)$. However, since
 1133 there is a one-to-one correspondence between $S\Sigma + \Sigma S$ and S for a given Σ , one may regard these
 1134 two definitions as equivalent by identifying the tangent vector with $v_0 = S$ whenever the velocity
 1135 $\dot{\gamma}(0) = S\Sigma + \Sigma S$ appears. One can change all corresponding quantities accordingly, and these two
 1136 definitions turned out to be equivalent. We have chosen our formulation because it leads to a simpler
 1137 algorithm (6.3) that avoids solving the Lyapunov equation.

1138 Lastly, we end up with the analysis of the curvature of $BW(\mathbb{R}^d)$. In particular, we show the result
 1139 that even $BW(\mathbb{R}^d)$ space does not allow the curvature upper bound, indicating that the 2-Wasserstein
 1140 space does not have the curvature upper bound as well.

1141 By applying Corollary A.22 and the flatness of Euclidean space, we obtain the following result:

1142 **Corollary A.42.** For any $\mu \in BW(\mathbb{R}^d)$ and $\{(a, S), (b, R)\}$ orthonormal vectors in $T_\mu BW(\mathbb{R}^d) =$
 1143 $\mathbb{R}^d \times \text{Sym}(d)$,

$$K_{BW(\mathbb{R}^d)}((a, S), (b, R)) = (\text{tr}(S\Sigma S) \text{tr}(R\Sigma R) - \text{tr}(S\Sigma R)^2) K_{\text{Sym}_+(\mathbb{R}^d \times d)}(S, R).$$

1144 Therefore, to analyze the curvature of $BW(\mathbb{R}^d)$, it is sufficient to analyze the space of positive
 1145 definite matrices, without accounting for the mean component. In this regard, without the loss of
 1146 generality we consider $\mu = N(0, \Sigma)$. Then, since Σ is a symmetric positive definite matrix, it is
 1147 diagonalizable, and therefore we can write $\Sigma = PD(\lambda_i)P^T$ with P being an orthogonal matrix

1148 and all real positive eigenvalues λ_i . Then, it is known that $\text{Sym}(d)$ is spanned by the following
 1149 orthonormal basis [Tak09]:

$$\left\{ e_+ = \frac{P(E_{11} + E_{dd})P^T}{\sqrt{\lambda_1 + \lambda_d}}, e_{ij} = \frac{P(E_{ii} - E_{jj})P^T}{\sqrt{\lambda_i + \lambda_j}}, f_{ij} = \frac{P(E_{ij} + E_{ji})P^T}{\sqrt{\lambda_i + \lambda_j}} \right\}_{1 \leq i, j \leq d}$$

1150 where E_{ij} is a matrix with only its (i, j) entry is 1 and 0 otherwise.

1151 Using this orthonormal basis, we can characterize all of the sectional curvature in $\text{SPD}(d)$ as follows:

1152 **Lemma A.43.** [Tak09][Sectional curvature of Bures-Wasserstein space]

$$\begin{aligned} K(e_+, f_{ij}) &= \frac{3\lambda_i\lambda_j}{(\lambda_i + \lambda_j)^2(\lambda_1 + \lambda_d)} \quad (i = 1 \text{ or } j = d), \\ K(e_{ik}, f_{ij}) &= \frac{3\lambda_i\lambda_j}{(\lambda_i + \lambda_j)^2(\lambda_i + \lambda_k)} \quad (j \neq k), \\ K(e_{ij}, f_{ij}) &= \frac{12\lambda_i\lambda_j}{(\lambda_i + \lambda_j)^3}, \\ K(f_{ij}, f_{ik}) &= \frac{3\lambda_j\lambda_k}{(\lambda_i + \lambda_j)(\lambda_j + \lambda_k)(\lambda_i + \lambda_k)} \quad (j \neq k), \\ K(\text{any other combinations}) &= 0. \end{aligned}$$

1153 This explicit form indicates that the curvature upper bound at μ depends on the smallest eigenvalue
 1154 of the covariance matrix Σ . Since the space of Gaussian distributions does not have the uniform
 1155 positive eigenvalue lower bound, $BW(\mathbb{R}^d)$ does not have the uniform curvature upper bound. See
 1156 [Tak09] for more discussions on the sectional curvature of $BW(\mathbb{R}^d)$ space.

1157 In general, the curvature of a submanifold and the curvature of its ambient manifold needs not be
 1158 the same. However, if the submanifold is totally geodesic, by Gauss formula [Lee18][Theorem 8.2]
 1159 and the fact that the second fundamental form vanishes [Lee18][Exercise 8.4], the curvature of the
 1160 submanifold coincides to the curvature of the ambient manifold. Since $BW(\mathbb{R}^d)$ is a totally geodesic
 1161 submanifold of the 2-Wasserstein space [CL20], Lemma A.43 implies that 2-Wasserstein space also
 1162 does not have the sectional curvature upper bound.

1163 B Deferred proofs

1164 B.1 Deferred proofs for Section 5

1165 This appendix contains the proofs of Section 5.

1166 Before we proceed, we introduce more convenient formulation of Q_{ij} . Using Proposition A.12, one
 1167 can write Q_{ij} as follows:

$$\begin{aligned} Q_{ij} &= 2f(x_i) - 2f(x_j) - 2 \left\langle \text{Grad } f(x_j), \log_{x_j} x_i \right\rangle_{x_j} \\ &\quad - \|\text{Grad } f(x_i)\|_{x_i}^2 - \|\text{Grad } f(x_j)\|_{x_j}^2 + 2 \left\langle \text{Grad } f(x_j), \Gamma_{x_i}^{x_j} \text{Grad } f(x_i) \right\rangle_{x_j}. \end{aligned}$$

1168 This formulation will be used frequently for the rest of the proof.

Proof of Lemma 5.1.

$$\begin{aligned} RHS &= -\|\log_{x_0} x_*\|_{x_0}^2 + \|\log_{x_n} x_*\|_{x_n}^2 + \frac{1}{4r_k^2} \|\text{Grad } f(x_n)\|_{x_n}^2 + \frac{1}{r_k} \langle \log_{x_n} x_*, \text{Grad } f(x_n) \rangle_{x_n} \\ &\leq -\|\log_{x_0} x_*\|_{x_0}^2 + \|\log_{x_{n-1}} x_*\|_{x_{n-1}}^2 + \|\log_{x_{n-1}} x_n\|_{x_{n-1}}^2 - 2 \langle \log_{x_{n-1}} x_*, \log_{x_{n-1}} x_n \rangle_{x_{n-1}} \end{aligned}$$

$$\begin{aligned}
& + \frac{1}{4r_k^2} \|\text{Grad } f(x_n)\|_{x_n}^2 + \frac{1}{r_k} \langle \text{Grad } f(x_n), \log_{x_n} x_* \rangle_{x_n} \\
& = \frac{1}{4r_k^2} \|\text{Grad } f(x_n)\|_{x_n}^2 + \frac{1}{r_k} \langle \text{Grad } f(x_n), \log_{x_n} x_* \rangle_{x_n} - \|\log_{x_0} x_*\|_{x_0}^2 + \|\log_{x_{n-1}} x_*\|_{x_{n-1}}^2 \\
& \quad + \eta_{n-1}^2 \|\text{Grad } f(x_{n-1})\|_{x_{n-1}}^2 + 2\eta_{n-1} \langle \log_{x_{n-1}} x_*, \text{Grad } f(x_{n-1}) \rangle_{x_{n-1}} \\
& \leq \frac{1}{4r_k^2} \|\text{Grad } f(x_n)\|_{x_n}^2 + \frac{1}{r_k} \langle \text{Grad } f(x_n), \log_{x_n} x_* \rangle_{x_n} \\
& \quad - \|\log_{x_0} x_*\|_{x_0}^2 + \|\log_{x_{n-2}} x_*\|_{x_{n-2}}^2 + \eta_{n-2}^2 \|\text{Grad } f(x_{n-2})\|_{x_{n-2}}^2 + 2\eta_{n-2} \langle \log_{x_{n-2}} x_*, \text{Grad } f(x_{n-2}) \rangle_{x_{n-2}} \\
& \quad + \eta_{n-1}^2 \|\text{Grad } f(x_{n-1})\|_{x_{n-1}}^2 + 2\eta_{n-1} \langle \log_{x_{n-1}} x_*, \text{Grad } f(x_{n-1}) \rangle_{x_{n-1}} \\
& \leq \dots (\text{inductively apply Lemma C.1 on } \|\log_{x_i} x_*\|_{x_i}^2) \\
& \leq \frac{1}{4r_k^2} \|\text{Grad } f(x_n)\|_{x_n}^2 + \frac{1}{r_k} \langle \text{Grad } f(x_n), \log_{x_n} x_* \rangle_{x_n} \\
& \quad + \sum_{i=1}^n \eta_{n-i}^2 \|\text{Grad } f(x_{n-i})\|_{x_{n-i}}^2 + 2 \sum_{i=1}^n \eta_{n-i} \langle \log_{x_{n-i}} x_*, \text{Grad } f(x_{n-i}) \rangle_{x_{n-i}} \\
& \quad - \underbrace{\|\log_{x_0} x_*\|_{x_0}^2 + \|\log_{x_0} x_*\|_{x_0}^2}_{=0} = LHS.
\end{aligned}$$

1169 Here, all the inequalities are obtained from repeatedly applying Lemma C.1 on each

1170 $\|\log_{x_n} x_*\|_{x_n}^2, \|\log_{x_{n-1}} x_*\|_{x_{n-1}}^2, \dots, \|\log_{x_1} x_*\|_{x_1}^2$. ■

