Infinite-Dimensional Game Optimization via Variational Transport

Abstract

Game optimization has been extensively studied when decision variables lie in a finite-dimensional space, of which solutions correspond to pure strategies at the Nash equilibrium (NE) and the gradient descent-ascent (GDA) method works widely in practice. In this paper, we consider infinite-dimensional games by a zero-sum distributional optimization problem over a space of probability measures defined on a continuous variable set, which is inspired by finding a mixed NE for finite-dimensional games. We then aim to answer a natural question: *Will GDA-type algorithms still be provably efficient when extended to infinite-dimensional games?* To answer this question, we propose a particle-based variational transport algorithm based on GDA in the functional spaces. Specifically, the algorithm performs multi-step functional gradient descent-ascent in the Wasserstein space via alternately pushing two sets of particles in the variable space. By characterizing the gradient estimation error from variational form maximization and the convergence behavior of each player with different objective landscapes, we prove rigorously that the generalized GDA algorithm converges to the NE or the value of the game efficiently for a class of games under the Polyak-Łojasiewicz (PL) condition. To conclude, we provide complete statistical and convergence guarantees for solving an infinite-dimensional zero-sum game via a provably efficient particle-based method under mild conditions. Additionally, our work provides the first thorough statistical analysis for the particle-based algorithm to learn an objective functional with a variational form using universal approximators (*i.e.*, neural networks (NNs)), which may be of independent interest.

1 Introduction

Recent years have witnessed a resurgence in zero-sum games for machine learning applications, where two players' strategies are usually parameterized with two finite-dimensional decision variables. Furthermore, the optimal strategies define the *pure* NE in the sense that they identify two deterministic strategies. Motivating examples include generative adversarial networks (GANs) [22, 46, 56, 58], reinforcement learning [13, 28], distributionally robust optimization (DRO) [21, 61], and learning exponential families [12], among others. Such zero-sum games have been extensively analyzed in convex-concave settings, where a global Nash equilibrium (NE) can be computed efficiently by gradient descent-ascent (GDA) type algorithms [18, 27, 42, 43]. Nonetheless, these methods mostly stagger in nonconvex landscapes for modern applications, for which a bunch of modified gradient-based methods [31, 56, 57] under different settings relax nonconvexity to find a first-order NE [49]. Despite the popularity in exploring variants of the pure NE in nonconvex-nonconcave game settings, another crucial problem arises from a game-theoretic perspective: *What if the pure NE does not exist [2, 31]?* The finite-dimensional formulation naturally excludes a potential better or even the only existential optimal mixed strategies, and meanwhile is restricted to local convergence without convexity.

To alleviate the concern above and to further understand the difficulty at the boundary of contemporary game optimization, we consider a class of zero-sum infinite-dimensional games where each decision variable is a probability measure representing the mixed strategies over the spaces of pure strategies. In addition, we

assume this distributional games to satisfy Riemannian Polyak-Łojasiewicz (PL) conditions and smoothness assumptions, which are satisfied by a range of nonconvex landscapes and the practical training objectives for GANs with regularization [2]. A natural approach to distributional optimization problems is the particle-based method [52, 64, 67], where stochastic gradient Langevin dynamics (SGLD) is usually adopted to draw a sample from the desired distribution via discretization of stochastic differential equations [30]. However, SGLD sampling is quite inefficient for reaching a stationary distribution at each step. Meanwhile, from the view of games, GDA-type algorithms have not been studied in full generality for infinite-dimensional settings. Motivated by the two facts above, we adapt the multi-step GDA-type algorithm to infinite-dimensional games through particle-based approximation, and provide the first set of theoretical guarantees by analyzing its new behavior under infinite-dimensional settings.

We conclude our contributions as follows. (1) To model the mixed NE of finite-dimensional games, we introduce the generic infinite-dimensional zero-sum games. We establish the GDA-type algorithm in the Wasserstein space, also named variational transport for infinite-dimensional games (VTIG), for such games via Riemannian gradient propositions (Proposition 3.1 and 3.2). (2) We provide the first thorough analysis of both statistical and optimization errors for VTIG in two scenarios. One is the convergence to the first-order NE under a Riemannian PL condition (Theorem 4.3), and the other is the convergence to the minimax value under a stronger two-sided PL condition (Theorem 4.5). (3) As a technical component, we provide statistical analysis for particle-based gradient estimation via a reverse Poincaré inequality, which upper bounds the ℓ_p-norm of the gradient by the ℓ_p-norm of the function for $p \geq 1$.

Related work. Finite-dimensional games under convex-concave settings [27, 32, 42, 43] are adequately studied with corresponding monotonic variational inequalities [14, 22] and solved by GDA [60]. Meanwhile, primal-dual schemes and negative momentum [7, 16, 23] are proposed to help GDA on convergence, which bypasses cyclic dynamics [15, 39, 40]. To tame nonconvexity, [31] proves the $\mathcal{O}(\theta^{-4})$ rate in gradient evaluations is required in the convergence to an θ-first order NE with Max-oracle; [37, 38] reached the same rate when the objective is concave *w.r.t.* the max-player strategy; improved rates of $\mathcal{O}(\theta^{-3.5})$ and $\mathcal{O}(\theta^{-2})$ are shown in [57] under PL-game conditions, which is similar to our setting. However, our results are derived for infinite dimensions as a mixed-strategy extension.

In machine learning literature, the notion of mixed NE for GANs was originally presented in [25] without an algorithm to find it. A line of work [2, 26, 30, 47] seeks to further understand and find mixed NEs of GANs. Nonetheless, the existing algorithm in [30] using SGLD is computationally demanding at each step and complicated in the idea of algorithm design without statistical analysis. Our analysis extends the GDA-type algorithm to the Wasserstein space, and shows the existence of a provably efficient particle-based algorithm that pushes a fix-sized set of particles instead of running SGLD repeatedly.

Optimizing functionals of probability measures was studied by Frank-Wolfe [20] and steepest descent algorithms [41] in earlier times. More recently, descent methods in the space of probability measures [19, 53] are getting popular in machine learning, where particle-based methods [8, 35] approximate measures for practical implementation. Similarly, two sets of particles in our algorithm also provide the Dirac measure approximation for probability measures.

Notations. We denote by $[n]$ the set of integers $\{0, 1, ..., n\}$. Let $\mathcal{C}(\mathbb{R}^d)$ be the set of continuous functions over the d-dimensional real space $\mathbb{R}^d$. Let $\mathcal{X}$ be a convex compact set in $\mathbb{R}^d$. Given a nonnegative measure μ on $\mathcal{X}$, we define the ℓ_p-norm of the function $f \in \mathcal{C}(\mathbb{R}^d)$ on $\mathcal{X}$ as $\|f\|_{L^p_\mu(\mathcal{X})} = (\int_{\mathcal{X}} |f|^p d\mu)^{1/p}$. Let $\mathcal{P}(\mathcal{X})$ denote the collection of all Borel probability measures on the measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, where $\mathcal{B}(\mathcal{X})$ is the Borel σ-algebra on $\mathcal{X}$. We denote by $\mathcal{P}_2(\mathcal{X}) \subseteq \mathcal{P}(\mathcal{X})$ the set of Borel probability measures with finite second moments. We define the metric space $(M, \|\cdot\|)$ by a vector space M and a metric induced by the norm $\|\cdot\|$.

2 Problem Formulation and Optimization over Wasserstein Spaces

Below we state the formulation and assumptions for infinite-dimensional games in the Wasserstein space.

2.1 From Finite-Dimensional to Infinite-Dimensional Games

Consider the classical formulation of a two-player zero-sum game as folows,

$$\min_{x^\mu \in \mathcal{X}_\mu} \max_{x^\nu \in \mathcal{X}_\nu} f(x^\mu, x^\nu), \tag{2.1}$$

where $\mathcal{X}_\mu, \mathcal{X}_\nu \subseteq \mathbb{R}^d$ with $d > 0$ are convex compact sets of pure strategies as d-dimensional vectors with periodic or zero-flux boundary conditions *w.r.t.* the vector fields specified in Proposition 3.2, and f is the objective function. In nonconvex-nonconcave regimes, as finding local Nash equilibria is NP-hard or even impossible [31], a weaker notion of *first-order* NE (FNE) for a pair $(x_*^\mu, x_*^\nu) \in \mathcal{X}_\mu \times \mathcal{X}_\nu$ is defined as

$$\begin{aligned} &\langle \nabla_{x^\mu} f(x_*^\mu, x_*^\nu), x^\mu - x_*^\mu \rangle \geq 0, \\ &\langle \nabla_{x^\nu} f(x_*^\mu, x_*^\nu), x^\nu - x_*^\nu \rangle \leq 0, \quad \forall x^\mu \in \mathcal{X}_\mu,\ \forall x^\nu \in \mathcal{X}_\nu, \end{aligned} \tag{2.2}$$

which corresponds to first-order necessary optimality conditions. Observing that without a probability representation (2.1) only admits pure Nash strategies, we lift (2.1) by considering distributions over $\mathcal{X}_\mu$ and $\mathcal{X}_\nu$ to allow mixed strategies. The infinite-dimensional distributional two-player zero-sum game is defined as

$$\min_{\mu \in \mathcal{M}(\mathcal{X}_\mu)} \max_{\nu \in \mathcal{M}(\mathcal{X}_\nu)} F(\mu, \nu). \tag{2.3}$$

Here $F\colon \mathcal{M}(\mathcal{X}_\mu) \times \mathcal{M}(\mathcal{X}_\nu) \to \mathbb{R}$ is the objective functional. Also, $\mathcal{M}(\mathcal{X}) = (\mathcal{P}_2(\mathcal{X}), \mathcal{W}_2)$ is the Wasserstein (W_2-) space, an infinite-dimensional manifold by [63], with the W_2-distance on $\mathcal{P}_2(\mathcal{X})$ defined as

$$\mathcal{W}_2(\mu, \nu) = \inf\Big\{ \mathbb{E}\big[||X - Y||^2\big]^{1/2} \mid \mathcal{L}(X) = \mu, \mathcal{L}(Y) = \nu \Big\},$$

where the infimum is taken over the random variables X and Y, and we denote by $\mathcal{L}(X)$ the law of a random variable X. Without loss of generality, we set $\mathcal{X}_\mu = \mathcal{X}_\nu = \mathcal{X}$ and write $\mathcal{M} = \mathcal{M}(\mathcal{X})$. Without specification, the domain of an integral is the set $\mathcal{X}$. We refer to the two players as player μ and player ν, respectively. To further characterize the properties of such a Wasserstein space $\mathcal{M}$, we introduce geodesics, tangent vectors and tangent spaces below.

Definition 2.1. Define the smooth curve $\gamma : [0, 1] \to \mathcal{P}_2(\mathcal{X})$. We call the curve γ a geodesic if there exists a constant $v \geq 0$ such that $\|\gamma(t_1) - \gamma(t_2)\| = v \cdot |t_1 - t_2|$ for any $t_1, t_2 \in [0, 1]$. A tangent vector at $\mu \in \mathcal{M}$ is an equivalence class of differentiable curves through μ with a prescribed velocity vector at μ. The tangent space at μ, denoted by $T_\mu\mathcal{M}$, consists of all tangent vectors at μ.