1171 *Proof of Lemma 5.3.* First, from the gradient update, one has $\log_{x_0} x_1 = -(\rho - 1) \text{Grad } f(x_0)$ as
 1172 $\eta_0 = \rho - 1$. Using Lemma C.4, Prop A.12, $\eta_0 = \rho - 1$, and $\text{Grad } f(x_*) = 0$, one can proceed as
 1173 follows:

$$\begin{aligned}
\sum_{i,j} \lambda_{ij} Q_{ij} & = \rho Q_{01} + Q_{10} + (\rho - 1) Q_{1*} + (\rho - 1) Q_{*0} + \frac{1}{2r_1} Q_{*1} \\
& = \frac{f(x_*) - f(x_1)}{r_1} - 2\rho \underbrace{\langle \text{Grad } f(x_1), \log_{x_1} x_0 \rangle_{x_1}}_{=-\langle \Gamma_{x_1}^{x_0} \text{Grad } f(x_1), \log_{x_0} x_1 \rangle_{x_0}} - \rho \underbrace{\|\text{Grad } f(x_0) \circ T_{1,0} - f(x_1)\|_{x_1}^2}_{\|\text{Grad } f(x_0) - \Gamma_{x_1}^{x_0} \text{Grad } f(x_1)\|_{x_0}^2} \\
& \quad - 2 \langle \text{Grad } f(x_0), \log_{x_0} x_1 \rangle_{x_0} - \|\Gamma_{x_1}^{x_0} \text{Grad } f(x_1) - \text{Grad } f(x_0)\|_{x_0}^2 - (\rho - 1) \underbrace{\|\Gamma_{x_*}^{x_1} \text{Grad } f(x_1)\|_{x_*}^2}_{=\|\text{Grad } f(x_1)\|_{x_1}^2} \\
& \quad - 2(\rho - 1) \langle \text{Grad } f(x_0), \log_{x_0} x_* \rangle_{x_0} - (\rho - 1) \|\text{Grad } f(x_0)\|_{x_0}^2 - \frac{1}{r_1} \langle \text{Grad } f(x_1), \log_{x_1} x_* \rangle_{x_1} \\
& \quad - \frac{1}{2r_1} \|\text{Grad } f(x_1)\|_{x_1}^2 \\
& = \frac{f(x_*) - f(x_1)}{r_1} - 2\rho(\rho - 1) \langle \Gamma_{x_1}^{x_0} \text{Grad } f(x_1), \text{Grad } f(x_0) \rangle_{x_0} - (\rho + 1) \|\text{Grad } f(x_0) - \Gamma_{x_1}^{x_0} \text{Grad } f(x_1)\|_{x_0}^2 \\
& \quad + (\rho - 1) \|\text{Grad } f(x_0)\|_{x_0}^2 - (\rho - 1) \|\text{Grad } f(x_1)\|_{x_1}^2 - 2(\rho - 1) \langle \text{Grad } f(x_0), \log_{x_0} x_* \rangle_{x_0} \\
& \quad - \frac{1}{r_1} \langle \text{Grad } f(x_1), \log_{x_1} x_* \rangle_{x_1} - \frac{1}{2r_1} \|\text{Grad } f(x_1)\|_{x_1}^2 \\
& = \frac{f(x_*) - f(x_1)}{r_1} - 2 \|\text{Grad } f(x_0)\|_{x_0}^2 - \underbrace{\left(2\rho + \frac{1}{2r_1}\right)}_{=\frac{1}{4r_1^2}} \|\text{Grad } f(x_1)\|_{x_1}^2 - 2(\rho - 1) \langle \text{Grad } f(x_0), \log_{x_0} x_* \rangle_{x_0}
\end{aligned}$$

$$\begin{aligned}
& - 2 \underbrace{(\rho^2 - 2\rho - 1)}_{=0} \langle \text{Grad } f(x_0), \Gamma_{x_1}^{x_0} \text{Grad } f(x_1) \rangle_{x_0} - \frac{1}{r_1} \langle \text{Grad } f(x_1), \log_{x_1} x_* \rangle_1 \\
& = \frac{f(x_*) - f(x_1)}{r_1} - 2 \|\text{Grad } f(x_0)\|_{x_0}^2 - \frac{1}{4r_1^2} \|\text{Grad } f(x_1)\|_{x_1}^2 \\
& \quad - 2(\rho - 1) \langle \text{Grad } f(x_0), \log_{x_0} x_* \rangle_{x_0} - \frac{1}{r_1} \langle \text{Grad } f(x_1), \log_{x_1} x_* \rangle_{x_1} \\
& = RHS.
\end{aligned}$$

1174

1175 *Proof of Lemma 5.4.* From the construction of σ_{ij} , we have

$$\sum_{i,j=0,\dots,2n+1,*} \sigma_{ij} Q_{ij} = \sum_{i,j=0,\dots,n,*} \lambda_{ij}^{(k)} Q_{ij} + (1+2\rho) \sum_{i,j=n+1,\dots,2n+1,*} \lambda_{i-n-1,j-n-1}^{(k)} Q_{ij}.$$

1176 We begin with subtracting $\sum_{ij} \sigma_{ij} Q_{ij}$ from RHS. Since we assumed the inequality (5.1),

$$\begin{aligned}
RHS - \sum_{ij} \sigma_{ij} Q_{ij} & \geq \left(\frac{1}{r_{k+1}} - \frac{2+2\rho}{r_k} \right) f(x_*) + \frac{1}{r_k} f(x_n) + \left(\frac{1+2\rho}{r_k} - \frac{1}{r_{k+1}} \right) f(x_{2n+1}) \\
& \quad + 2\rho \sum_{i=n+1}^{2n} \eta_i^2 \|\text{Grad } f(x_i)\|_{x_i}^2 + 4\rho \sum_{i=n+1}^{2n} \eta_i \langle \log_{x_i} x_*, \text{Grad } f(x_i) \rangle_{x_i} \\
& \quad - \left(\eta_n^2 - \frac{1}{4r_k^2} \right) \|\text{Grad } f(x_n)\|_{x_n}^2 - \left(2\eta_n - \frac{1}{r_k} \right) \langle \log_{x_n} x_*, \text{Grad } f(x_n) \rangle_{x_n} \\
& \quad - \left(\frac{1}{4r_{k+1}^2} - \frac{1+2\rho}{4r_k^2} \right) \|\text{Grad } f(x_{2n+1})\|_{x_{2n+1}}^2 \\
& \quad - \left(\frac{1}{r_{k+1}} - \frac{1+2\rho}{r_k} \right) \langle \log_{x_{2n+1}} x_*, \text{Grad } f(x_{2n+1}) \rangle_{x_{2n+1}}.
\end{aligned}$$

1177 We want to remove inner product terms so that we can express the formula in terms of norms (to
1178 show non-negativity). To this end, we consider

$$A := -2\rho \sum_{j=n+1}^{2n} \eta_j Q_{*,j} + \left(\frac{1}{2r_{k+1}} - \frac{1+2\rho}{2r_k} \right) Q_{*,2n+1} + \left(1 + \rho^{k-1} - \frac{1}{2r_k} \right) Q_{*,n}.$$

1179 Then, by subtracting A , one gets

$$\begin{aligned}
RHS - \sum_{ij} \sigma_{ij} Q_{ij} - A & \geq 2(1 + \rho^{k-1}) (f(x_n) - f(x_*)) + 4\rho \sum_{i=n+1}^{2n} \eta_i (f(x_i) - f(x_*)) \\
& \quad + 2\rho \sum_{i=n+1}^{2n} \eta_i (\eta_i - 1) \|\text{Grad } f(x_i)\|_{x_i}^2 - \left(1 + \rho^{k-1} - \frac{1}{2r_k} \right) \left(\rho^{k-1} + \frac{1}{2r_k} \right) \|\text{Grad } f(x_n)\|_{x_n}^2 \\
& \quad - \left(\frac{1}{2r_{k+1}} - \frac{1+2\rho}{2r_k} - \frac{1}{4r_{k+1}^2} + \frac{1+2\rho}{4r_k^2} \right) \|\text{Grad } f(x_{2n+1})\|_{x_{2n+1}}^2 \\
& := B.
\end{aligned}$$

1180 If $B \geq 0$, then the claimed inequality follows with coefficients in the theorem, i.e.,

$$\lambda_{ij}^{(k+1)} = \sigma_{ij} + \begin{cases} -2\rho\eta_j & i = *, j = n+1, \dots, 2n \\ 1 + \rho^{k-1} - \frac{1}{2r_k} & i = *, j = n \\ \left(\frac{1}{2r_{k+1}} - \frac{1+2\rho}{2r_k} \right) & i = *, j = 2n+1 \\ 0 & \text{otherwise.} \end{cases}$$

1181 Therefore, we show $B \geq 0$ for the rest of the proof. Since x_* is a minimizer and $\eta_i \geq 1$, we have
 1182 $2(1 + \rho^{k-1})(f(x_n) - f(x_*)) + 4\rho \sum_{i=n+1}^{2n} \eta_i (f(x_i) - f(x_*)) \geq 0$. In addition, again $\eta_i \geq 1$
 1183 implies $2\rho \sum_{i=n+1}^{2n} \eta_i (\eta_i - 1) \|\text{Grad } f(x_i)\|_{x_i}^2 \geq 0$.

1184 Therefore, if $1 + \rho^{k-1} - \frac{1}{2r_k} \leq 0$ and $\frac{1}{2r_{k+1}} - \frac{1+2\rho}{2r_k} - \frac{1}{4r_{k+1}^2} + \frac{1+2\rho}{4r_k^2} \leq 0$, then $B \geq 0$. We show
 1185 these inequalities hold under our choice of r_k . For simplicity, let $a_k = \frac{1}{2r_k} = \frac{1+\sqrt{4\rho^{2k}-3}}{2}$.

1186 For $1 + \rho^{k-1} \leq a_k$, observe the following calculations:

$$\begin{aligned} 1 + \rho^{k-1} \leq \frac{1 + \sqrt{4\rho^{2k}-3}}{2} &\Leftrightarrow (2\rho^{k-1} + 1)^2 \leq 4\rho^{2k} - 3 \Leftrightarrow \rho^{2k-2}(\rho^2 - 1) \geq \rho^{k-1} + 1 \\ &\stackrel{(i)}{\Leftrightarrow} 2\rho^{2k-1} \geq \rho^{k-1} + 1 \Leftrightarrow 2\rho^k \geq 1 + \frac{1}{\rho^{k-1}}. \end{aligned}$$

1187 (i) comes from $\rho^2 - 1 = 2\rho$. Now, one can see the last inequality is true, as $LHS \geq 2 \geq RHS$. This
 1188 proves the coefficient of $\|\text{Grad } f(x_n)\|_{x_n}^2$ is non-negative.