Furthermore, the manifold $\mathcal{M}$ is equipped with a weak Riemannian structure in the following sense [63]. Given any tangent vectors u, v at $\mu \in \mathcal{M}$ and the vector fields $\widetilde{u}, \widetilde{v}$ satisfying *continuity equations* $u = -\operatorname{div}(\mu\widetilde{u})$ and $v = -\operatorname{div}(\mu\widetilde{v})$, respectively, we define the inner product of u and v as $\langle u, v \rangle_\mu = \int \langle \widetilde{u}, \widetilde{v} \rangle \, \mathrm{d}\mu$, where $\langle \widetilde{u}, \widetilde{v} \rangle$ is the inner product in $\mathcal{X}$. Such a metric induces a norm $\|u\|_\mu = \langle u, u \rangle_\mu^{1/2}$ for any $u \in T_\mu\mathcal{M}$. Under such a structure, we define the directional derivative *w.r.t.* $u \in T_\mu\mathcal{M}$ of a differentiable functional $g\colon \mathcal{M} \to \mathbb{R}$ as $\nabla_v g(\mu) = \frac{\mathrm{d}}{\mathrm{d}t} g[\gamma(t)]\big|_{t=0}$, where $\gamma(0) = \mu \in \mathcal{M}$, $\gamma'(0) = u$. In addition, we say g is W_2-differentiable at μ if there exists $u' \in T_\mu\mathcal{M}$ such that $\nabla_u g(\mu) = \langle u', u \rangle_\mu$ for any $u \in T_\mu\mathcal{M}$, and write $\operatorname{grad} g(\mu) = u'$ as the (weak) Riemannian gradient of g at μ. The partial gradient $\operatorname{grad}_\mu F(\mu, \nu)$ is defined similarly for a functional $F\colon \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ when fixing ν. The exponential map at μ, denoted by Exp_μ, sends any $u \in T_\mu\mathcal{M}$ to $\mu' = \gamma_u(1)$[1], where γ_u is a geodesic such that $\gamma_u(0) = \mu$, $\gamma_u'(0) = u$. For any $\mu, \nu \in \mathcal{M}$, the parallel transport $\Gamma_\mu^\nu\colon T_\mu\mathcal{M} \to T_\nu\mathcal{M}$ is the map such that $\langle u, v \rangle_\mu = \langle \Gamma_\mu^\nu u, \Gamma_\mu^\nu v \rangle_\nu$ for any $u, v \in T_\mu\mathcal{M}$. Also, as $\mathcal{X}$ is separable and complete, $\mathcal{M}$ is geodesically complete [62] in the sense that the exponential map is defined on the whole tangent space. See §B for more formal definitions.

[1] Hence, for $\mu_1, \mu_2 \in \mathcal{M}$, $\operatorname{Exp}_{\mu_1}^{-1}(\mu_2)$ is an analogy to $x_2 - x_1$ for $x_1, x_2 \in \mathcal{X}$.

We also assume the objective functional F in (2.3) to satisfy the following variational forms,

$$F(\mu,\nu)=F_\nu(\mu)=\sup_{f\in\mathcal{F}}\left\{\int_{\mathcal{X}}f d\mu-F_\nu^*(f)\right\},\quad F(\mu,\nu)=F_\mu(\nu)=-\sup_{f\in\mathcal{F}}\left\{\int_{\mathcal{X}}f d\nu-F_\mu^*(f)\right\}, \tag{2.4}$$

where $\mathcal{F}$ is the family of square-integrable functions over $\mathcal{X}$, $F_\mu^*\colon\mathcal{F}\to\mathbb{R}$ and $F_\nu^*\colon\mathcal{F}\to\mathbb{R}$ are strongly convex and smooth functional *w.r.t.* the ℓ_2-norm. In fact, (2.4) follows from definitions of the conjugate function, and the example in § C.2 shows that a wide class of f-divergences admits such forms.

For theoretical analysis, we impose the following assumptions on the objective functional F.

Assumption 2.2. We assume that F is Lipschitz continuous and smooth *w.r.t.* the Wasserstein distance in the sense that

$$\begin{aligned}
&\left|F(\mu_1,\nu)-F(\mu_2,\nu)\right|\le L_\mu\mathcal{W}_2(\mu_1,\mu_2),\\
&\left|F(\mu,\nu_1)-F(\mu,\nu_2)\right|\le L_\nu\mathcal{W}_2(\nu_1,\nu_2),\\
&\boldsymbol{d}\big(\operatorname{grad}F_\nu(\mu_1),\operatorname{grad}F_\nu(\mu_2)\big)\le L_1\cdot\mathcal{W}_2(\mu_1,\mu_2),\quad \boldsymbol{d}\big(\operatorname{grad}F_\mu(\nu_1),\operatorname{grad}F_\mu(\nu_2)\big)\le L_2\cdot\mathcal{W}_2(\nu_1,\nu_2),\\
&\boldsymbol{d}\big(\operatorname{grad}F_{\mu_1}(\nu),\operatorname{grad}F_{\mu_2}(\nu)\big)\le L_0\cdot\mathcal{W}_2(\mu_1,\mu_2),\quad \boldsymbol{d}\big(\operatorname{grad}F_{\nu_1}(\mu),\operatorname{grad}F_{\nu_2}(\mu)\big)\le L_0\cdot\mathcal{W}_2(\nu_1,\nu_2)
\end{aligned} \tag{2.5}$$

for any $\mu,\mu_1,\mu_2,\nu,\nu_1,\nu_2\in\mathcal{M}$. Here L_μ,L_ν,L_1,L_2, and L_0 are absolute constants and $\boldsymbol{d}^2(u,v)=\left\langle u-\Gamma_\nu^\mu v,u-\Gamma_\nu^\mu v\right\rangle_\mu$ for any $\mu,\nu\in\mathcal{M}$, $u\in T_\mu\mathcal{M}$, and $v\in T_\nu\mathcal{M}$.

Assumption 2.2 is a natural extension of Lipschitz continuity and smoothness for Euclidean space to $\mathcal{M}$, where the Euclidean distance is replaced by $\mathcal{W}_2$. The following assumption extends the notion of PL condition, also known as gradient domination [44, 51, 57], to infinite-dimensional spaces.

Assumption 2.3. (Riemannian PL condition). A W_2-differentiable functional $g:\mathcal{M}\to\mathbb{R}$ with minimum value $g^*=\inf_{\mu\in\mathcal{M}}g(\mu)$ is called ξ-PL (ξ-gradient dominated) if for all $\mu\in\mathcal{M}$ we have

$$\langle\operatorname{grad}g(\mu),\operatorname{grad}g(\mu)\rangle_\mu\ge 2\xi\left(g(\mu)-g^*\right). \tag{2.6}$$

We call (2.3) a ξ-PL game, or simply a PL game, if $H_\mu(\nu)\triangleq-F(\mu,\nu)$ is ξ-PL. We assume (2.3) to be a ξ-PL game.

In particular, Assumption 2.3 implies that if the norm of the gradient is small at $\mu\in\mathcal{M}$, then the functional value at μ will be close to the optimum. In addition, it is not restrictive since a non-convex functional can still satisfy the PL condition [33]. To justify all the above assumptions, we provide the following example stemming from learning GANs, where the pure strategies in (2.3) correspond to parameters x^μ and x^ν of the GAN.

Example 2.4. Consider the mixed NE of WGANs [1] with Kullback-Leibler (KL) divergence regularization,

$$\min_{\mu\in\mathcal{M}}\max_{\nu\in\mathcal{M}}\mathbb{E}_{x^\nu\sim\nu}\mathbb{E}_{\zeta\sim\mathbb{P}_{\text{real}}}[h_{x^\nu}(\zeta)]-\mathbb{E}_{x^\nu\sim\nu}\mathbb{E}_{x^\mu\sim\mu}\mathbb{E}_{\zeta\sim\mathbb{P}_{x^\mu}}[h_{x^\nu}(\zeta)]-\mathrm{KL}(\nu\|\mu_0)+\mathrm{KL}(\mu\|\mu_0), \tag{2.7}$$

where $\mathrm{KL}(\mu\|\lambda)=\int_{\mathcal{X}}\log(\mathrm{d}\mu/\mathrm{d}\lambda)\,\mathrm{d}\mu$ with Lebesgue measures μ and λ, and μ_0 is the probability measure of standard Gaussian. Also, h_v denotes the discriminator parameterized by NNs, of which the input is $\zeta\in\mathcal{X}$. Without the expectations of x^μ and x^ν, (2.7) is reduced to the original regularized WGAN objective that admits only finite-dimensional pure Nash strategies. Further, we define the linear operator $D:\mathcal{M}\to\mathcal{F}$ by $(D\mu)(x^\nu)=\mathbb{E}_{x^\mu\sim\mu}\mathbb{E}_{\zeta\sim\mathbb{P}_{x^\mu}}[h_{x^\nu}(\zeta)]$ for any $x^\nu\in\mathcal{X}$ and some continuous function $h_{x^\nu}\in\mathcal{F}$. We also define $g(x^\nu)=\mathbb{E}_{\zeta\sim\mathbb{P}_{\text{real}}}[h_{x^\nu}(\zeta)]$. Then the objective F in (2.7) can be rewritten as $F(\mu,\nu)=\langle\nu,g\rangle-\langle\nu,D\mu\rangle-\mathrm{KL}(\nu\|\mu_0)+\mathrm{KL}(\mu\|\mu_0)$, It follows from the logarithmic Sobolev inequality (LSI) [48] in W_2-space that player μ meets the PL condition. Since the KL divergence is an f-divergence,

the variantional forms are guaranteed as follows,

$$F_\nu(\mu) = \sup_{f\in\mathcal{F}} \Big\{ \int f d\mu - \int \exp\big\{f(x^\mu) + \mathbb{E}_{x^\nu\sim\nu}\mathbb{E}_{\zeta\sim\mathbb{P}_{x^\mu}}[h_{x^\nu}(\zeta)]\big\}\, \mathrm{d}\mu_0(x^\mu) + \widehat{F}_\nu \Big\}, \tag{2.8}$$

$$F_\mu(\nu) = -\sup_{f\in\mathcal{F}} \Big\{ \int f d\nu - \int \exp\big\{f(x^\nu) + g(x^\nu) - (D\mu)(x^\nu)\big\}\, \mathrm{d}\mu_0(x^\nu) + \widehat{F}_\mu \Big\}. \tag{2.9}$$

Here $\widehat{F}_\nu = 1 - \mathrm{KL}(\nu\|\mu_0) + \mathbb{E}_{x^\nu\sim\nu}\mathbb{E}_{\zeta\sim\mathbb{P}_{\mathrm{real}}}[h_{x^\nu}(\zeta)]$ and $\widehat{F}_\mu = 1 - \mathrm{KL}(\mu\|\mu_0)$ are constants when fixing ν and μ, respectively. See §C.3 for details. We remark that in practical GAN training, KL regularization terms exist to prevent the mode collapse. More generally, the KL-regularized distributional bilinear game $\min_{\mu\in\mathcal{M}}\max_{\nu\in\mathcal{M}}\langle\nu, A\mu\rangle - \mathrm{KL}(\nu\|\mu_0) + \mathrm{KL}(\mu\|\mu_0)$ given a linear operator $A:\mathcal{M}\to\mathcal{F}$ is widely studied in games. Similarly, we write its variational forms as $F_\nu(\mu) = \sup_{f\in\mathcal{F}}\{\int f d\mu - \int \exp\big\{f(x^\nu) - A^*\nu(x^\nu)\big\}\,\mathrm{d}\mu_0(x^\nu) + 1 - \mathrm{KL}(\nu\|\mu_0)\}$ and $F_\mu(\nu) = -\sup_{f\in\mathcal{F}}\{\int f d\nu - \int \exp\big\{f(x^\mu) + A\mu(x^\mu)\big\}\,\mathrm{d}\mu_0(x^\mu) + 1 - \mathrm{KL}(\mu\|\mu_0)\}$, where A^* is the adjoint of A.

2.2 Measurement of Solutions

To quantify the accuracy of solutions to (2.3), we generalize the NE of finite-dimensional games to our infinite-dimensional distributional games. Given the numerical accuracy of iterative algorithms in practice, we define the notion of infinite-dimensional first-order NE (IFNE) as a performance measure.