1189 Next, to observe $a_{k+1} - a_k^2 \leq (1 + 2\rho)(a_k - a_k^2)$, we write $\sqrt{4\rho^{2k}-3} := S_k$ for simplicity.
 1190 Then, observe the following calculation:

$$\begin{aligned} a_k &= \frac{1 + S_k}{2} \\ a_k^2 &= \frac{1 + 2S_k + S_k^2}{4} = \frac{4\rho^{2k} - 2 + 2S_k}{4} = \rho^{2k} + \frac{S_k - 1}{2} \\ &\Rightarrow a_k - a_k^2 = 1 - \rho^{2k}. \\ \therefore a_{k+1} - a_{k+1}^2 &\leq (1 + 2\rho)(a_k - a_k^2) \stackrel{(ii)}{\Leftrightarrow} 1 - \rho^{2k+2} \leq \rho^2(1 - \rho^{2k}) \\ &\Leftrightarrow 1 \leq \rho^2. \end{aligned}$$

1191 Since the last inequality holds, the coefficient of $\|\text{Grad } f(x_{2n+1})\|_{x_{2n+1}}^2$ is also non-negative.

1192 In sum, we have $B := RHS - LHS \geq 0$. This proves the desired inequality.

1193 Lastly, to establish the non-negativity of $\lambda_{ij}^{(k)}$, note that if we initialize with $\lambda_{ij}^{(1)}$ as in Lemma 5.3, then
 1194 at each index where our coefficients are nonzero, they match those of [AP24c]. The non-negativity of
 1195 these coefficients was already proven in that work.

1196 ■

1197 *Proof of Theorem 4.1.* First consider the case $L = 1$. Lemma 5.3 and 5.4 together imply the
 1198 inequality in (5.1), i.e., Lemma 5.2. Then, applying Lemma 5.1 to RHS of (5.1) leads to the desired
 1199 result.

1200 For general L , let $g = \frac{1}{L}f$. Then, by the linearity of the Riemannian gradient and parallel transport,
 1201 g satisfies (3.1) with $L = 1$. By applying $L = 1$ case on g one gets

$$\frac{1}{L} (f(x_n) - f(x_*)) = g(x_n) - g(x_*) \leq r_k d^2(x_0, x_*).$$

1202 ■

1203 B.2 Deferred proofs for Section 6

1204 This appendix contains the proofs for the results in Section 6.

1205 *Proof of Corollary 6.1.* The proof goes exactly same as in Theorem 4.1 and 4.2, once one substitutes
 1206 the following quantities in the proof of Lemma 5.1, 5.3, 5.4, and Theorem 4.1, 4.2 accordingly.

- 1207 • Set $M = N = \mathcal{P}_{2,ac}(\mathbb{R}^d)$.

- 1208 • Change the Riemannian metric by $\langle \cdot, \cdot \rangle_\mu = \langle \cdot, \cdot \rangle_{\mathcal{L}^2(\mu)} = \mathbb{E}_{x \sim \mu} [\langle \cdot(x), \cdot(x) \rangle]$.
- 1209 • Substitute the notion of generalized geodesic convexity and smoothness to Definition A.35.
- 1210 • Take $\exp_\mu(v) = (id + v)_{\# \mu}$.
- 1211 • Take $\log_\mu \nu = T_{\mu, \nu} - id$.
- 1212 • Set $\Gamma_\mu^\nu v = v \circ T_{\nu, \mu}$.
- 1213 • Set $\text{Grad} f(x)$ to $\text{Grad}_{W_2} \mathcal{F}(\mu)$, introduced in Definition A.34.
- 1214 • Substitute Lemma C.1 to Lemma C.3.
- 1215 • Substitute Lemma C.4 to Lemma C.5.

1216 Note Assumption 3.1 is satisfied in this case due to the global well-definedness of the exponential
 1217 map and logarithmic map in 2-Wasserstein space (see Definition A.28). One part we need to verify
 1218 is the fact that \mathcal{F} being generalized geodesically convex and geodesically L -smooth implies the
 1219 inequality (3.1) in Wasserstein sense. To obtain this result, it is sufficient to verify whether the
 1220 condition $z \in N$ in Proposition 3.8 holds in this case. In fact, it turns out that in Wasserstein
 1221 space this is true regardless of the choice of the functional \mathcal{F} , as long as its Wasserstein gradient is
 1222 well-defined. For any $\mu, \nu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$, consider

$$\begin{aligned} \pi &:= \exp_\nu \left(-\frac{1}{L} (\text{Grad}_{W_2} \mathcal{F}(\nu) - \Gamma_\mu^\nu \text{Grad}_{W_2} \mathcal{F}(\mu)) \right) \\ &= \left(id - \frac{1}{L} \text{Grad}_{W_2} \mathcal{F}(\nu) + \frac{1}{L} \text{Grad}_{W_2} \mathcal{F}(\mu) \circ T_{\nu, \mu} \right)_{\# \nu}. \end{aligned}$$

1223 Since $\text{Grad}_{W_2} \mathcal{F}(\nu), \text{Grad}_{W_2} \mathcal{F}(\mu) \circ T_{\nu, \mu} \in \mathcal{L}^2(\nu)$, $\pi \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$. Therefore, the same logic in
 1224 Proposition 3.8 (or Lemma D.10) yields the inequality (3.1). Then, the proof is an exact duplicate of
 1225 the proofs of our main theorems. ■

1226 **Remark B.1** (Proof for Bures-Wasserstein space). *The proof of Corollary 6.1 holds the same if we*
 1227 *replace N to be $BW(\mathbb{R}^d)$, as $BW(\mathbb{R}^d)$ is a totally geodesic submanifold of $\mathcal{P}_{2,ac}(\mathbb{R}^d)$. This justifies*
 1228 *our choice of N in Section 6.1.*

1229 C Auxiliary lemmas

1230 This section aggregates the required lemmas for intermediate calculations.

1231 **Lemma C.1.** *For any $x, y, z \in N$, one has*

$$\|\log_y z\|^2 \leq \|\log_x z\|^2 + \|\log_x y\|^2 - 2 \langle \log_x y, \log_x z \rangle.$$

1232 *Proof.* In [KY22][Lemma 5.2], plug-in $p_A = x, p_B = y, x = z, v_A = \log_x y, v_B = 0, r = 1$, and
 1233 $\zeta = 1$ (due to the non-negativity of the curvature). Then, expanding the formula leads to the desired
 1234 bound. ■

1235 **Remark C.2.** *The constant ζ comes from the Hessian comparison theorem [AOBL20, KY22]. In*
 1236 *their theorem, they assumed the curvature upper bound as well as the diameter bound of the set.*
 1237 *However, by carefully analyzing the proof of [AOBL20][Lemma 2], one can check the one side*
 1238 *inequality involving ζ only requires K_{\min} , and does not require K_{\max} as well as the diameter bound*
 1239 *D . This is why our analysis requires neither curvature upper bound nor the diameter bound.*

1240 One can write Wasserstein space version of Lemma C.1 without bringing the curvature of Wasserstein
 1241 space.

1242 **Lemma C.3.** *For any $\mu, \nu, \pi \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$, one has*

$$\|T_{\nu, \pi} - id\|_\nu^2 \leq \|T_{\mu, \pi} - id\|_\mu^2 + \|T_{\mu, \nu} - id\|_\mu^2 + 2 \langle T_{\mu, \nu} - id, T_{\mu, \pi} - id \rangle_\mu.$$

1243 *Proof.* Observe $T_{\mu,\pi} \circ T_{\nu,\mu}$ is also a transport map from ν to π . By the optimality of the optimal
 1244 transport map $T_{\nu,\pi}$ (Theorem A.24),

$$\begin{aligned} \|T_{\nu,\pi} - id\|_\nu^2 &\leq \|T_{\mu,\pi} \circ T_{\nu,\mu} - id\|_\nu^2 \stackrel{(i)}{=} \|T_{\mu,\pi} - T_{\mu,\nu}\|_\mu^2 \\ &= \|T_{\mu,\pi} - id\|_\mu^2 + \|T_{\mu,\nu} - id\|_\mu^2 + 2 \langle T_{\mu,\nu} - id, T_{\mu,\pi} - id \rangle_\mu. \end{aligned}$$

1245 For (i) we used Proposition A.30. ■

1246 Next, we show how logarithmic map changes under the parallel transport.

1247 **Lemma C.4.** *For all $x, y \in N$, let Γ_x^y be a parallel transport from x to y induced from the geodesic*
 1248 *connecting x and y . Then,*

$$\Gamma_x^y \log_x y = -\log_y x.$$

1249 This result is analogous result of $y - x = -(x - y)$ in Euclidean case.

1250 *Proof.* Let $\gamma : [0, 1] \rightarrow M$ be a geodesic curve such that $\gamma(0) = x$ and $\gamma(1) = y$. Then, by definition
 1251 of logarithmic map, one gets $\gamma'(0) = \log_x y$.

1252 Now, consider the reversed geodesic $\sigma(t) := \gamma(1 - t)$. Then, $\sigma'(0) = -\gamma'(1) = \log_y x$. By the
 1253 property of the geodesic and the parallel transport,

$$\Gamma_x^y \log_x y = \Gamma_x^y \gamma'(0) = \gamma'(1) = -\sigma'(0) = -\log_y x.$$

1254 ■

1255 Again, we provide a Wasserstein space version of Lemma C.4.

1256 **Lemma C.5.** *For all $\mu, \nu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$,*

$$(T_{\mu,\nu} - id) \circ T_{\nu,\mu} = -(T_{\nu,\mu} - id).$$

1257 *Proof.* This is a direct consequence of Theorem A.24. ■

1258 D Additional discussions

1259 D.1 Generalized geodesic convexity

1260 The notion of generalized geodesic convexity was originally introduced in optimal transport and has
 1261 found various usages in Wasserstein geometry, including the theoretical analysis of the proximal
 1262 operator in the 2-Wasserstein space [AGS08][Lemma 9.2.7], [SKL20, DBCS23], and its connection
 1263 to Γ -convergence [AGS08][Lemma 9.2.9]. To the best of our knowledge, this notion has not yet
 1264 been explored in the Riemannian geometry literature. We therefore expect that introducing it in this
 1265 context could provide new tools for analyzing proximal operators and Γ -convergence on Riemannian
 1266 manifolds, as it has in the 2-Wasserstein setting-areas that, to date, remain underdeveloped.