Definition 2.5 (IFNE)**.** For any $\mu_1, \nu_1 \in \mathcal{M}$, we define

$$\mathcal{J}_\mu(\mu_1,\nu_1) \triangleq -\min_{\mathcal{W}_2(\mu,\mu_1)\le 1} \big\langle \mathrm{grad}_\mu F(\mu_1,\nu_1), \mathrm{Exp}_{\mu_1}^{-1}(\mu)\big\rangle_{\mu_1},$$

$$\mathcal{J}_\nu(\mu_1,\nu_1) \triangleq \max_{\mathcal{W}_2(\nu,\nu_1)\le 1} \big\langle \mathrm{grad}_\nu F(\mu_1,\nu_1), \mathrm{Exp}_{\nu_1}^{-1}(\nu)\big\rangle_{\nu_1}$$

as the first-order errors (FEs). Then a point $(\mu^*,\nu^*)\in\mathcal{M}\times\mathcal{M}$ is called a θ-IFNE of (2.3) if

$$\mathcal{J}_\mu(\mu^*,\nu^*) \le \theta \quad \text{and} \quad \mathcal{J}_\nu(\mu^*,\nu^*) \le \theta. \tag{2.10}$$

When $\theta = 0$, we call (μ^*,ν^*) an IFNE. Definition 2.5 characterizes how far the solutions are from the FNE in the W_2-space. Also, we characterize the upper bound θ in terms of the problem parameters for convergence rates in §4.

3 Variational Transport Algorithm for Infinite-Dimensional Games

In what follows, we introduce the variational transport algorithm to characterize GDA for the infinite-dimensional game defined in (2.3). Our idea is based on the multi-step GDA algorithm in [57] with nested loops, where multiple gradient ascent steps are run for estimating the gradient of the *inner maximization functional* defined as $G(\mu) = \max_{\nu\in\mathcal{M}} F(\mu,\nu)$ *w.r.t.* μ, which provides a descent direction for the outer minimization problem. Without specifying, statements below hold for both μ and ν although they are presented by $\mu\in\mathcal{M}$.

3.1 Gradient Descent beyond the Euclidean Space

We first show the connection between functional gradient descent in the Wasserstein Space $\mathcal{M}$ and transportation maps in the variable space $\mathcal{X}$. Specifically, we expect to update the current iterate $\mu\in\mathcal{M}$ of the gradient descent in the direction of $\mathrm{grad}\, F_\nu(\mu)$ along the geodesic. Therefore, in the ideal case, the gradient update is given by

$$\mu \leftarrow \mathrm{Exp}_\mu\big[-\eta\cdot\mathrm{grad}\, F_\nu(\mu)\big], \tag{3.1}$$

where $\eta > 0$ is the stepsize. The proposition below bridges the Riemannian gradient of a W_2-differentiable functional on $\mathcal{M}$ and its functional gradient *w.r.t.* the ℓ_2-norm. We denote by $f_\mu^*\in\mathcal{F}$ the optimal solution to (2.4) for $F_\nu(\mu)$.

Proposition 3.1 (Riemannian Gradients to Functional Gradients)**.** Let $F : \mathcal{M} \to \mathbb{R}$ be a W_2-differentiable functional, with its functional gradient *w.r.t.* the ℓ_2-norm written as $\delta F/\delta\mu$. Then, it follows that

$$\operatorname{grad} F(\mu) = -\operatorname{div}\left[\mu \cdot \nabla\left(\frac{\delta F}{\delta \mu}\right)\right], \tag{3.2}$$

where div is the divergence operator on $\mathcal{X}$. Furthermore, by the variational form of (2.3), we have $\delta F_\nu/\delta\mu = f_\mu^*$ and $\operatorname{grad} F_\nu(\mu) = -\operatorname{div}(\mu \cdot \nabla f_\mu^*)$.

Proof. See §C.1 for a detailed proof. □

By Proposition 3.1, to obtain a descent direction in W_2-space for $F_\nu(\mu)$, we first solve (2.4) for $f_\mu^* \in \mathcal{F}$ and then, compute the divergence in (3.2). Also, Exp_μ in (3.1) needs to be specified. As in practice we only have access to samples, or particles, from μ, we establish the proposition below to perform approximate gradient updates in (3.1) via particles.

Proposition 3.2 (Pushing particles as an exponential map)**.** For any $\mu \in \mathcal{M}$ and any $s \in T_\mu\mathcal{M}$, suppose the elliptic equation $-\operatorname{div}(\mu \cdot \nabla u) = s$ admits a unique solution $u\colon \mathcal{X} \to \mathbb{R}$ such that $\nabla u\colon \mathcal{X} \to \mathbb{R}^d$ is h-Lipschitz continuous. Then, for any $t \in [0, 1/h)$, we have

$$\left[\operatorname{Exp}_{\mathcal{X}}(t \cdot \nabla u)\right]_\sharp \mu = \operatorname{Exp}_\mu(t \cdot s), \tag{3.3}$$

where we use $\operatorname{Exp}_{\mathcal{X}}(t \cdot \nabla u)$ to denote the transportation map on $\mathcal{X}$ that sends any $x \in \mathcal{X}$ to a point $\operatorname{Exp}_x(t \cdot \nabla u(x)) \in \mathcal{X}$, which is also the exponential map over $\mathcal{X}$. We denote by $T_\sharp : \mathcal{P}_2(\mathcal{X}) \to \mathcal{P}_2(\mathcal{X})$ the push-forward map of a transportation map $T : \mathcal{X} \to \mathcal{X}$ such that for any $\mu \in \mathcal{M}$ and any measurable set $A \in \mathcal{X}$, we have $T_\sharp\mu(A) = \mu(T^{-1}(A))$.

Proof. See §C.2 for a detailed proof. □

Hence, if ∇f_μ^* is h-Lipschitz, by Proposition 3.1 and 3.2, for any $t \in [0, 1/h)$ we obtain $\operatorname{Exp}_\mu[-t \cdot \operatorname{grad} F(\mu)] = \left[\operatorname{Exp}_{\mathcal{X}}\left(-t \cdot \nabla f_\mu^*\right)\right]_\sharp \mu$. given $\mu \in \mathcal{M}$. This identifies the gradient descent update in the Wasserstein spaces with the push-forward map of probability measures over the Euclidean space, which can be approximated by pushing a set of particles. We illustrate such correspondence in Figure 1.

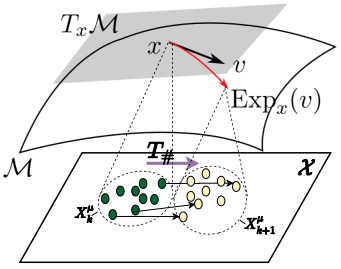

Figure 1: Equivalence between particle pushing in the Euclidean space $\mathcal{X}$ and the exponential map in the Wasserstein space $\mathcal{M}$. The tangent vector $v \in T_x\mathcal{M}$ at x induces the exponential map Exp_x and its correspondence in $\mathcal{X}$, the push-forward map $T_\sharp = \left[\operatorname{Exp}_{\mathcal{X}}\left(-t \cdot \nabla f_\mu^*\right)\right]_\sharp$. X_k^μ is the set of μ-particles at timestep k in Algorithm 1.

Further, we are left with the variational form maximization (VFM) problem in (2.4), where difficulties lie in the following aspects. (i) Firstly, our approach is expected to provide the reasonable statistical error incurred by estimating f_μ^* by $\widetilde{f}_\mu^*$ from the empirical version of VFM,

$$\widetilde{f}_\mu^* = \operatorname*{argmax}_{f \in \mathcal{F}} \left\{ \int_{\mathcal{X}} f \ \mathrm{d}\widehat{\mu} - F_\nu^*(f) \right\} = \operatorname*{argmax}_{f \in \mathcal{F}} \left\{ \frac{1}{N} \sum_{i=1}^N f(x_i) - F_\nu^*(f) \right\}, \tag{3.4}$$

where we replace μ in (2.4) by the empirical measure $\widehat{\mu} = 1/N \sum_{i=1}^N \delta_{x_i}$, *i.e.,* an average of Dirac measures over samples x_i's. (ii) Secondly, maximization over $\mathcal{F}$ is computationally intractable. To this end, we perform stochastic gradient descent (SGD) to learn f_μ^* from the following class $\widetilde{\mathcal{F}}$ of neural networks (NNs) instead of $\mathcal{F}$, which is a rich class by the universal approximation theorem [11, 29].

Neural Network Parametrization. We consider the following class of NNs,

$$\widetilde{\mathcal{F}} = \left\{ \widetilde{f} \,\middle|\, \widetilde{f}(x) = \frac{1}{\sqrt{w}} \sum_{i=1}^w b_i \cdot \sigma\left([\beta]_i^\top x\right) \right\}, \tag{3.5}$$

where w is the width of the neural network, $[\beta]_i \in \mathbb{R}^d$, $\beta = ([\beta]_1^\top, \cdots, [\beta]_w^\top)^\top \in \mathbb{R}^{wd}$ are input weights, $\sigma(\cdot)$ denotes a smooth activation function, and $b_i \in \{-1, 1\}$ $(i \in [w])$ are the output weights. As shown

in Algorithm 3, only β is updated during training while b_i ($i \in [w]$) is fixed. In addition, at each iteration we project the input weights β to an ℓ_2-ball centered at $\beta(0)$ with radius r_f defined as $\mathcal{B}^0(r_f) = \{\beta : \|\beta - \beta(0)\|_2 \le r_f\}$. See §D.1 for more details of $\widetilde{\mathcal{F}}$.

3.2 Algorithm for Two-Player Infinite-Dimensional Games

We now put together two nested loops of gradient descent/ascent updates approximated by particles as the variational transport algorithm for infinite-dimensional games (VTIG) in Algorithm 1. In detail, we maintain two sets of N_μ μ-particles and N_ν ν-particles for player μ and player ν. Also, VTIG output the corresponding probability measures approximated by these two sets as the solutions to (2.3), respectively. At outer-loop timestep k, we denote the set for player μ by $X_k^\mu = \{x_{i,k}^\mu\}_{i\in[N_\mu]}$ and the set for player ν at inner-loop timestep l of outer-loop timestep k by $X_l^\nu(\widetilde{\mu}_k) = \{x_{i,l}^\nu(\widetilde{\mu}_k)\}_{i\in[N_\nu]}$. Here we write $X_l^\nu(\widetilde{\mu}_k)$ and $x_{i,l}^\nu(\widetilde{\mu}_k)$ to emphasize that we fix X_k^μ (resp. $\widetilde{\mu}_k$) when updating X_l^ν (resp. $\widetilde{\nu}_l$) in Line 7 of Algorithm 1. Also, $\{\widetilde{\mu}_k\}_{k\ge0}$ and $\{\widetilde{\nu}_l(\widetilde{\mu}_k)\}_{k,l\ge0}$ are sequences of probability measures of $\{X_k^\mu\}_{k\ge0}$ and $\{X_l^\nu(\widetilde{\mu}_k)\}_{k,l\ge0}$ constructed implicitly by VTIG, which is specified later. Further, the set of μ-particles X_k^μ is updated as follows given X_0^μ for $k \ge 1$. At the outer-loop timestep k, VTIG computes the solution to (3.4) following Line 10 in Algorithm 1

$$\widetilde{f}_k^* \leftarrow \mathbf{VFM}\big(X_k^\mu, F^*_{\widetilde{\nu}_{k+1}}, N_\mu\big) \tag{3.6}$$

based on the current μ-particle set X_k^μ, the functional $F^*_{\widetilde{\nu}_{k+1}}$ defined in the variational form (2.4), and the number of μ-particles N_μ. As shown in Algorithm 3, the VFM problem is solved by learning a neural network $\widetilde{f}_k^*$ belonging to the class $\widetilde{\mathcal{F}}$ defined in (3.5) via SGD. With the obtained $\nabla\widetilde{f}_k^*$, VTIG push μ-particles in this direction as follows (Line 11 of Algorithm 1),

$$x_{i,k+1}^\mu \leftarrow \mathrm{Exp}_{x_{i,k}^\mu}\big[-\eta^\mu \cdot \nabla\widetilde{f}_k^*(x_{i,k}^\mu)\big] \tag{3.7}$$

for all $i \in [N_\mu]$. Here $\eta^\mu > 0$ are the stepsize specified in Theorem 4.3. This is equivalent to updating the empirical measure $\widehat{\mu} = N_\mu^{-1}\sum_{i\in[N_\mu]}\delta_{x_{i,k}}$ by the pushforward measure $[\mathrm{Exp}_{\mathcal{X}}(-\eta^\mu\cdot\nabla\widetilde{f}_{l,k}^*)]_\sharp\widehat{\mu}$, which approximates the Riemannian gradient update in (3.1) with stepsize η^μ. Also, the exponential map in Euclidean space is reduced to a gradient descent step on $x_{i,k}^\mu \in \mathbb{R}^d$.