1267 In this appendix, we provide some examples of generalized geodesically convex functionals for
 1268 readers who are not familiar with the concept. Then, we prove Proposition 3.8, which is one of our
 1269 main findings.

1270 First, recall the notion of generalized geodesic convexity.

1271 **Definition D.1** (Generalized geodesic convexity). *A differentiable function $f : N \rightarrow \mathbb{R}$ is called*
 1272 *generalized geodesically α -strongly convex with base $z \in M$ if for all $x, y \in N$*

$$f(y) \geq f(x) + \langle \Gamma_x^z \text{Grad } f(x), \log_z y - \log_z x \rangle_z + \frac{\alpha}{2} \|\log_z y - \log_z x\|_z^2.$$

1273 *If $\alpha = 0$, we say f is generalized geodesically convex with base z . If f is generalized geodesically*
 1274 *α -strongly convex for all $z \in M$, then f is called generalized geodesically α -strongly convex.*

1275 D.1.1 Examples of generalized geodesically convex functional

1276 We start with the trivial example: Euclidean space.

1277 **Example D.2.** A differentiable, α -strongly convex function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is generalized geodesically
1278 α -strongly convex.

1279 *Proof.* In Euclidean space, $\exp_x(v) = x + v$ and $\log_x y = y - x$. Since f is differentiable and
1280 α -strongly convex, for all $x, y, z \in \mathbb{R}^d$

$$\begin{aligned} f(y) &\geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2} \|y - x\|^2 \\ &= f(x) + \langle \nabla f(x), (y - z) - (x - z) \rangle + \frac{\alpha}{2} \|(y - z) - (x - z)\|^2. \end{aligned}$$

1281

1282 Now, we move to nontrivial examples: non-Euclidean manifolds. As mentioned in the main body,
1283 this concept has already been widely discussed in the Wasserstein space. Therefore, there are some
1284 known examples in 2-Wasserstein space. We first introduce some generalized geodesically convex
1285 functionals in Wasserstein space: potential energy functional and internal energy functional.

1286 **Example D.3** (Potential energy). Consider a function $V : \mathbb{R}^d \rightarrow \mathbb{R}$. A functional $\mathcal{V}(\mu) :=$
1287 $\mathbb{E}_{X \sim \mu}[V(X)]$ is called a potential functional. If V is α -strongly convex (L -smooth) in \mathbb{R}^d , then
1288 \mathcal{V} geodesically α -strongly convex (resp. L -smooth).

1289 This is duplicate of Proposition 6.2.

1290 **Example D.4** (Internal energy). Let $F : [0, \infty) \rightarrow (-\infty, \infty]$ be a proper, lower semi-continuous
1291 convex function such that

$$F(0) = 0, \quad \liminf_{s \downarrow 0} \frac{F(s)}{s^\alpha} > -\infty \text{ for some } \alpha > \frac{d}{d+2}.$$

1292 Consider a functional $\mathcal{H}_F : \mathcal{P}_{2,ac}(\mathbb{R}^d) \rightarrow \mathbb{R}$ defined by

$$\mathcal{H}_F(\mu) := \int_{\mathbb{R}^d} F(\mu(x)) dx.$$

1293 If the map $s \mapsto s^d F(s^{-d})$ is convex and non-increasing in $(0, \infty)$, then the functional \mathcal{H}_F is
1294 generalized geodesically convex.

1295 We refer to [AGS08][Proposition 9.3.9] for the proof.

1296 **Remark D.5.** Some widely used choice of F satisfying the conditions are as follows:

- 1297 1. $F(s) = s \log s$. This choice leads to \mathcal{H}_F being the differential entropy functional.
- 1298 2. For any $q > 1$, $F(s) = s^q$.
- 1299 3. For $m \geq 1 - 1/d$, $F(s) = \frac{1}{m-1} s^m$.

1300 Now, we present examples on Riemannian manifolds. We begin by providing sufficient conditions
1301 for generalized geodesic convexity, which turns out to be useful in verifying the generalized geodesic
1302 convexity for a given functional.

1303 **Lemma D.6** (Criteria for generalized geodesic convexity). Fix $z \in N$. For any $x, y \in N$, let $\gamma(t)$ be
1304 any curve such that $\gamma(0) = x$, $\gamma(1) = y$, and $\dot{\gamma}(0) = \Gamma_z^x(\log_z y - \log_z x)$. If a differentiable function
1305 $f : N \rightarrow \mathbb{R}$ satisfies either one of the following conditions, then f is generalized geodesically convex
1306 with base $z \in N$.

- 1307 1. Zeroth-order criterion: $(1-t)f(x) + tf(y) \geq (f \circ \gamma)(t)$ for all $t \in [0, 1]$.
- 1308 2. Second-order criterion: $\frac{d^2}{dt^2} (f \circ \gamma)(t) \geq 0$ for all $t \in (0, 1)$.

1309 **Proof. 1. Zeroth-order criterion:** Since f is differentiable, differentiate the both hand sides with
 1310 respect to t and plug-in $t = 0$. Then,

$$f(y) - f(x) \geq \left. \frac{d}{dt} \right|_{t=0} (f \circ \gamma)(t) = \langle \text{Grad } f(x), \Gamma_z^x(\log_z y - \log_z x) \rangle = \langle \Gamma_x^z \text{Grad } f(x), \log_z y - \log_z x \rangle.$$

1311 **2. Second-order criterion:** By Taylor's theorem,

$$\begin{aligned} f(y) &= f(x) + \left. \frac{d}{dt} \right|_{t=0} (f \circ \gamma)(t) + \int_0^1 (1-t) \frac{d^2}{dt^2} (f \circ \gamma)(t) dt \\ &\geq f(x) + \langle \text{Grad } f(x), \Gamma_z^x(\log_z y - \log_z x) \rangle = f(x) + \langle \Gamma_x^z \text{Grad } f(x), \log_z y - \log_z x \rangle. \end{aligned}$$

1312 ■

1313 **Remark D.7** (Existence of γ). *It is natural to ask whether such curve $\gamma(t)$ exists. In fact, as long as*
 1314 *the exponential map is defined for sufficiently large neighborhood of x , there always exists a curve*
 1315 *satisfying the conditions. Let $v(t) := t\Gamma_z^x(\log_z y - \log_z x) + t^2(\log_x y - \Gamma_z^x(\log_z y - \log_z x))$, and*
 1316 *define $\gamma(t) = \exp_x(v(t))$. Observe $\gamma(0) = x$ and $\gamma(1) = y$. Furthermore, since the differential of*
 1317 *the exponential map is the identity at the origin, by the chain rule*

$$\dot{\gamma}(0) = d\exp_x(v(0))[v'(0)] = \Gamma_z^x(\log_z y - \log_z x).$$

1318 *In certain Riemannian manifolds with a particularly well-behaving exponential map, simpler curves*
 1319 *can be used. For instance, in the 2-Wasserstein space, a more natural choice of curve is available.*
 1320 *Fix a base $\pi \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$. For any $\mu, \nu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$, let $\gamma(t) := \exp_\pi((1-t)\log_\pi \mu + t\log_\pi \nu) =$
 1321 $((1-t)T_{\pi,\mu} + tT_{\pi,\nu})_{\#}\pi$ *be a curve. Then, $\gamma(0) = \mu, \gamma(1) = \nu$, and the velocity vector field*
 1322 *corresponding to $\gamma(t)$ is $v_t = (T_{\pi,\nu} - T_{\pi,\mu}) \circ T_{\gamma(t),\pi}$ [DBCS23][Appendix B.2].**

1323 As a specific example, we consider the entropy functional on $SPD(d)$ space. This example will
 1324 show how one can verify the generalized geodesic convexity using Lemma D.6.

1325 **Example D.8** (Entropy of Gaussian). *Consider a functional $\mathcal{H} : SPD(d) \rightarrow \mathbb{R}$ defined by*
 1326 $\mathcal{H}(A) = -\frac{1}{2} \log \det A$. *This functional is in fact the entropy functional of the multivariate Gaussian*
 1327 *distribution $N(0, A)$ (up to an affine transformation). There are two natural Riemannian metrics in*
 1328 *$SPD(d)$ space [FAP⁺05, PFA05, BH06, HMJG21, Ngu22, TP22, KPB25].*

1329 1. *Affine invariant metric:* $d_{AI}(A, B) := \|\log A^{-1/2} B A^{-1/2}\|_F$, *and $\langle S, R \rangle_A =$*
 1330 $\text{tr}(A^{-1} S A^{-1} R)$ *for $S, R \in \text{Sym}(d)$. This metric induces non-positively curved geome-*
 1331 *try on $SPD(d)$.*

1332 2. *Bures-Wasserstein metric:* $d_{BW}^2(A, B) := \text{tr}(A) + \text{tr}(B) - 2\text{tr}(A^{1/2} B A^{1/2})^{1/2}$, *and*
 1333 $\langle S, R \rangle_A = \text{tr}(S A R)$ *for $S, R \in \text{Sym}(d)$. This metric induces non-negatively curved*
 1334 *geometry on $SPD(d)$.*

1335 *Both geometries originate from the geometry of zero-mean Gaussian distributions. The metric d_{AI}*
 1336 *arises from the Fisher information metric associated with zero-mean Gaussians [Nie23], while the*
 1337 *metric d_{BW} corresponds to the Wasserstein geometry of zero-mean Gaussians, as described in*
 1338 *Appendix A.2.1. Under both geometries, $\mathcal{H}(A)$ is generalized geodesically convex.*

1339 Note that d_{BW} corresponds to the 2-Wasserstein distance between Gaussians, so the result for d_{BW}
 1340 is a special case of Example D.4. Nonetheless, we present the proof entirely in the language of
 1341 Riemannian geometry to demonstrate that the notion of generalized geodesic convexity remains valid
 1342 purely within the Riemannian framework.