Similarly, to update the set of ν-particles $X_l^\nu(\widetilde{\mu}_k)$, VTIG computes the solution to (3.4) following Line 6 in Algorithm 1 at inner-loop timestep l of outer-loop timestep k. Then, the ν-particles are pushed by

$$x_{i,l+1}^\nu(\widetilde{\mu}_k) \leftarrow \mathrm{Exp}_{x_{i,l}^\nu(\widetilde{\mu}_k)}\Big[\eta^\nu \cdot \nabla\widetilde{f}_{l,k}^*\big(x_{i,l}^\nu(\widetilde{\mu}_k)\big)\Big] \tag{3.8}$$

for all $i \in [N_\nu]$ in Line 7 of Algorithm 1, with fixed $\widetilde{\mu}_k$. In particular, the sequences of probability measures $\{\widetilde{\mu}_k\}_{k\ge0}$ and $\{\widetilde{\nu}_l(\widetilde{\mu}_k)\}_{k,l\ge0}$ are constructed as below. We define sequences of transportation maps $\{T_k^\mu\colon \mathcal{X}\to\mathcal{X}\}_{k=0}^{K_\mu}$ with $T_0^\mu = \mathrm{id}$ and $\{T_m^\nu\colon \mathcal{X}\to\mathcal{X}\}_{m=0}^{K_\mu K_\nu}$ with $T_0^\nu = \mathrm{id}$, by

$$T_{k+1}^\mu = \mathrm{Exp}_{\mathcal{X}}(-\eta^\mu\cdot\nabla\widetilde{f}_k^*)\circ T_k^\mu \quad \text{and} \quad T_{kl+1}^\nu = \mathrm{Exp}_{\mathcal{X}}(-\eta^\nu\cdot\nabla\widetilde{f}_{l,k}^*)\circ T_{kl}^\nu, \tag{3.9}$$

respectively for $k \in [K_\mu], l \in [K_\nu]$. Here K_μ and K_ν are the numbers of timesteps of the inner and outer loops, respectively. Then for each $k \ge 1$ we define $\widetilde{\mu}_k = (T_k^\mu)_\sharp\widetilde{\mu}_0$ and $\widetilde{\nu}_k = (T_k^\nu)_\sharp\widetilde{\nu}_0$, where μ_0 and ν_0 are initial probability measures. Hence, we have $\widetilde{\nu}_l(\widetilde{\mu}_k) = \widetilde{\nu}_{lk}$. Also, $x_{i,k}^\mu \overset{\text{i.i.d.}}{\sim} \widetilde{\mu}_k$ and $x_{i,l}^\nu(\widetilde{\mu}_k) \overset{\text{i.i.d.}}{\sim} \widetilde{\nu}_l(\widetilde{\mu}_k)$ are independent samples. Such implicit construction of transportation maps and probability measures also induces a theoretical version of VTIG via resampling. See Algorithm 2 for details. Additionally, we adopt the constructed measure $\widetilde{\nu}_{k+1}$ to compute $F^*_{\widetilde{\nu}_{k+1}}$ in (3.6) since for most objectives F such as that in Example 2.4, we can always sample many enough particles to approximate the expectation terms *w.r.t.* $\widetilde{\nu}_{k+1}$ for $k \ge 0$.

Algorithm 1 Multi-Step Variational Transport Algorithm for Infinite-Dimensional Games (VTIG)

Input: Functional $F\colon \mathcal{M} \times \mathcal{M} \to \mathbb{R}$; initial probability measures $\widetilde{\mu}_0, \widetilde{\nu}_0 \in \mathcal{M}$; numbers of particles N_μ, N_ν; numbers of iterations K_μ, K_ν; and stepsizes $\eta^\mu \in \big(0, \min\{1/h, 2/\widetilde{L}\}\big)$, $\eta^\nu \in \big(0, \min\{1/(4L_2), 1/h\}\big)$.
Initialize N_μ (N_ν) particles $X_0^\mu = \{x_{i,0}^\mu\}_{i\in[N_\mu]}$ $\big(X_0^\nu\big)$ by drawing N_μ (N_ν) i.i.d. samples from $\widetilde{\mu}_0$ ($\widetilde{\nu}_0$).
for $k = 0, 1, 2, \ldots, K_\mu - 1$ **do**
  Set $X_0^\nu(\widetilde{\mu}_k) = X_k^\nu$
  for $l = 0, 1, 2, \ldots, K_\nu - 1$ **do**
    $\widetilde{f}_{l,k}^* \leftarrow \mathbf{VFM}\big(X_l^\nu(\widetilde{\mu}_k), F_{\widetilde{\mu}_k}^*, N_\nu\big)$
    Push ν-particles: $x_{i,l+1}^\nu(\widetilde{\mu}_k) \leftarrow \mathrm{Exp}_{x_{i,l}^\nu(\widetilde{\mu}_k)}\big[-\eta^\nu \cdot \nabla \widetilde{f}_{l,k}^*\big(x_{i,l}^\nu(\widetilde{\mu}_k)\big)\big]$ for all $i \in [N_\nu]$
  end for
  Set $X_{k+1}^\nu = X_{K_\nu}^\nu(\widetilde{\mu}_k)$
  $\widetilde{f}_k^* \leftarrow \mathbf{VFM}\big(X_k^\mu, F_{\widetilde{\nu}_{k+1}}^*, N_\mu\big)$
  Push μ-particles: $x_{i,k+1}^\mu \leftarrow \mathrm{Exp}_{x_{i,k}^\mu}[-\eta^\mu \cdot \nabla \widetilde{f}_k^*(x_{i,k}^\mu)]$ for all $i \in [N_\mu]$
  Set $X_{k+1}^\mu = \{x_{i,k+1}^\mu\}_{i\in[N_\mu]}$
end for
Output: $\widetilde{\mu}^* = N_\mu^{-1} \sum_{i\in[N_\mu]} \delta_{x_{i,K_\mu}}$, $\widetilde{\nu}^* = N_\nu^{-1} \sum_{i\in[N_\nu]} \delta_{x_{i,K_\nu}}$

4 Main Results

To ensure the independence of the particles for statistical analysis, we adopt Algorithm 2 for theoretical analysis. We characterize the statistical error induced by estimating Riemannian gradients using finite particle samples in §4.1 for both players. In §4.2 we establish the convergence rate of VTIG to the IFNE under the PL condition for one player. Furthermore, we present in §4.3 that under a stronger assumption on the objective F, *i.e.,* the two-sided PL condition, a linear convergence rate to the minimax value of the game is achieved.

4.1 Statistical Analysis

For each player, VTIG can be viewed as a Riemannian gradient descent method with biased gradient estimates. We characterize the bias in terms of the generalization error of function approximators, where lie the essential difficulties in theory. In this section, we present the analysis for player μ. The analysis of player ν is similar.

Gradient estimation. Recall that by Proposition 3.1, the desired descent direction for timestep $k \geq 0$ is $\mathrm{grad}\, F(\widetilde{\mu}_k) = -\operatorname{div}(\widetilde{\mu}_k \cdot \nabla f_k^*)$. However, with only finite samples, we obtain an estimator $\widetilde{f}_k^*$ of f_k^*. Hence, the gradient estimate at $\widetilde{\mu}_k$ is $-\operatorname{div}(\widetilde{\mu}_k \cdot \nabla \widetilde{f}_k^*)$, and the difference between $\mathrm{grad}\, F(\widetilde{\mu}_k)$ and its estimate is denoted by $\delta_k = -\operatorname{div}[\widetilde{\mu}_k \cdot (\nabla \widetilde{f}_k^* - \nabla f_k^*)]$. By observing that $\delta_k \in T_{\widetilde{\mu}_k}\mathcal{M}$ and that the randomness of δ_k comes from the intial samples X_0^μ, we define

$$\bar{\varepsilon}_k = \mathbb{E}_{X_0^\mu}\langle \delta_k, \delta_k \rangle_{\widetilde{\mu}_k} = \mathbb{E}_{X_0^\mu} \int_{\mathcal{X}} \big\|\nabla \widetilde{f}_k^*(x) - \nabla f_k^*(x)\big\|_2^2 \,\mathrm{d}\widetilde{\mu}_k(x) \tag{4.1}$$

as the (expected) gradient error. In general, it is hard to derive upper bounds of gradients for general functions. Nevertheless, we upper bounds function gradients by function values for a specific function class, $\widetilde{\mathcal{F}}$ defined in Section 3.1. Below we provide a generic assumption on $\widetilde{\mathcal{F}}$ to derive the desired upper bounds of gradients.

Assumption 4.1. The set $\nabla\widetilde{\mathcal{F}} = \{\nabla f : f \in \widetilde{\mathcal{F}}\}$ is closed, bounded in $(\mathcal{C}(\mathcal{X}), \ell_\infty)$. For each $\nabla f \in \nabla\widetilde{\mathcal{F}}$, ∇f is h-Lipschitz for some $h > 0$.

Such an assumption can be achieved by function classes with uniformly bounded and Lipschitz continuous gradients, which includes the class of neural networks defined in (3.5) with bounded parameters. See §D.1 for more details. Then, we state the following inequality for gradient error bounds.

Lemma 4.2 (Reverse Poincaré inequality)**.** Under Assumption 4.1, for every $f \in \widetilde{\mathcal{F}}$ and $p \geq 1$, there exists a constant $\widetilde{K}$ such that

$$\int_{\mathcal{X}} \|\nabla f\|^p d\mu \leq \widetilde{K} \int_{\mathcal{X}} |f|^p d\mu, \tag{4.2}$$

where $\mathcal{X}$ is a compact set, and μ is a nonnegative measure on $\mathcal{X}$.

Proof. See §D.1 for a detailed proof. □

We identify Lemma 4.2 as a new type of reverse Poincaré inequality [3] based on the fundamental topology and analysis property of $\mathcal{X}$ and $\widetilde{\mathcal{F}}$, which can be extended to non-Euclidean space $\mathcal{X}$.

Generalization error of VFM. By setting $p = 2$ and $f(x) = \widetilde{f}_k^*(x) - f_k^*(x)$ in Lemma 4.2, we are able to upper bound the gradient errors by the generalization errors of NNs, which is bounded in §D.3 with the orders of

$$\bar{\varepsilon}_\mu = \mathcal{O}(N_\mu^{-1/2}), \quad \bar{\varepsilon}_\nu = \mathcal{O}(N_\nu^{-1/2}) \tag{4.3}$$

by wide enough NNs for player μ and player ν, respectively. Here N_μ and N_ν are the numbers of particles for player μ and player ν, respectively. Such results are standard for the stochastic gradient descent (SGD) over neural networks, since the number of iterations t in Algorithm 3 is also the sample size N_μ (resp. N_ν) in our algorithm.

4.2 Convergence to the IFNE for PL Games

Recall that $G(\mu) = \max_{\nu \in \mathcal{M}} F(\mu, \nu)$. We define $L_G = \max_{\mu \in \mathcal{M}} \|\operatorname{grad} G(\mu)\|_\mu$, which is upper bounded since G is Lipschitz (Lemma E.1) on a compact domain $\mathcal{M}$ (Proposition E.1). We assume that there exist constants $M_H > 0$ and $M_G > 0$ such that $M_H = \max_{\mu,\nu_0 \in \mathcal{M}} [G(\mu) - F(\mu, \nu_0)]$ and $M_G = \max_{\mu_0 \in \mathcal{M}} G(\mu_0) - G(\mu^*)$, where $\mu^* \in \operatorname{argmin}_{\mu \in \mathcal{M}} G(\mu)$. Under Assumption 2.3 for ξ-PL games, we characterize the following sublinear rate to find an IFNE defined in Definition 2.5 by VTIG with sample sizes N_μ, N_ν and numbers of iterations K_μ, K_ν. Recall that L_0, L_1, and L_2 are Lipschitz constants defined in Assumption 2.2. The constant ξ for the PL condition is defined in Assumption 2.3. Also, $\sigma = 1 - \xi\eta^\nu/2 \in (0, 1)$ is a contraction factor from Lemma E.5.

Theorem 4.3 (Convergence of Infinite-Dimensional PL Games)**.** Suppose that the objective F admits a variational form under Assumption 2.2 and 2.3. Also, the function class $\widetilde{\mathcal{F}}$ satisfies Assumption 4.1. We set the stepsizes to be $\eta^\mu \in \big[0, \min\{1/h, 2/\widetilde{L}\}\big)$ and $\eta^\nu \in \big(0, \min\{1/(4L_\nu), 1/h\}\big)$, where $\widetilde{L} = L_1 + L_0^2/\xi$. Then, for any $\theta > 0$, if

$$K_\nu \geq K_\nu(\theta) = \mathcal{O}\Big(\log \frac{(1-\sigma)\widehat{M}_H - \eta^\nu \bar{\varepsilon}_\nu}{\theta} \Big/ \log \frac{1}{\sigma} \Big), \quad \text{where } \widehat{M}_H = \max\Big\{ M_H, \frac{\eta^\nu \bar{\varepsilon}_\nu + 1}{1-\sigma} \Big\}, \tag{4.4}$$

there exists an iteration $k \in [K_\mu]$ such that

$$\mathbb{E}_{X_0}\big[\mathcal{J}_\mu^2(\widetilde{\mu}_k, \widetilde{\nu}_{k+1})\big] = \mathcal{O}\bigg(\Big(\Delta + \sqrt{\bar{\varepsilon}_\mu}\Big)^2 \cdot \Big((\Delta + \sqrt{\bar{\varepsilon}_\mu}) + \frac{M_G}{K_\mu} \Big) \bigg), \quad \mathbb{E}_{X_0}\big[\mathcal{J}_\nu(\widetilde{\mu}_k, \widetilde{\nu}_{k+1})\big] = \mathcal{O}\Big(\frac{L_2 \Delta}{L_0} \Big). \tag{4.5}$$

Here $\Delta = L_0 \sqrt{\dfrac{\eta_\nu \bar{\varepsilon}_\nu + \theta}{2\xi(1-\sigma)}}$, and the gradient error terms $\bar{\varepsilon}_\mu$ and $\bar{\varepsilon}_\nu$ are characterized in (4.3).

Proof. See §E.3 for a detailed proof and more dependencies on other constants. □

The proof of Theorem 4.3 is based on Lemma E.5 and a Danskin-type lemma in §E.1 which ensures an appropriate estimate of $\operatorname{grad} G$ provided by inner loops and the smoothness of the objective defined in Assumption 2.2. Such properties imply that VTIG behaves as the gradient descent over the inner maximization value functional G, which concludes the proof. The bounds for the first-order errors $\mathcal{J}_\mu$ and $\mathcal{J}_\nu$ are composed of the optimization error θ of player ν, the optimization error $\mathcal{O}(K_\mu^{-1})$ of player μ, and the gradient errors $\bar{\varepsilon}_\mu$ and $\bar{\varepsilon}_\nu$ characterized in (4.3) due to finite samples. Specifically, the term Δ encapsulates both the statistical

error and the optimization error of player ν, which are added to $\sqrt{\bar{\varepsilon}_\mu}$ and $\mathcal{O}(K_\mu^{-1})$ in the error bound for player μ. Considering N_μ, N_ν, K_μ, and K_ν as dominating terms in the bounds, if we set $N_\mu = N_\nu = \mathcal{O}(\theta^{-4})$, $K_\mu \geq K_\mu(\theta) = \mathcal{O}(\theta^{-2})$, and $K_\nu \geq K_\nu(\theta) = \mathcal{O}(\log(\theta^{-1}))$, by Definition 2.5 we achieve a θ-IFNE. In this sense, VTIG converges at a sublinear rate to the IFNE defined in (2.10) under the PL game condition.

4.3 Convergence to the Minimax Value under the Two-Sided PL Condition

In this section, we aim to achieve a stronger convergence result leading to the minimax value of the game by a stronger assumption. We give the definition of two-sided Riemannian PL games below.

Assumption 4.4 (Two-Sided Riemannian PL Game)**.** We define functionals $H_\mu(\nu) = -F(\mu, \nu)$ and $F_\nu(\mu) = F(\mu, \nu)$ for fixed μ and ν, respectively. We assume (2.3) to be a two-sided Riemannian PL game, or simply a two-sided PL game, in the sense that $F_\nu(\mu)$ is ξ_1-PL and $H_\mu(\nu)$ is ξ_2-PL for some $\xi_1, \xi_2 > 0$.

Note that the definition of two-sided PL games relaxes that of the convex-concave games even in infinite-dimensional spaces. In fact, Example 2.4 provides a two-sided PL game by KL regularization for both players, which is ubiquitous in training GANs. Assumption 4.4 also guarantees a PL condition on $G(\mu) = \max_{\nu \in \mathcal{M}} F(\mu, \nu)$ according to Lemma F.1. By using such a landscape, we establish the linear convergence rate of finding the minimax value of the game as below.

Theorem 4.5 (Convergence to the Minimax Value of Two-Sided PL Games)**.** Let the objective F satisfy (2.4), Assumption 2.2 and 4.4. Suppose that $\widetilde{\mathcal{F}}$ satisfies Assumption 4.1. With the outer-loop stepsize $\eta^\mu \in \big[0, \min\{1/h, 1/(4\widetilde{L})\}\big)$ and inner-loop stepsize $\eta^\nu \in \big(0, \min\{1/(4L_\nu), 1/h\}\big)$, for $k \geq 1$ it holds that

$$\mathbb{E}_{X_0}\Big[F\big(\widetilde{\mu}_k, \nu^*(\widetilde{\mu}_k)\big)\Big] - F(\mu^*, \nu^*) \leq \underbrace{\widetilde{\sigma}^k \cdot \Big(\mathbb{E}_{X_0}\big[F(\widetilde{\mu}_0, \widetilde{\nu}_1)\big] - F(\mu^*, \nu^*)\Big)}_{\text{(i)}} + \underbrace{\frac{1 - \widetilde{\sigma}^k}{1 - \widetilde{\sigma}} \cdot \eta^\mu \big(\bar{\varepsilon}_\mu + \widetilde{\Delta}^2\big)}_{\text{(ii)}}, \tag{4.6}$$

where $\widetilde{\mu}_k$ and $\widetilde{\nu}_{k+1}$ are probability measure iterates defined in Algorithm 2, gradient error terms $\bar{\varepsilon}_\mu$ and $\bar{\varepsilon}_\nu$ are given in (4.3). The expectation is taken *w.r.t.* the initial sample X_0. The contraction factor is $\widetilde{\sigma} = 1 - \xi_1 \eta^\mu / 2$, and we define the total error term for player ν as $\widetilde{\Delta}^2 = L_0^2 / 2\xi_2 \cdot \Big(\sigma^{K_\nu} \cdot M_H + \eta^\nu \bar{\varepsilon}_\nu \cdot \frac{1 - \sigma^{K_\nu}}{1 - \sigma}\Big)$, where M_H is the upper bound of $F(\mu, \nu^*(\mu)) - F(\mu, \nu_0(\mu))$ defined in §4.2, and K_ν denotes the number of timesteps for player ν in Algorithm 2.

Proof. See §F.2 for a detailed proof. □

The proof of Theorem 4.5 differs from that of Theorem 4.3 mainly by the lower bounds of gradient norms provided by ξ_1-PL condition on functional G. Under the two-sided PL condition in Assumption 4.4, Theorem 4.5 characterizes a linear convergence rate for VTIG of the objective functional value to the minimax value $F(\mu^*, \nu^*)$ of the game, with an accumulated statistical error term (ii). In detail, the optimization error (i) decays by a factor of $\widetilde{\sigma}$ linearly. Additionally, our statistical error is composed of the gradient error $\bar{\epsilon}_\mu$ of player μ and the error term $\widetilde{\Delta}^2$, which is further decomposed into the linearly decaying optimization error $\sigma^{K_\nu} M_H$ and the gradient error $\bar{\epsilon}_\nu$ of player ν scaled by $(1 - \sigma^{K_\nu})/(1 - \sigma)$. Specifically, in the total bound (4.6) $\bar{\epsilon}_\mu$ scales at a rate of $(1 - \widetilde{\sigma}^k)/(1 - \widetilde{\sigma})$ with the iteration k and $\bar{\epsilon}_\nu$ scales at the rate of $(1 - \widetilde{\sigma}^k)/(1 - \widetilde{\sigma}) \cdot (1 - \sigma^{K_\nu})/(1 - \sigma)$, which implies the error accumulation from the the inner loop of Algorithm 2. Also, we adopt the objective value instead of IFNE in Theorem 4.3 to measure the error of convergence to the minimax value. Although we suffer from the finite-sample error to approximate probability measures, it is flexible to tune parameters N_μ, N_ν, K_μ, and K_ν according to their corresponding error terms in the bound to optimize the algorithm in practice, especially when some parameters are restricted.

References

[1] ARJOVSKY, M., CHINTALA, S. and BOTTOU, L. (2017). Wasserstein gan. *arXiv preprint arXiv:1701.07875* .

[2] ARORA, S., GE, R., LIANG, Y., MA, T. and ZHANG, Y. (2017). Generalization and equilibrium in generative adversarial nets (gans). In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org.

[3] BAUDOIN, F. and BONNEFONT, M. (2016). Reverse poincaré inequalities, isoperimetry, and riesz transforms in carnot groups. *Nonlinear Analysis* **131** 48–59.

[4] BERNHARD, P. and RAPAPORT, A. (1995). On a theorem of danskin with an application to a theorem of von neumann-sion. *Nonlinear Analysis: Theory, Methods & Applications* **24** 1163–1181.

[5] BURAGO, D., IVANOV, S. and BURAGO, Y. (2001). Course in metric geometry .

[6] CARLEN, E. A. and GANGBO, W. (2003). Constrained steepest descent in the 2-Wasserstein metric. *Annals of mathematics* 807–846.

[7] CHAMBOLLE, A. and POCK, T. (2016). On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming* **159** 253–287.

[8] CHEN, C., ZHANG, R., WANG, W., LI, B. and CHEN, L. (2018). A unified particle-optimization framework for scalable bayesian sampling. *arXiv preprint arXiv:1805.11659* .

[9] CHERN, S.-S., CHEN, W.-H. and LAM, K. S. (1999). *Lectures on differential geometry*, vol. 1. World Scientific Publishing Company.

[10] COTTER, N. E. (1990). The stone-weierstrass theorem and its application to neural networks. *IEEE Transactions on Neural Networks* **1** 290–295.

[11] CSÁJI, B. C. ET AL. (2001). Approximation with artificial neural networks. *Faculty of Sciences, Etvs Lornd University, Hungary* **24** 7.

[12] DAI, B., DAI, H., GRETTON, A., SONG, L., SCHUURMANS, D. and HE, N. (2018). Kernel exponential family estimation via doubly dual embedding. *arXiv preprint arXiv:1811.02228* .