1343 **Proof of Example D.8.** In both cases, we apply the second-order criterion from Lemma D.6. The
 1344 general strategy is to construct a curve that satisfies the required conditions with respect to a fixed
 1345 starting point, endpoint, and base point. The specific choice of curve should reflect the underlying
 1346 geometry. Once the curve is chosen, we compute the time derivative of the functional along the curve;

1347 this can be carried out entirely using matrix calculus, without explicitly invoking the Riemannian
1348 structure.

1349 We will use N to denote the arbitrary base point, and M_0, M_1 to denote the starting point and the
1350 endpoint of the curve.

1351 **1. Affine invariant metric:** We first construct a curve satisfying the desired property. We consider a
1352 curve on $\text{SPD}(d)$ defined by

$$M(t) := N^{1/2} \exp \left((1-t) \log \left(N^{-1/2} M_0 N^{-1/2} \right) + t \log \left(N^{-1/2} M_1 N^{-1/2} \right) \right) N^{1/2}.$$

1353 Here, \exp and \log are just matrix exponential and logarithm, not the Riemannian operators. Observe
1354 $M(0) = M_0$ and $M(1) = M_1$. Now, we check $M'(0)$. For simplicity, denote $c(t) = (1 -$
1355 $t) \log \left(N^{-1/2} M_0 N^{-1/2} \right) + t \log \left(N^{-1/2} M_1 N^{-1/2} \right)$. Then,

$$\begin{aligned} M'(0) &= N^{1/2} \exp(c(0)) c'(0) N^{1/2} = M_0 N^{-1/2} \left(\log(N^{-1/2} M_1 N^{-1/2}) - \log(N^{-1/2} M_0 N^{-1/2}) \right) N^{1/2} \\ &= (M_0 N^{-1})^{1/2} (M_0 N^{-1})^{1/2} N^{1/2} \left(\log(N^{-1/2} M_1 N^{-1/2}) - \log(N^{-1/2} M_0 N^{-1/2}) \right) N^{1/2} \\ &= (M_0 N^{-1})^{1/2} \left[(M_0 N^{-1})^{1/2} N^{1/2} \left(\log(N^{-1/2} M_1 N^{-1/2}) - \log(N^{-1/2} M_0 N^{-1/2}) \right) N^{1/2} \right] \\ &= (M_0 N^{-1})^{1/2} \left[(M_0 N^{-1})^{1/2} N^{1/2} \left(\log(N^{-1/2} M_1 N^{-1/2}) - \log(N^{-1/2} M_0 N^{-1/2}) \right) N^{1/2} \right]^T \\ &= (M_0 N^{-1})^{1/2} \left[N^{1/2} \left(\log(N^{-12} M_1 N^{-1/2}) - \log(N^{-1/2} M_0 N^{-1/2}) \right) N^{1/2} \right] ((M_0 N^{-1})^{1/2})^T. \end{aligned}$$

1356 This exactly coincides to $\Gamma_N^{M_0}(\log_N M_1 - \log_N M_0)$ on $(\text{SPD}(d), d_{AI})^2$. Thus, the curve $M(t)$
1357 satisfies the conditions in Lemma D.6.

1358 Now, we compute $\frac{d^2}{dt^2} \mathcal{H}(M_t)$. First, observe

$$\begin{aligned} \mathcal{H}(M_t) &= -\frac{1}{2} \log(\det N \det[\exp(c(t))]) = -\frac{1}{2} \log \det N - \frac{1}{2} \log \exp(\text{tr}[c(t)]) \\ &= -\frac{1}{2} \log \det N - \frac{1}{2} \text{tr}(c(t)). \end{aligned}$$

1359 Then,

$$\frac{d}{dt} \mathcal{H}(M_t) = -\frac{1}{2} \text{tr}(c'(t)) = -\frac{1}{2} \text{tr} \left(\log(N^{-1/2} M_1 N^{-1/2}) - \log(N^{-1/2} M_0 N^{-1/2}) \right).$$

1360 Note this formula does not involve t anymore. Therefore, we have $\frac{d^2}{dt^2} \mathcal{H}(M_t) = 0$ for all $t \in (0, 1)$.
1361 Since this result holds for arbitrary base point N , we have the generalized geodesic convexity³.

1362 **2. Bures-Wasserstein metric:** We again start with constructing a curve satisfying the desired
1363 properties. As noted in Remark A.41, in this setting we must match the *tangent vector corresponding*
1364 *to* $M'(0)$ with $\Gamma_N^{M_0}(\log_N M_1 - \log_N M_0)$, rather than matching $M'(0)$ directly. We consider
1365 $\nu = N(0, N)$, $\mu_0 = N(0, M_0)$, and $\mu_1 = N(0, M_1)$. From Appendix A.2.1, the optimal transport
1366 map between 0-mean Gaussians is a linear map. Therefore, for any π_0, π_1 , we denote B_{L_0, L_1} to be
1367 the matrix corresponding to the optimal transport map between $\pi_0 = N(0, L_0)$, $\pi_1 = N(0, L_1)$, i.e.,
1368 $T_{\pi_0, \pi_1}(x) = B_{L_0, L_1} x$. Now, consider a curve on $\text{SPD}(d)$ defined by

$$M(t) := ((1-t)I + tB_{N, M_1} B_{M_0, N}) M_0 ((1-t)I + tB_{N, M_1} B_{M_0, N})^T {}^4.$$

1369 Then, $M(0) = M_0$ trivially and $M(1) = M_1$; for any $X \sim N(0, M_0)$, on the one hand
1370 $B_{N, M_1} B_{M_0, N} X = T_{\nu, \mu_1} \circ T_{\mu_0, \nu}(X) \sim N(0, M_1)$, and on the other hand $B_{N, M_1} B_{M_0, N} X \sim$

²For the formula of the parallel transport and Riemannian logarithmic map on (SPD, d_{AI}) , see [Ngu22][Supplement 1.1].

³In fact, this means the functional \mathcal{H} is generalized geodesically linear.

⁴While $B_{N, M_1} B_{M_0, N} - I$ may not be symmetric, the formula on the right hand side is still well-defined. Consequently, there is no harm in defining the curve via this formula.

1371 $N(0, (B_{N,M_1} B_{M_0,N}) M_0 (B_{N,M_1} B_{M_0,N})^T)$, meaning $(B_{N,M_1} B_{M_0,N}) M_0 (B_{N,M_1} B_{M_0,N})^T =$
1372 M_1 . In addition, since $M'(0) = B_{N,M_1} B_{M_0,N} M_0 + M_0 B_{N,M_1} B_{M_0,N}$, from the identifica-
1373 tion in Remark A.41 the tangent vector corresponding to $M'(0)$ is $V_0 = B_{N,M_1} B_{M_0,N} - I =$
1374 $\Gamma_N^{M_0}(B_{N,M_1} - B_{N,M_0})$. Therefore, the curve $M(t)$ satisfies the conditions in Lemma D.6.

1375 Now, we compute $\frac{d^2}{dt^2} \mathcal{H}(M_t)$. First, since $M_t = A_t M_0 A_t^T$, $\mathcal{H}(M_t) = -\log \det(A_t) - \frac{1}{2} \log \det M_0$.
1376 Then, for all $t \in (0, 1)$,

$$\begin{aligned} \frac{d^2}{dt^2} \mathcal{H}(M_t) &= -\frac{d^2}{dt^2} \log \det(A_t) = -\frac{d}{dt} \operatorname{tr} \left(A_t^{-1} \dot{A}_t \right) = -\frac{d}{dt} \operatorname{tr} \left(A_t^{-1} (B_{N,M_1} B_{M_0,N} - I) \right) \\ &= -\operatorname{tr} \left(\frac{d}{dt} A_t^{-1} (B_{N,M_1} B_{M_0,N} - I) \right) = \operatorname{tr} \left(A_t^{-1} \dot{A}_t A_t^{-1} (B_{N,M_1} B_{M_0,N} - I) \right) \\ &= \operatorname{tr} \left(A_t^{-1} (B_{N,M_1} B_{M_0,N} - I) A_t^{-1} (B_{N,M_1} B_{M_0,N} - I) \right) \\ &\stackrel{(i)}{=} \operatorname{tr} \left(\left[A_t^{-1/2} (B_{N,M_1} B_{M_0,N} - I) A_t^{-1/2} \right]^2 \right) \geq 0 \end{aligned}$$

1377 which is the desired inequality. For (i), we claim that $A_t^{-1/2}$ is well-defined as the principal square
1378 root for all $t \in (0, 1)$. This follows from the fact that both $B_{N,M_1}, B_{M_0,N}$ are optimal transport maps
1379 and thus, by Brenier's Theorem A.24, they are non-negative definite. Consequently, the product
1380 $B_{N,M_1} B_{M_0,N}$ also has non-negative eigenvalues. Since A_t is a convex combination of the identity
1381 matrix I and a matrix with non-negative eigenvalues, it follows that all eigenvalues of A_t are strictly
1382 positive on $t \in (0, 1)$. Hence, all eigenvalues of A_t^{-1} are positive for $t \in (0, 1)$, and then $A_t^{-1/2}$ is
1383 well-defined as the principal square root.

1384 Again, since the inequality holds for arbitrary base N , we obtain the generalized geodesic convexity
1385 of \mathcal{H} . ■

1386 D.1.2 Proof of Proposition 3.8

1387 Next, we prove Proposition 3.8. To prove Proposition 3.8, we need to introduce the notion of
1388 co-coercivity.

1389 **Definition D.9** (Geodesic co-coercivity). *A differentiable function $f : N \rightarrow \mathbb{R}$ is called geodesically*
1390 *co-coercive if for all $x, y \in N$*

$$\langle \Gamma_y^x \operatorname{Grad} f(y) - \operatorname{Grad} f(x), \log_x y \rangle \geq \frac{1}{L} \left\| \Gamma_y^x \operatorname{Grad} f(y) - \operatorname{Grad} f(x) \right\|^2.$$

1391 The geodesic co-coercivity condition links L -smoothness and (3.1). The next lemma is a general
1392 version of Proposition 3.8, which shows the relationship between L -smoothness, co-coercivity, and
1393 (3.1).