[13] DAI, B., SHAW, A., LI, L., XIAO, L., HE, N., LIU, Z., CHEN, J. and SONG, L. (2017). Sbeed: Convergent reinforcement learning with nonlinear function approximation. *arXiv preprint arXiv:1712.10285* .

[14] DANG, C. D. and LAN, G. (2015). On the convergence properties of non-euclidean extragradient methods for variational inequalities with generalized monotone operators. *Computational Optimization and applications* **60** 277–310.

[15] DASKALAKIS, C. and PANAGEAS, I. (2018). Last-iterate convergence: Zero-sum games and constrained min-max optimization. *arXiv preprint arXiv:1807.04252* .

[16] DASKALAKIS, C. and PANAGEAS, I. (2018). The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in Neural Information Processing Systems*.

[17] EVANS, L. (2010). *Partial Differential Equations*. American Mathematical Society.

[18] FACCHINEI, F. and PANG, J.-S. (2007). *Finite-dimensional variational inequalities and complementarity problems*. Springer Science & Business Media.

[19] FROGNER, C. and POGGIO, T. (2018). Approximate inference with wasserstein gradient flows. *arXiv preprint arXiv:1806.04542* .

[20] GAIVORONSKI, A. (1986). Linearization methods for optimization of functionals which depend on probability measures. In *Stochastic Programming 84 Part II*. Springer, 157–181.

[21] GHOSH, S., SQUILLANTE, M. and WOLLEGA, E. (2018). Efficient stochastic gradient descent for distributionally robust learning. *arXiv preprint arXiv:1805.08728* .

[22] GIDEL, G., BERARD, H., VIGNOUD, G., VINCENT, P. and LACOSTE-JULIEN, S. (2018). A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551* .

[23] GIDEL, G., HEMMAT, R. A., PEZESHKI, M., LEPRIOL, R., HUANG, G., LACOSTE-JULIEN, S. and MITLIAGKAS, I. (2018). Negative momentum for improved game dynamics. *arXiv preprint arXiv:1807.04740* .

[24] GIGLI, N. (2011). On the inverse implication of brenier-mccann theorems and the structure of $(p_2(\mathcal{M}), w_2)$. *Methods and Applications of Analysis* **18** 127–158.

[25] GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A. and BENGIO, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*.

[26] GRNAROVA, P., LEVY, K. Y., LUCCHI, A., HOFMANN, T. and KRAUSE, A. (2017). An online learning approach to generative adversarial networks. *arXiv preprint arXiv:1706.03269* .

[27] HAMEDANI, E. Y., JALILZADEH, A., AYBAT, N. and SHANBHAG, U. (2018). Iteration complexity of randomized primal-dual methods for convex-concave saddle point problems. *arXiv preprint arXiv:1806.04118* .

[28] HO, J. and ERMON, S. (2016). Generative adversarial imitation learning. In *Advances in neural information processing systems*.

[29] HOFMANN, T., SCHÖLKOPF, B. and SMOLA, A. J. (2008). Kernel methods in machine learning. *Annals of Statistics* 1171–1220.

[30] HSIEH, Y.-P., LIU, C. and CEVHER, V. (2018). Finding mixed nash equilibria of generative adversarial networks. *arXiv preprint arXiv:1811.02002* .

[31] JIN, C., NETRAPALLI, P. and JORDAN, M. I. (2019). Minmax optimization: Stable limit points of gradient descent ascent are locally optimal. *arXiv preprint arXiv:1902.00618* .

[32] JUDITSKY, A. and NEMIROVSKI, A. (2016). Solving variational inequalities with monotone operators on domains given by linear minimization oracles. *Mathematical Programming* **156** 221–256.

[33] KARIMI, H., NUTINI, J. and SCHMIDT, M. (2016). Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conf erence on Machine Learning and Knowledge Discovery in Databases*. Springer.

[34] LEE, J. M. (2001). *Introduction to smooth manifolds*. Springer.

[35] LIU, C., ZHUO, J., CHENG, P., ZHANG, R., ZHU, J. and CARIN, L. (2018). Understanding and accelerating particle-based variational inference. *arXiv preprint arXiv:1807.01750* .

[36] LIU, Y., SHANG, F., CHENG, J., CHENG, H. and JIAO, L. (2017). Accelerated first-order methods for geodesically convex optimization on riemannian manifolds. In *Advances in Neural Information Processing Systems*.

[37] LU, S., TSAKNAKIS, I. and HONG, M. (2019). Block alternating optimization for non-convex min-max problems: algorithms and applications in signal processing and communications. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

[38] LU, S., TSAKNAKIS, I., HONG, M. and CHEN, Y. (2019). Hybrid block successive approximation for one-sided non-convex min-max problems: algorithms and applications. *arXiv preprint arXiv:1902.08294* .

[39] MAI, T., MIHAIL, M., PANAGEAS, I., RATCLIFF, W., VAZIRANI, V. and YUNKER, P. (2018). Cycles in zero-sum differential games and biological diversity. In *Proceedings of the 2018 ACM Conference on Economics and Computation*.

[40] MESCHEDER, L., GEIGER, A. and NOWOZIN, S. (2018). Which training methods for gans do actually converge? *arXiv preprint arXiv:1801.04406* .

[41] MOLCHANOV, I. and ZUYEV, S. (2001). Variational calculus in the space of measures and optimal design. In *Optimum design 2000*. Springer, 79–90.

[42] MONTEIRO, R. D. and SVAITER, B. F. (2010). On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM Journal on Optimization* **20** 2755–2787.

[43] NEMIROVSKI, A. (2004). Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization* **15** 229–251.

[44] NESTEROV, Y. and POLYAK, B. T. (2006). Cubic regularization of newton method and its global performance. *Mathematical Programming* **108** 177–205.

[45] NGUYEN, X., WAINWRIGHT, M. J. and JORDAN, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory* **56** 5847–5861.

[46] NOWOZIN, S., CSEKE, B. and TOMIOKA, R. (2016). f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*.

[47] OLIEHOEK, F. A., SAVANI, R., GALLEGO, J., VAN DER POL, E. and GROSS, R. (2018). Beyond local nash equilibria for adversarial networks. In *Benelux Conference on Artificial Intelligence*. Springer.

[48] OTTO, F. and VILLANI, C. (2000). Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis* **173** 361–400.

[49] PANG, J.-S. and RAZAVIYAYN, M. (2016). A unified distributed algorithm for non-cooperative games.

[50] PETERSEN, P., AXLER, S. and RIBET, K. (2006). *Riemannian geometry*, vol. 171. Springer.

[51] POLYAK, B. T. (1963). Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki* **3** 643–653.

[52] RAGINSKY, M., RAKHLIN, A. and TELGARSKY, M. (2017). Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. *arXiv preprint arXiv:1702.03849* .

[53] RICHEMOND, P. H. and MAGINNIS, B. (2017). On wasserstein reinforcement learning and the fokker-planck equation. *arXiv preprint arXiv:1712.07185* .

[54] ROCKAFELLAR, R. T. (1970). *Convex analysis*. 28, Princeton university press.

[55] RUDIN, W. (1976). Principles of mathematical analysis. .

[56] SANJABI, M., BA, J., RAZAVIYAYN, M. and LEE, J. D. (2018). On the convergence and robustness of training gans with regularized optimal transport. In *Advances in Neural Information Processing Systems*.

[57] SANJABI, M., RAZAVIYAYN, M. and LEE, J. D. (2018). Solving non-convex non-concave min-max games under polyak-Łojasiewicz condition. *arXiv preprint arXiv:1812.02878* .

[58] SINHA, A., NAMKOONG, H. and DUCHI, J. (2017). Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571* .

[59] STEWART, J. (2008). Chapter 15.2 limits and continuity, multivariable calculus.

[60] THEKUMPARAMPIL, K. K., JAIN, P., NETRAPALLI, P. and OH, S. (2019). Efficient algorithms for smooth minimax optimization. In *Advances in Neural Information Processing Systems*.

[61] VAN PARYS, B. P., ESFAHANI, P. M. and KUHN, D. (2017). From data to decisions: Distributionally robust optimization is optimal. *arXiv preprint arXiv:1704.04118* .

[62] VILLANI, C. (2003). *Topics in optimal transportation*. 58, American Mathematical Society.

[63] VILLANI, C. (2008). *Optimal transport: old and new*, vol. 338. Springer.

[64] WIBISONO, A. (2018). Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem. *arXiv preprint arXiv:1802.08089* .

[65] ZHANG, H., REDDI, S. J. and SRA, S. (2016). Riemannian svrg: Fast stochastic optimization on riemannian manifolds. In *Advances in Neural Information Processing Systems*.

[66] ZHANG, H. and SRA, S. (2018). An estimate sequence for geodesically convex optimization. In *Conference On Learning Theory*.

[67] ZOU, D., XU, P. and GU, Q. (2018). Subsampled stochastic variance-reduced gradient langevin dynamics. In *International Conference on Uncertainty in Artificial Intelligence*.

A Additional Algorithms

Algorithm 2 Theoretical Version of Multi-Step Variational Transport Algorithm for Infinite-Dimensional Games (VTIG) with Resampling

Input: Functional $F\colon \mathcal{M} \times \mathcal{M} \to \mathbb{R}$; initial measures $\widetilde{\mu}_0 \in \mathcal{M}$, $\widetilde{\nu}_0 \in \mathcal{M}$; numbers of particles N_μ, N_ν; numbers of iterations K_μ, K_ν; and stepsizes $\eta_k^\mu \in \big(0, \min\{1/h, 2/\widetilde{L}\}\big)$, $\eta^\nu \in \big(0, \min\{1/(4L_2), 1/h\}\big)$.
Initialize the transportation maps $T_0^\mu \leftarrow \mathrm{id}$, $T_0^\nu \leftarrow \mathrm{id}$.
Initialize N_μ particles $X_0^\mu = \{x_i^\mu\}_{i\in[N_\mu]}$ by drawing N_μ i.i.d. samples from $\widetilde{\mu}_0$.
Initialize N_ν particles $X_0^\nu = \{x_i^\nu\}_{i\in[N_\nu]}$ by drawing N_ν i.i.d. samples from $\widetilde{\nu}_0$.
for $k = 0, 1, 2, \ldots, K_\mu - 1$ **do**
  Generate N_μ particles $X_0^\mu = \{x_i^\mu\}_{i\in[N_\mu]}$ by drawing N_μ i.i.d. samples from $\widetilde{\mu}_0$.
  Push μ-particles by letting $X_k^\mu \leftarrow T_k^\mu(X_0^\mu)$.
  for $l = 0, 1, 2, \ldots, K_\nu - 1$ **do**
    Generate N_ν particles $X_0^\nu(\widetilde{\mu}_k) = \{x_i^\nu\}_{i\in[N]}$ by drawing N_ν i.i.d. samples from $\widetilde{\nu}_0$.
    Push ν-particles by letting $X_l^\nu(\widetilde{\mu}_k) \leftarrow T_{kK_\nu+l}^\nu(X_0^\nu(\widetilde{\mu}_k))$.
    $\widetilde{f}_{l,k}^* \leftarrow \mathbf{VFM}\big(X_l^\nu(\widetilde{\mu}_k), F_{\widetilde{\mu}_k}^*, N_\nu\big)$.
    Update the transportation map by letting $T_{kK_\nu+l+1}^\nu = [\mathrm{Exp}_{\mathcal{X}}(-\eta^\nu \cdot \nabla \widetilde{f}_{l,k}^*)] \circ T_{kK_\nu+l}^\nu$.
  end for
  $\widetilde{f}_k^* \leftarrow \mathbf{VFM}\big(X_k^\mu, F_{\widetilde{\nu}_{k+1}}^*, N_\mu\big)$.
  Update the transportation map by letting $T_{k+1}^\mu = [\mathrm{Exp}_{\mathcal{X}}(-\eta^\mu \cdot \nabla \widetilde{f}_k^*)] \circ T_k^\mu$.
end for
Output: The final transportation maps $T_{K_\mu}^\mu$ and $T_{K_\mu K_\nu}^\nu$.