1394 **Lemma D.10.** *For a differentiable function $f : N \rightarrow \mathbb{R}$, The below relationship holds:*

$$(3.1) \stackrel{(i)}{\Rightarrow} \text{geodesic co-coercivity} \stackrel{(ii)}{\Rightarrow} \text{geodesic } L\text{-smoothness}$$

1395 *In addition, suppose for all $x, y \in N$, f satisfies $z := \exp_y \left(-\frac{1}{L} (\operatorname{Grad} f(y) - \Gamma_x^y \operatorname{Grad} f(x)) \right) \in$*
1396 *N . Then, if f is generalized geodesically convex,*

$$\text{geodesic } L\text{-smoothness} \stackrel{(iii)}{\Rightarrow} (3.1).$$

1397 **Proof. (i):** By applying (3.1) for (x, y) and (y, x) and using Lemma C.4, one gets

$$\begin{aligned} f(y) - f(x) - \langle \operatorname{Grad} f(x), \log_x y \rangle - \frac{1}{2L} \left\| \Gamma_y^x \operatorname{Grad} f(y) - \operatorname{Grad} f(x) \right\|^2 &\geq 0, \\ f(x) - f(y) + \langle \Gamma_y^x \operatorname{Grad} f(y), \log_x y \rangle - \frac{1}{2L} \left\| \Gamma_y^x \operatorname{Grad} f(y) - \operatorname{Grad} f(x) \right\|^2 &\geq 0. \end{aligned}$$

1398 Summing up two inequalities, one gets

$$\langle \Gamma_y^x \operatorname{Grad} f(y) - \operatorname{Grad} f(x), \log_x y \rangle \geq \frac{1}{L} \left\| \Gamma_y^x \operatorname{Grad} f(y) - \operatorname{Grad} f(x) \right\|^2.$$

1399 **(ii):** Using Cauchy-Schwartz inequality on the co-coercivity condition, one gets

$$\frac{1}{L} \|\Gamma_y^x \text{Grad } f(y) - \text{Grad } f(x)\|^2 \leq \|\Gamma_y^x \text{Grad } f(y) - \text{Grad } f(x)\| \|\log_x y\|.$$

1400 Since $\|\log_x y\| = d(x, y)$, one gets the result.

1401 **(iii):** Take $z = \exp_y(-\frac{1}{L}(\text{Grad } f(y) - \Gamma_x^y \text{Grad } f(x)))$. Write $f(x) - f(y) = f(x) - f(z) +$
 1402 $f(z) - f(y)$. Then, using generalized geodesic convexity with base y and Lemma A.18,

$$\begin{aligned} f(x) - f(y) &= f(x) - f(z) + f(z) - f(y) \\ &\leq -\langle \Gamma_x^y \text{Grad } f(x), \log_y z - \log_y x \rangle + \langle \text{Grad } f(y), \log_y z \rangle + \frac{L}{2} \|\log_y z\|^2 \\ &= -\left\langle \Gamma_x^y \text{Grad } f(x), -\frac{1}{L}(\text{Grad } f(y) - \Gamma_x^y \text{Grad } f(x)) - \log_y x \right\rangle \\ &\quad + \left\langle \text{Grad } f(y), -\frac{1}{L}(\text{Grad } f(y) - \Gamma_x^y \text{Grad } f(x)) \right\rangle \\ &\quad + \frac{1}{2L} \|\text{Grad } f(y) - \Gamma_x^y \text{Grad } f(x)\|^2 \\ &= \langle \Gamma_x^y \text{Grad } f(x), \log_y x \rangle - \frac{1}{2L} \|\text{Grad } f(y) - \Gamma_x^y \text{Grad } f(x)\|^2 \\ &= -\langle \text{Grad } f(x), \log_x y \rangle - \frac{1}{2L} \|\Gamma_y^x \text{Grad } f(y) - \text{Grad } f(x)\|^2. \end{aligned}$$

1403 Here, we again used Lemma C.4 for the last equality. This is equivalent to the desired inequality. ■

1404 D.2 Moving to strongly convex smooth functional: Restarting method

1405 We now turn our attention to the geodesically strongly convex case. Although an alternative silver
 1406 step-size scheme has been proposed for strongly convex, smooth problems in the Euclidean setting
 1407 [AP24b], the co-coercivity condition it relies on does not carry over to geodesically strongly convex,
 1408 smooth problems on Riemannian manifolds. In contrast, for convex, smooth functions the co-
 1409 coercivity condition admits a natural Riemannian interpretation via generalized geodesic convexity
 1410 and geodesic smoothness (see Proposition 3.8 and Lemma D.10).

1411 Nevertheless, as noted in the main text, one can still employ the silver step-size in the convex, smooth
 1412 setting by combining it with the restarting technique of [OC15]. Theorem 4.2 shows that applying
 1413 the restarting method [OC15] to our silver step-size RGD yields an algorithm that also applies to
 1414 geodesically strongly convex problems.

1415 *Proof of Theorem 4.2.* Since f is geodesically α -strongly convex,

$$f(x_m) - f(x_*) \geq \frac{\alpha}{2} d^2(x_m, x_*)$$

1416 from the geodesic strong convexity and stationarity condition.

1417 Therefore, for $m = 2^k - 1$, one gets

$$d^2(x_m, x_*) \leq \frac{2}{\alpha} (f(x_m) - f(x_*)) \leq 2\kappa r_k d^2(x_0, x_*).$$

1418 Now, we iterate this algorithm, *i.e.*, $m = 2^k - 1$ silver step-size gradient descent, ℓ times, by restarting
 1419 the algorithm from the very last update of the previous runs. The total number of iterations becomes
 1420 $n = m\ell = (2^k - 1)\ell$. Then, one gets the following bound for n number of iterations:

$$d^2(x_n, x_*) \leq (2\kappa r_k)^\ell d^2(x_0, x_*).$$

1421 The term $(2\kappa r_k)^\ell$ is the rate we obtain for this algorithm. Now, one can optimize the choice of k, ℓ to
 1422 get the tightest convergence rate, by solving

$$\min_{\ell, k} (2\kappa r_k)^\ell \quad \text{given} \quad (2^k - 1)\ell = n.$$

Specifically, we plug-in $k^* = \lceil \log_\rho \kappa \rceil + 1$. Observe $\rho^{k^*} + 1 \geq 1 + \rho^{\log_\rho \kappa} = 1 + \rho\kappa \geq \rho\kappa$. Then,

$$2\kappa r_{k^*} = \frac{2\kappa}{1 + \sqrt{4\rho^{2k^*} - 3}} \leq \frac{2\kappa}{\rho^{k^*} + 1} \leq \frac{2}{\rho} < 1$$

Now, since $\ell = \frac{n}{2^{k^*} - 1}$,

$$(2\kappa r_{k^*})^\ell = \exp(\ell \log(2\kappa r_{k^*})) \leq \exp\left(\left(\log \frac{2}{\rho}\right) \frac{n}{2^{k^*} - 1}\right) \leq \exp\left(-\left(\log \frac{\rho}{2}\right) \frac{n}{\kappa^{\log_\rho 2}}\right)$$

which is the claimed rate.

For the ϵ -approximate error, $d^2(x_n, x_*) \leq \epsilon$ holds whenever

$$\exp\left(-\left(\log \frac{\rho}{2}\right) \frac{n}{\kappa^{\log_\rho 2}}\right) d^2(x_0, x_*) \leq \epsilon.$$

This is equivalent to

$$n \geq \frac{\kappa^{\log_\rho 2}}{\log(\rho/2)} \log \frac{d^2(x_0, x_*)}{\epsilon} = \Theta(\kappa^{\log_\rho 2} \log(1/\epsilon)).$$

This completes the proof.

■

D.3 Analysis on possibly negatively curved manifolds

For the last theoretical part of the paper, we provide a heuristic reasoning why our silver step-size analyses do not directly extend to possibly negatively curved spaces.

To this end, we drop the non-negative curvature assumption, and take N to be a geodesically convex subset of M with the sectional curvature lower bound $K_{\min} > -\infty$ and diameter bound $\text{diam}(N) = D < \infty$. We define the K_{\min} related constant ζ , which is 1 if $K_{\min} \geq 0$ and $\sqrt{-K_{\min}} D \coth(\sqrt{-K_{\min}} D) \geq 1$ otherwise. Then, Lemma C.1 in fact admits more general formulation in terms of ζ .

Lemma D.11. *For any $x_n, x_{n+1}, x_* \in N$, one has*

$$\|\log_{x_{n+1}} x_*\|^2 \leq \zeta \|\log_{x_n} x_*\|^2 + \|\log_{x_n} x_{n+1}\|^2 - 2 \langle \log_{x_n} x_{n+1}, \log_{x_n} x_* \rangle.$$