Algorithm 3 Variational Form Maximization via SGD $\big(\mathbf{VFM}\big(\{x_i\}_{i=1}^t, F^*, t\big)\big)$

Require: Initial weights b_i, $[\beta(0)]_i$ $(i \in [w])$, number of iterations t, sample $\{x_i\}_{i=1}^t$, and projection radius r_f.
Set stepsize $\eta \leftarrow t^{-1/2}$
for $s = 0, \ldots, t-1$ **do**
  $x \leftarrow x_{i+1}$
  $\beta(s+1/2) \leftarrow \beta(s) - \eta\left(\nabla_\beta F^*(f_{\beta(s)}(x)) - \nabla_\beta f_{\beta(s)}(x)\right)$
  $\beta(s+1) \leftarrow \mathrm{argmin}_{\beta\in\mathcal{B}^0(r_f)}\big\{\|\beta - \beta(s+1/2)\|_2\big\}$
end for
Average over path $\widehat{\beta} \leftarrow 1/t \cdot \sum_{s=0}^{t-1} \beta(s)$
Output: $f_{\widehat{\beta}}$

Algorithm 2 provide the resampling version of VTIG (Algorithm 1) for theoretical analysis in §4. As mentioned above, here we explicitly use two sequences of transportation maps $\{T_k^\mu\}_{k=0}^{K_\mu}$ and $\{T_k^\nu\}_{k=0}^{K_\mu K_\nu}$ to record the transportation plans of mapping the initial probability measures to the current iterates of Player μ and player ν, respectively. For simplicity, we put the particle pushing step before the VFM step, while the essential difference between Algorithm 2 and Algorithm 1 lies only in the resampling procedures at the beginning of each player's loop, *i.e.,* Line 5 and Line 8 in Algorithm 2, after which we adopt the recorded transportation plan from time-step 0 to the current timestep k, T_k^μ or T_k^ν, to push the resampled initial particles into their states at the current timestep through k steps of transportation maps. In contrast, in Algorithm 1 we push the two sets of particles for one step at each timestep based on their current states, with

only one sampling procedure before the beginning of the two loops. Hence, We remark that Algorithm 1 is deterministic in essence; the only randomness comes from the initialization of the particles. In addition, note that even in resampling cases, our algorithm only requires to call an oracle that is able to sample from a fixed distribution, which is more efficient than SGLD requiring Markov-chain type sampling for different probability measure iterates. We remark that Algorithm 2 is considered only for the sake of theoretical analysis; random sampling is unnecessary in practical implementation.

We also remark that the exponential maps $\mathrm{Exp}_{\mathcal{X}}(-\eta^{\mu} \cdot \nabla \widetilde{f}_k^*)$ in Algorithm 1 and 2 are essentially translations of particles in the direction of $-\nabla \widetilde{f}_k^*$ with a stepsize η^{ν} in $\mathcal{X}$, which implies that they are computationally efficient for infinite-dimensional problems.

B Properties of Optimization over Riemannian Manifolds

B.1 Background Details

In this section, we present basic properties of functions and gradient descent on a Riemannian manifold $(\mathcal{M}, g)$ with a Riemannain metric g, supposing that any two points on $\mathcal{M}$ uniquely determines a geodesic. By the nature of Wasserstein spaces, these propositions also apply to our optimization process over the family of distributions. With such background, we are better prepared to demonstrate the convergence results by extending the classical optimization arguments to Riemannian manifolds. For further extensions of definitions and propositions within this section, we refer to other literatures on geodesically convex optimization. See, e.g., [36, 65, 66] and the references therein.

B.2 Riemannian Manifold

Let $\mathcal{M}$ be a d-dimensional differential manifold, *i.e.,* a topological space that is locally homeomorphic to the Euclidean space $\mathbb{R}^d$ and has a globally defined differential structure [9].

Definition B.1 (Tangent vector)**.** A tangent vector at $x \in \mathcal{M}$ is an equivalence class of differentiable curves through x with a prescribed velocity vector at x. The tangent space at x, denoted by $T_x\mathcal{M}$, consists of all tangent vectors at x.

In what follows, we denote by $f : \mathcal{M} \to \mathbb{R}$ a differentiable function over $\mathcal{M}$, and define its directional derivative as $\nabla_v f(x) = \frac{\mathrm{d}}{\mathrm{d}t} f[\gamma(t)]\big|_{t=0}$, where γ is any smooth curve such that $\gamma(0) = x$ and $\gamma'(0) = v$.

To compare two tangent vectors in a meaningful way, we consider a Riemannian manifold $(\mathcal{M}, g)$, which is a real smooth manifold equipped with an Riemannian metric g_x on the tangent space $T_x\mathcal{M}$ for any $x \in \mathcal{M}$ [50]. The inner product of any two tangent vectors $u_1, u_2 \in T_x\mathcal{M}$ is defined as $\langle u_1, u_2 \rangle_x = g_x(u_1, u_2)$, with an induced norm $\|u_1\|_x = \sqrt{\langle u_1, u_1 \rangle_x}$. We specify how to move a point along a direction in the following definition.

Definition B.2 (Exponential map)**.** The exponential map at x, denoted by Exp_x, sends any tangent vector $u \in T_x\mathcal{M}$ to $y = \gamma_u(1) \in \mathcal{M}$, where $\gamma_u : [0, 1] \to \mathcal{M}$ is the unique geodesic determined by $\gamma_u(0) = x$ and $\gamma_u'(0) = u$.

Moreover, since γ_u is the unique geodesic connecting x and y, the exponential map is invertible and we have $u = \mathrm{Exp}_x^{-1}(y)$. The distance between x and y satisfies $d(x, y) = [\langle \mathrm{Exp}_x^{-1}(y), \mathrm{Exp}_x^{-1}(y) \rangle_x]^{1/2}$, which is also called the geodesic distance. For any two points $x, y \in \mathcal{M}$, the parallel transport $\Gamma_x^y \colon T_x\mathcal{M} \to T_y\mathcal{M}$ specifies that how a tangent vector of x is identified with an element in $T_y\mathcal{M}$. Moveover, we have $\langle u, v \rangle_x =$

$\langle \Gamma_x^y u, \Gamma(\gamma)_x^y v\rangle_y$ for any $u, v \in T_x\mathcal{M}$. For simplicity, we denote $\boldsymbol{d}^2(u,v) = \langle u - \Gamma_y^x v, u - \Gamma_y^x v\rangle_x$ for any $x, y \in \mathcal{M}$ and any $u \in T_x\mathcal{M}$, $v \in T_y\mathcal{M}$. Below we define the gradient *w.r.t.* the Riemannian metric, in contrast with the Euclidean space.

Definition B.3 (Gradient)**.** Suppose there exists $u \in T_x\mathcal{M}$ such that $\nabla_v f(x) = \langle u, v\rangle_x$ for any $v \in T_x\mathcal{M}$, then f is differentiable at x, and u is called the gradient of f at x, denoted by $\operatorname{grad} f(x)$. We also define partial gradient of f at (x, y) *w.r.t.* x by $\operatorname{grad}_x f(x, y) = \operatorname{grad} f_y(x)$ as a natural extension, where $f_y(x) = f(x, y)$ for any fixed $y \in \mathcal{M}$. We call function f W_2-differentiable if $\operatorname{grad} f$ exists over its domain.

Definition B.4 (Geodesic space)**.** A metric space $(\mathcal{X}, d)$ consists of a set $\mathcal{X}$ and a distance function $d\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfying a few simple properties [5]. A curve γ on $\mathcal{X}$ is a continuous function from $[0,1]$ to $\mathcal{X}$, whose length is defined as $L(\gamma) = \sup \sum_{i=1}^n d[\gamma(t_{i-1}), \gamma(t_i)]$, where the supremum is taken over $n \geq 1$ and all partitions $0 = t_0 < t_1 < \ldots < t_n = 1$ of $[0,1]$. By this definition, for any curve γ, we have $L(\gamma) \geq d[\gamma(0), \gamma(1)]$. If there exists a constant $v \geq 0$ such that, $d[\gamma(t_1), \gamma(t_2)] = v \cdot |t_1 - t_2|$ for any $t_1, t_2 \in [0,1]$, then curve γ is called a geodesic. In this case, for any $0 \leq t_1 < t_2 \leq 1$, the length of γ restricted to $[t_1, t_2]$ is equal to $d[\gamma(t_1), \gamma(t_2)]$. Thus, a geodesic is everywhere locally a distance minimizer. Moreover, $(\mathcal{X}, d)$ is called a geodesic space if any two points $x, y \in \mathcal{X}$ are connected by a geodesic γ such that $\gamma(0) = x$ and $\gamma(1) = y$.

Definition B.5 (Geodesic convexity)**.** A function $f\colon \mathcal{M} \to \mathbb{R}$ is called geodesically convex if for any $x, y \in \mathcal{M}$ and a geodesic $\gamma\colon [0,1] \to \mathcal{M}$ such that $\gamma(0) = x$ and $\gamma(1) = y$, we have

$$f[\gamma(t)] \leq t \cdot f[\gamma(0)] + (1-t) \cdot f[\gamma(1)], \qquad \forall t \in [0,1]. \tag{B.1}$$

The following lemma characterizes the geodesic convexity using the gradient of f.

Lemma B.6. If $f : \mathcal{M} \to \mathbb{R}$ is differentiable, then it is geodesically convex if and only if

$$f(y) \geq f(x) + \left\langle \operatorname{grad} f(x), \operatorname{Exp}_x^{-1}(y)\right\rangle_x, \qquad \forall x, y \in \mathcal{M}. \tag{B.2}$$

Proof. For any $x, y \in \mathcal{M}$, let let $\gamma : [0,1] \to \mathcal{M}$ be the unique geodesic satisfying $\gamma(0) = x$ and $\gamma(1) = y$. By the definition of exponential map, we have $\operatorname{Exp}_x[\gamma'(0)] = y$, i.e., $\gamma'(0) = \operatorname{Exp}_x^{-1}(y)$. In addition, (B.1) shows that $f[\gamma(t)]$ is a convex and differentiable function on $[0,1]$, which implies

$$f[\gamma(1)] \geq f[\gamma(0)] + \frac{\mathrm{d}}{\mathrm{d}t} f[\gamma(t)]\Big|_{t=0}. \tag{B.3}$$

By the definition of directional derivative, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} f[\gamma(t)]\Big|_{t=0} = \nabla_{\gamma'(0)} f(x) = \left\langle \operatorname{grad} f(x), \gamma'(0)\right\rangle_x = \left\langle \operatorname{grad} f(x), \operatorname{Exp}_x^{-1}(y)\right\rangle_x. \tag{B.4}$$

Thus, combining (B.3) and (B.4) we obtain (B.2).

It remains to show (B.1) assuming (B.2) holds. For any geodesic $\gamma : [0,1] \to \mathcal{M}$, we will show that $f[\gamma(t)]$ is convex on $[0,1]$. To see this, for any $0 \leq t_1 \leq t_2 \leq 1$, let $x = \gamma(t_1)$ and $y = \gamma(t_2)$. Note that we can reparametrize γ to obtain a new geodesic $\widehat{\gamma}$ with $\widehat{\gamma}(0) = x$ and $\widehat{\gamma}(1) = y$ by letting

$$\widehat{\gamma}(t) = \gamma[t_1 + (t_2 - t_1) \cdot t]$$

for any $t \in [0,1]$. Since $\widehat{\gamma}$ is a geodesic, by the definition of exponential map, we have

$$\operatorname{Exp}_x^{-1}(y) = \widehat{\gamma}'(0) = (t_2 - t_1) \cdot \gamma'(t_1).$$

Thus, by (B.2), we have

$$\begin{aligned} f[\gamma(t_2)] = f(y) &\geq f(x) + \left\langle \operatorname{grad} f(x), \operatorname{Exp}_x^{-1}(y)\right\rangle_x \\ &= f[\gamma(t_1)] + \left\langle \operatorname{grad} f[\gamma(t_1)], (t_2 - t_1) \cdot \gamma'(t_1)\right\rangle_{\gamma(t_1)} = f[\gamma(t_1)] + (t_2 - t_1) \cdot \frac{\mathrm{d}}{\mathrm{d}t} f[\gamma(t)]\Big|_{t=t_1}, \end{aligned}$$

which implies that $f[\gamma(t)]$ is a convex function on $[0,1]$. Thus, (B.1) holds and we conclude the proof of this lemma. □

In the following, we extend the concepts of strong convexity and smoothness to manifold optimizaiton.