Proof. The proof is exactly same as Lemma C.1, except keeping ζ . ■

Note Lemma C.1 is a special case of this result. If one tries to apply the same method as in our analysis, the best inequality one can achieve is something like this (assuming $L = 1$):

$$f(x_n) - f(x_*) \leq r_k \zeta^n \|\log_{x_0} x_*\|^2.$$

The reason is as follows: Since we now need to repeatedly use Lemma D.11 for Lemma 5.1, in this general case one would get

$$\begin{aligned} & \zeta^n \|\log_{x_0} x_*\|^2 - \left\| \log_{x_n} x_* + \frac{1}{2r_k} \text{Grad } f(x_n) \right\|^2 + \frac{f(x_*) - f(x_n)}{r_k} \\ &= \zeta^n \|\log_{x_0} x_*\|^2 - \|\log_{x_n} x_*\|^2 - \frac{1}{4r_k^2} \|\text{Grad } f(x_n)\|^2 - \frac{1}{r_k} \langle \log_{x_n} x_*, \text{Grad } f(x_n) \rangle + \frac{f(x_*) - f(x_n)}{r_k} \\ &\geq \frac{f(x_*) - f(x_n)}{r_k} - \frac{1}{4r_k^2} \|\text{Grad } f(x_n)\|^2 - \frac{1}{r_k} \langle \log_{x_n} x_*, \text{Grad } f(x_n) \rangle \\ &\quad + \zeta^n \|\log_{x_0} x_*\|^2 - \zeta \|\log_{x_{n-1}} x_*\|^2 - \|\log_{x_n} x_{n+1}\|^2 + 2 \langle \log_{x_n} x_{n+1}, \log_{x_n} x_* \rangle \end{aligned}$$

$$\begin{aligned}
&= \frac{f(x_*) - f(x_n)}{r_k} - \frac{1}{4r_k^2} \|\text{Grad } f(x_n)\|^2 - \frac{1}{r_k} \langle \log_{x_n} x_*, \text{Grad } f(x_n) \rangle \\
&\quad + \zeta^n \|\log_{x_0} x_*\|^2 - \zeta \|\log_{x_{n-1}} x_*\|^2 - \eta_{n-1}^2 \|\text{Grad } f(x_{n-1})\|^2 - 2\eta_{n-1} \langle \text{Grad } f(x_{n-1}), \log_{x_{n-1}} x_* \rangle \\
&\geq \frac{f(x_*) - f(x_n)}{r_k} - \frac{1}{4r_k^2} \|\text{Grad } f(x_n)\|^2 - \frac{1}{r_k} \langle \log_{x_n} x_*, \text{Grad } f(x_n) \rangle \\
&\quad + \zeta^n \|\log_{x_0} x_*\|^2 - \eta_{n-1}^2 \|\text{Grad } f(x_{n-1})\|^2 - 2\eta_{n-1} \langle \text{Grad } f(x_{n-1}), \log_{x_{n-1}} x_* \rangle \\
&\quad - \zeta \left(\zeta \|\log_{x_{n-2}} x_*\|^2 + \eta_{n-2}^2 \|\text{Grad } f(x_{n-2})\|^2 + 2\eta_{n-2} \langle \text{Grad } f(x_{n-2}), \log_{x_{n-2}} x_* \rangle \right) \\
&= \frac{f(x_*) - f(x_n)}{r_k} - \frac{1}{4r_k^2} \|\text{Grad } f(x_n)\|^2 - \frac{1}{r_k} \langle \log_{x_n} x_*, \text{Grad } f(x_n) \rangle \\
&\quad + \zeta^n \|\log_{x_0} x_*\|^2 - \eta_n^2 \|\text{Grad } f(x_n)\|^2 - 2\eta_n \langle \text{Grad } f(x_n), \log_{x_n} x_* \rangle \\
&\quad - \zeta^2 \|\log_{x_{n-2}} x_*\|^2 - \zeta \eta_{n-1}^2 \|\text{Grad } f(x_{n-1})\|^2 - 2\zeta \eta_{n-1} \langle \text{Grad } f(x_{n-1}), \log_{x_{n-1}} x_* \rangle \\
&\geq \dots \\
&\geq \frac{f(x_*) - f(x_n)}{r_k} - \frac{1}{4r_k^2} \|\text{Grad } f(x_n)\|^2 - \frac{1}{r_k} \langle \log_{x_n} x_*, \text{Grad } f(x_n) \rangle \\
&\quad - \sum_{i=i}^n \zeta^{i-1} \eta_{n-i}^2 \|\text{Grad } f(x_{n-i})\|^2 - 2 \sum_{i=1}^n \zeta^{i-1} \eta_{n-i} \langle \text{Grad } f(x_{n-i}), \log_{x_{n-i}} x_* \rangle \\
&\quad + \zeta^n \|\log_{x_0} x_*\|^2 - \zeta^n \|\log_{x_0} x_*\|^2 \\
&= \frac{f(x_*) - f(x_n)}{r_k} - \frac{1}{4r_k^2} \|\text{Grad } f(x_n)\|^2 - \frac{1}{r_k} \langle \log_{x_n} x_*, \text{Grad } f(x_n) \rangle \\
&\quad - \sum_{i=i}^n \zeta^{i-1} \eta_{n-i}^2 \|\text{Grad } f(x_{n-i})\|^2 - 2 \sum_{i=1}^n \zeta^{i-1} \eta_{n-i} \langle \text{Grad } f(x_{n-i}), \log_{x_{n-i}} x_* \rangle.
\end{aligned}$$

1444 Therefore, using the same approach, one can only get up to

$$\sum_{ij} \lambda_{ij} Q_{ij} \leq \zeta^n d^2(x_0, x_*) - \left\| \log_{x_n} x_* + \frac{1}{2r_k} \text{Grad } f(x_n) \right\|^2 + \frac{f(x_*) - f(x_n)}{r_k}.$$

1445 Note we are missing one more ingredient here: non-negativeness of λ_{ij} is no longer guaranteed and
1446 should depend on ζ . That said, even if one assumes non-negative coefficients, one only gets

$$f(x_n) - f(x_*) \leq r_k \zeta^n d^2(x_0, x_*).$$

1447 If the space admits a negative curvature, then $\zeta > 1$, so that $\zeta^n r_k \rightarrow \infty$.

1448 This makes sense intuitively at least; if one has a negative curvature, the gradient update changes
1449 more rapidly for the small changes of the step-size. Therefore, if one takes very large step-sizes (as
1450 in silver step-size), its effect to the update is harder to control under the negative curvature.

1451 E Implementation detail and additional experiments

1452 This section includes implementation detail and more experiments of our algorithm under different
1453 settings. We conduct additional experiments on the problems in Section 6, to show the robustness
1454 of our algorithm. In particular, in this appendix we elaborate the following points that were briefly
1455 mentioned in the main body.

- 1456 1. We show the number of step-size needs not be in the form of $n = 2^k - 1$, by numerically
1457 showing our algorithm works under other choices of $n \neq 2^k - 1$.
- 1458 2. Because the silver step-size schedule sometimes uses very large step-sizes, one might ask
1459 whether simply increasing RGD's constant step-size could match its performance. We show

1460 this is not the case: using a constant step-size above the critical threshold $2/L$ causes RGD
 1461 to diverge, while silver step-size shows the improved performance.

1462 3. We conducted experiments using multiple random seeds and demonstrate that our algorithm’s
 1463 performances are statistically significant.

1464 Furthermore, to demonstrate our method’s versatility, we include experiments on one additional
 1465 optimization problem in the Wasserstein space: the mean-field training of a two-layer neural network.
 1466 This problem showcases the applicability of our algorithm, and of Wasserstein-based optimization
 1467 more broadly, to neural network training.

1468 E.1 Implementation detail

1469 All experiments in our paper were conducted on the free version of Google Colab using a T4 GPU.
 1470 Each task took no more than 5 minutes.

1471 **Wasserstein potential functional optimization** For the potential functional optimization problem
 1472 in Section 6.1, we used Python packages `numpy`, `scipy` for the implementation. We generated
 1473 m_* from the uniform distribution on the unit cube $[0, 1]^d$. For Σ_* , since we conducted experiments
 1474 with fixed $L = 1$ and $\alpha = 10^{-1}, 10^{-3}, 10^{-7}, 10^{-13}$, we have $\lambda_{\min} = 1/L = 1$ and $\lambda_{\max} = 1/\alpha$.
 1475 We placed d points evenly on a log-scale over the interval $[1/L, 1/\alpha]$ and used those values as
 1476 the eigenvalues to construct a diagonal matrix Λ . Then, we uniformly sampled an orthogonal
 1477 matrix P from the uniform distribution on the orthogonal group $O(d)$ (using Haar measure), and set
 1478 $\Sigma_* = P\Lambda P^T$. We used $m_0 = 0$ and $\Sigma_0 = I$ as the initialization for all experiments.

1479 **Rayleigh quotient maximization** We used the package `pymanopt` [TKW16] to model the spherical
 1480 data and compute geometric quantities. As mentioned in our main body, we conduct experiments on
 1481 two cases of H : (1) $H = \frac{1}{2}(A + A^T)$ where the entries of A are randomly generated from $N(0, 1/d)$
 1482 as in [KY22] (corresponding to small eigenvalue gaps); and (2) a randomly generated symmetric
 1483 matrix with $\lambda_{\max} = d$ and $\lambda_{\min} = -d$ (corresponding to large eigenvalue gaps). In the second
 1484 case, we reused the code for generating Σ_* in the Wasserstein potential optimization problem, but
 1485 generated eigenvalues at $d/2$ points evenly spaced on a log-scale over $[-d, -1]$ and the other $d/2$
 1486 over $[1, d]$. We excluded the interval $(-1, 1)$ to avoid some eigenvalues being close to 0. We used the
 1487 uniform random initialization on the sphere for all experiments.

1488 E.2 Additional experiments

1489 E.2.1 Potential functional optimization

1490 We solve the same task as in Section 6.1. To verify that our algorithm remains effective with a general
 1491 choice of iteration count, we set the number of iterations $n = 1500$, which is neither of the form
 1492 $2^k - 1$ nor close to $2^{10} - 1$ or $2^{11} - 1$. For the inner-iterations in the strongly convex setting for
 1493 the restarting, we chose $m = 20$ for $\alpha = 10^{-1}$ and $m = 500$ for $\alpha = 10^{-3}$, selecting values near
 1494 the $2^{k^*} - 1$ in Theorem 4.2 while ensuring divisibility by 1,500. We compared our silver step-size
 1495 RGD with constant step-size RGD using $\eta = 1/L$ (the standard choice), $\eta = 1.99/L$ (just below
 1496 the theoretical threshold), and $\eta = 2.01/L$ (just above it). The experiment was repeated over 100
 1497 random seeds, and we report the mean error curves along with 95% confidence intervals. Here, using
 1498 different seeds can be understood as solving instances of a stochastic optimization problem. In this
 1499 regard, comparing the errors across different seeds is a reasonable evaluation.

1500 The results are displayed in Figure 4. Figure 4 provides evidence supporting our claims:

- 1501 1. The algorithm performs well even when the number of iterations is not of the form $2^k - 1$.
- 1502 2. Our method is not equivalent to simply increasing the constant step-size in RGD; it consis-
- 1503 tently outperforms all tested step-size choices. In particular, the large step-size RGD, unlike
- 1504 silver step-size RGD, diverges.

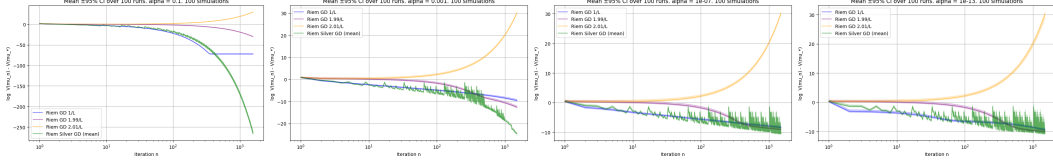


Figure 4: Comparison between silver step-size method and RGD for potential functional optimization in $BW(\mathbb{R}^d)$ with different convexity parameters. For each task, we conduct 100 simulations with different seeds and plot the mean and 95% confidence interval of the error over the iterates. **Columns:** From left to right, each column corresponds to $\kappa = 10^1, 10^3, 10^7, 10^{13}$.

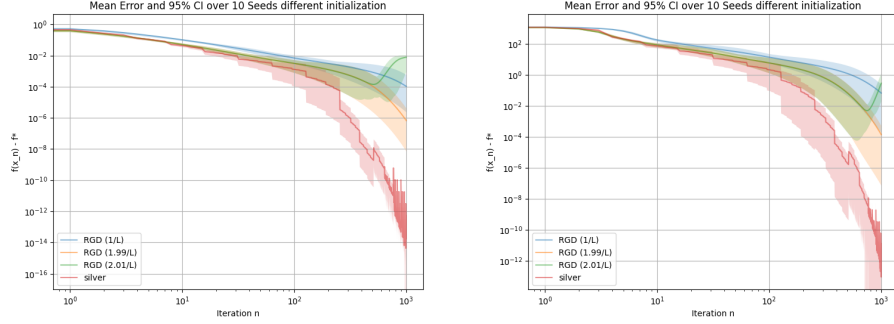


Figure 5: Comparison between silver step-size method and RGD for Rayleigh quotient maximization problem on \mathbb{S}^{2500} . For each task, we conduct 10 simulations with different seeds and plot the mean and 95% confidence interval of the error over the iterates. **Left:** H with small eigenvalue gaps. **Right:** H with large eigenvalue gaps.

3. The performances of our algorithm are statistically significant.

E.2.2 Rayleigh quotient maximization

As in Appendix E.2.1, we conduct additional experiments on Rayleigh quotient maximization problem under similar settings: using multiple random seeds, setting the number of iterations to a value not of the form $2^k - 1$, and comparing our method with RGD using various constant step-size choices. Due to the higher computational cost compared to the Wasserstein potential experiments, we fix the target matrix H , and conduct experiments with 10 different random seeds, varying only the initialization. We also reduce the number of iterations to $n = 1000$ (still not of the form $2^k - 1$). The step-size choices remain the same: $\eta = 1/L, 1.99/L$, and $2.01/L$.

The results are summarized in Figure 5. Figure 5 again validates the points discussed in the main text.

E.2.3 Mean-Field Two-Layer Network Training via Wasserstein gradient

Finally, we numerically demonstrate the effectiveness of our algorithms for two-layer neural network training. We first introduce the mean-field training formulation for a two-layer neural network, which enables us to view neural network training as a Wasserstein optimization problem, and then present our experimental results. For further details, we refer the interested reader to [CB18, MMN18, Woj20, FRF22].

Problem formulation One way to interpret two-layer neural networks is to view their function space as a space of probability measures. In particular, we adopt the Barron space formulation studied in [Bar93, WE20, Woj20]. In Barron space formulation, a (possibly infinitely wide) two-layer neural network is represented as

$$f_\pi(x) := \mathbb{E}_{(a,w,b) \sim \pi} [a\sigma(w^T x + b)]$$

where σ denoting a fixed activation function (e.g., ReLU). For instance, a m -width two-layer neural network corresponds to f_{π_m} , where $\pi_m = \frac{1}{m} \sum_{i=1}^m \delta_{(a_i, w_i, b_i)}$.

This formulation enables us to view neural network training as an optimization over probability measures. In particular, it becomes the following risk-functional minimization problem:

$$\pi_* := \operatorname{argmin}_{\pi \in \mathcal{P}_{2,ac}(\mathbb{R}^d)} R(\pi) := \mathbb{E}_{x \sim \mathbb{P}} [\ell(f_\pi(x), f^*(x))] \quad (\text{E.1})$$

where f^* is the target function, f_π is the two-layer neural network, and ℓ is a loss function (e.g., squared loss). The neural network f_{π_*} is the risk-functional minimizer and thus the desired solution. Since (E.1) is now just the optimization problem on the Wasserstein space, it is possible to consider Wasserstein gradient descent algorithms (6.2) to solve (E.1):

$$\pi_{n+1} = (id - \eta_n \operatorname{Grad}_{W_2} R(\pi_n))_{\#} \pi_n. \quad (\text{E.2})$$

In practice, this update operates over the space of functions and is thus not directly implementable. Instead, one typically uses a particle approximation of the probability measure, *i.e.*,

$$\pi_n = \frac{1}{m} \sum_{i=1}^m \delta_{(a_i^{(n)}, w_i^{(n)}, b_i^{(n)})},$$

where m is the number of particles chosen by the user [SKL20, WL22]. Under this approximation, the Wasserstein gradient update becomes

$$\begin{aligned} \pi_{n+1} &= (id - \eta_n \operatorname{Grad}_{W_2} R(\pi_n))_{\#} \pi_n \\ &= \frac{1}{m} \sum_{i=1}^m \delta_{(a_i^{(n)}, w_i^{(n)}, b_i^{(n)}) - \eta_n \operatorname{Grad}_{W_2} R(\pi_n)(a_i^{(n)}, w_i^{(n)}, b_i^{(n)})}. \end{aligned}$$

Using Definition A.34, it is known from [Woj20] that

$$\operatorname{Grad}_{W_2} R(\pi)(a, w, b) = \mathbb{E}_{x \sim \mathbb{P}} [\nabla_{(a,w,b)} \ell(f_\pi(x), f^*(x))].$$

Therefore, the particle approximation of the Wasserstein gradient update for a two-layer neural network takes the form

$$(a_i^{(n+1)}, w_i^{(n+1)}, b_i^{(n+1)}) = (a_i^{(n)}, w_i^{(n)}, b_i^{(n)}) - \eta_n \mathbb{E}_{x \sim \mathbb{P}} [\nabla_{(a_i^{(n)}, w_i^{(n)}, b_i^{(n)})} \ell(f_{\pi_n}(x), f^*(x))] \quad (\text{E.3})$$

for $i = 1, \dots, m$. Observe (E.3) exactly coincides with the standard gradient descent update of the parameters.

In conclusion, the silver step-size (and, respectively, constant step-size) parameter updates in two-layer neural networks (E.3) can be interpreted as the particle approximation of silver step-size (*resp.* constant step-size) Wasserstein gradient descent (E.2) applied to the risk minimization problem (E.1).

Numerical experiments To evaluate the effectiveness of the silver step-size for this task, we conduct experiments on learning a target function using a two-layer neural network with ReLU activation. Specifically, we consider the simple task of learning a univariate function $f^* : [-1, 1] \rightarrow \mathbb{R}$. We consider two target functions:

1. $f^*(x) = \frac{1}{30} \sum_{i=1}^{30} a_i^* \sigma(w_i^* x + b_i^*)$, *i.e.*, a 30-width two-layer neural network with fixed parameters a_i^*, w_i^*, b_i^* . Here, σ is the ReLU activation.
2. $f^*(x) = \sin(2\pi x)$.

We use $N = 200$ samples, with 70% of the data used for training and the remaining 30% for testing. The model is a two-layer neural network with width $m = 100$, trained using mean squared loss. We set the smoothness parameter to $L = 100$, and the number of training iterations to $n = 2000$.

Figure 6 shows the results of our experiments for solving (E.3) using different step-size schedules. Consistent with previous findings, the silver step-size algorithm outperforms constant step-size RGDs with various step-sizes in solving (E.1). While the figure displays results for a specific random seed, we observed similar trends across multiple seeds.

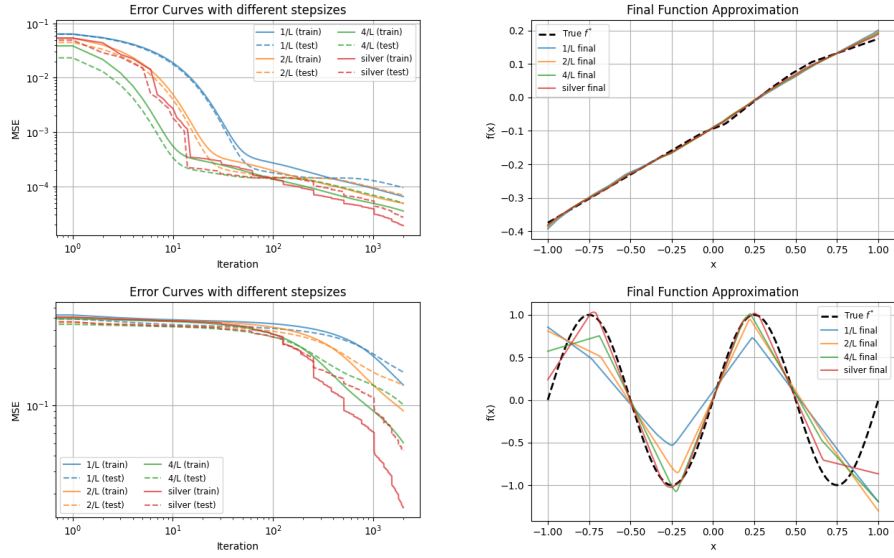


Figure 6: Mean-field training (E.3) of two-layer neural networks. **Rows:** The first row is the results from $f^*(x) = \frac{1}{30} \sum_{i=1}^{30} a_i^* \sigma(w_i^* x + b^*)$, and the second row is the results from $f^*(x) = \sin(2\pi x)$. **Columns:** The first column is the training and test error curve, and the second column is the function graph of the learned function.