Definition B.7 (Geodesic strong convexity and smoothness)**.** For any $\mu > 0$, a differentiable function $f\colon \mathcal{M} \to \mathbb{R}$ is called geodesically μ-strongly convex if

$$f(y) \geq f(x) + \left\langle \operatorname{grad} f(x), \operatorname{Exp}_x^{-1}(y) \right\rangle_x + \mu/2 \cdot d^2(x,y),$$

where d is the distance function induced by the Riemannian metric. Moreover, f is said to be geodesically L-smooth if $\operatorname{grad} f$ is L-Lipschitz continuous. That is, for any $x, y \in \mathcal{M}$, we have

$$\left\langle \operatorname{grad} f(x) - \Gamma_y^x[\operatorname{grad} f(y)], \operatorname{grad} f(x) - \Gamma_y^x[\operatorname{grad} f(y)] \right\rangle_x \leq L^2 \cdot d^2(x,y), \tag{B.5}$$

where $\Gamma_y^x : \mathcal{T}_y\mathcal{M} \to \mathcal{T}_x\mathcal{M}$ is the parallel transport from the tangent space at y to that at x.

Note that we apply the parallel transport in (B.5) to compare $\operatorname{grad} f(x)$ and $\operatorname{grad} f(y)$, which belong to two different tangent spaces. In the following, we introduce the notion of gradient dominated function.

Definition B.8 (Gradient dominance)**.** Let $\mu > 0$ and $f\colon \mathcal{M} \to \mathbb{R}$ be a differentiable function with $f^* = \min_{x\in\mathcal{M}} f(x)$. Then f is μ-gradient dominated if

$$2\mu \cdot [f(x) - f^*] \leq \left\langle \operatorname{grad} f(x), \operatorname{grad} f(x) \right\rangle_x, \qquad \forall x \in \mathcal{M}. \tag{B.6}$$

In the following lemma, we show that, similar to functions in the Euclidean space, $\operatorname{grad} f$ being Lipschitz smooth implies that f can be upper bounded by the distance function. More importantly, we show that, gradient dominance is implied by geodesically strong convexity and thus is a weaker condition.

Lemma B.9. If $f : \mathcal{M} \to \mathbb{R}$ is geodesically μ-strongly convex, then f is also μ-gradient dominated. In addition, if f has L-Lipschitz continuous gradient, then, we have

$$f(y) \leq f(x) + \left\langle \operatorname{grad} f(x), \operatorname{Exp}_x^{-1}(y) \right\rangle_x + L/2 \cdot d^2(x,y), \qquad \forall x, y \in \mathcal{M}. \tag{B.7}$$

Proof. For the first part, let f be a geodesically μ-strongly convex function. Since $(\mathcal{M}, g)$ is a geodesic space, we have $d^2(x,y) = \langle \operatorname{Exp}_x^{-1}(y), \operatorname{Exp}_x^{-1}(y) \rangle_x$. Thus, by direct computation, we have

$$\begin{aligned} f(y) &\geq f(x) + \left\langle \operatorname{grad} f(x), \operatorname{Exp}_x^{-1}(y) \right\rangle_x + \mu/2 \cdot \left\langle \operatorname{Exp}_x^{-1}(y), \operatorname{Exp}_x^{-1}(y) \right\rangle_x \\ &= f(x) + \mu/2 \cdot \left\langle \operatorname{Exp}_x^{-1}(y) + 1/\mu \cdot \operatorname{grad} f(x), \operatorname{Exp}_x^{-1}(y) + 1/\mu \cdot \operatorname{grad} f(x) \right\rangle_x \\ &\quad - 1/(2\mu) \cdot \left\langle \operatorname{grad} f(x), \operatorname{grad} f(x) \right\rangle_x \\ &\geq f(x) - 1/(2\mu) \cdot \left\langle \operatorname{grad} f(x), \operatorname{grad} f(x) \right\rangle_x. \end{aligned} \tag{B.8}$$

Setting $y = x^*$ such that $f(x^*) = f^*$ in (B.8), we establish (B.6).

For the second part of Lemma B.9, for any $x, y \in \mathcal{M}$, let γ be the unique geodesic satisfying $\gamma(0) = x$ and $\gamma(1) = y$. Then we have $\operatorname{Exp}_x^{-1}(y) = \gamma'(0)$. Moreover, for any $t \in [0,1]$, note that $\gamma'(t) \in \mathcal{T}_{\gamma(t)}\mathcal{M}$. By the definition of the parallel transport, we have $\Gamma_{\gamma(t)}^x \gamma'(t) = \gamma'(0) = \operatorname{Exp}_x^{-1}(y)$. Thus, by (B.4) it holds that

$$\begin{aligned} f(y) - f(x) - \left\langle \operatorname{grad} f(x), \operatorname{Exp}_x^{-1}(y) \right\rangle_x &= f[\gamma(1)] - f[\gamma(0)] - \frac{\mathrm{d}}{\mathrm{d}t} f[\gamma(t)] \Big|_{t=0} \\ &= \int_0^1 \left\{ \left\langle \operatorname{grad} f[\gamma(t)], \gamma'(t) \right\rangle_{\gamma(t)} - \left\langle \operatorname{grad} f(x), \gamma'(0) \right\rangle_x \right\} \mathrm{d}t \\ &= \int_0^1 \left(\left\langle \Gamma_{\gamma(t)}^x \left\{ \operatorname{grad} f[\gamma(t)] \right\} - \operatorname{grad} f(x), \operatorname{Exp}_x^{-1}(y) \right\rangle_x \right) \mathrm{d}t, \end{aligned} \tag{B.9}$$

where in the last equality we transport the tangent vectors to $\mathcal{T}_x\mathcal{M}$. Besides, by (B.5) and Cauchy-Schwarz inequality, for any $z \in \mathcal{M}$, we have

$$\begin{aligned}
&\left|\left\langle \operatorname{grad} f(x) - \Gamma_z^x[\operatorname{grad} f(z)], \operatorname{Exp}_x^{-1}(y)\right\rangle_x\right| \\
&\quad \le \left\{\left\langle \operatorname{grad} f(x) - \Gamma_z^x[\operatorname{grad} f(z)], \operatorname{grad} f(x) - \Gamma_z^x[\operatorname{grad} f(z)]\right\rangle_x\right\}^{1/2} \cdot \left[\left\langle \operatorname{Exp}_x^{-1}(y), \operatorname{Exp}_x^{-1}(y)\right\rangle_x\right]^{1/2} \\
&\quad \le L \cdot d(x,y) \cdot d(z,x). \qquad \text{(B.10)}
\end{aligned}$$

Finally, combining (B.9) and (B.10), we have

$$f(y) - f(x) - \left\langle \operatorname{grad} f(x), \operatorname{Exp}_x^{-1}(y)\right\rangle_x \le L \cdot \int_0^1 d(x,y) \cdot d[\gamma(t), x] \,\mathrm{d}t = L/2 \cdot d^2(x,y),$$

where the last equality follows from the fact that $d[\gamma(t), x] = d[\gamma(t), \gamma(0)] = t \cdot d(x,y)$. Therefore, we establish (B.7) and conclude the proof of Lemma B.9. □

C Auxiliary Results

In this section, we collect a bunch of supportive proofs and concrete examples to characterize gradient descent over Wasserstein spaces by pushing particles in Euclidean spaces and to illustrate the feasibility of the variational form. First, we show that the Riemannain gradient in distribution spaces can be expressed in the functional gradient in vector variable spaces.

C.1 Proof of Proposition 3.1

Proof. We start from the definition of directional derivative to introduce the link between Riemannain gradients and derivatives *w.r.t.* ℓ_2-norm. For any $s \in T_\mu\mathcal{M}$ and any $\mu \in \mathcal{M}$ with corresponding density p, suppose $\gamma\colon [0,1] \to \mathcal{M}$ represents a curve satisfying $\gamma(0) = p$ and $\gamma'(0) = s$. Then, the directional derivative of F gives (Definition B.3)

$$\frac{\mathrm{d}}{\mathrm{d}t} F\big[\gamma(t)\big]\bigg|_{t=0} = \big\langle \operatorname{grad} F(\mu), s\big\rangle_\mu. \qquad \text{(C.1)}$$

Furthermore, by the chain rule of functional gradient of F with respect to the ℓ_2-Euclidean structure, the directional derivative at p in the direction of s can be expressed as

$$\frac{\mathrm{d}}{\mathrm{d}t} F\big[\gamma(t)\big]\bigg|_{t=0} = \int_{\mathcal{X}} \frac{\delta F}{\delta p}(x) \cdot s(x) \,\mathrm{d}x. \qquad \text{(C.2)}$$

On the other hand, let $u\colon \mathcal{X} \to \mathbb{R}$ be the unique solution to elliptic equation

$$-\operatorname{div}\big[p(x) \cdot (\nabla u)(x)\big] = s(x), \qquad \forall x \in \mathcal{X}, \qquad \text{(C.3)}$$

where div is the divergence operator on $\mathcal{X}$. Thus, plugging (C.2) into (C.3) we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} F\big[\gamma(t)\big]\bigg|_{t=0} &= -\int_{\mathcal{X}} \frac{\delta F}{\delta p}(x) \cdot \operatorname{div}\big[p(x) \cdot (\nabla u)(x)\big] \,\mathrm{d}x \qquad \text{(C.4)} \\
&= -\int_{\mathcal{X}} \left\{\operatorname{div}\left[p(x) \cdot \frac{\delta F}{\delta p}(x) \cdot (\nabla u)(x)\right] - \left\langle \nabla\left(\frac{\delta F}{\delta p}\right)(x), (\nabla u)(x)\right\rangle \cdot p(x)\right\} \mathrm{d}x,
\end{aligned}$$

where the second equation follows from integration by parts and the property of the divergence operator that

$$\operatorname{div}(f \cdot v) = \langle \nabla f, v\rangle + f \cdot \operatorname{div}(v)$$

holds for any scalar function f and any vector field v. It is well known that given some regularity condition, one can show that the first term on the right-hand side of (C.4) vanishes. For example, when $\mathcal{X}$ is a convex compact region with periodic boundary condition, this is implied by the divergence theorem [55]. Therefore,

again combining (C.1) and (C.4) we obtain that

$$\big\langle \operatorname{grad} F(\mu), s \big\rangle_{\mu} = \int_{\mathcal{X}} \left\langle \nabla \left(\frac{\delta F}{\delta p} \right)(x), (\nabla u)(x) \right\rangle \cdot p(x) \ \mathrm{d}x. \tag{C.5}$$

Meanwhile we rewrite (C.3) in $\operatorname{grad} F(\mu)$, *i.e.,* another tangent vector at μ, with $v \colon \mathcal{X} \to \mathbb{R}$ as the solution to elliptic equation

$$-\operatorname{div}\big[p(x) \cdot (\nabla v)(x)\big] = [\operatorname{grad} F(\mu)](x), \qquad \forall x \in \mathcal{X}.$$

Since both $\operatorname{grad} F(\mu)$ and s lie in the same tangent space $T_{\mu}\mathcal{M}$, their inner product is characterized by the Riemannian metric on $(\mathcal{M}, W_2)$, that is,

$$\big\langle \operatorname{grad} F(\mu), s \big\rangle_{\mu} = \int_{\mathcal{X}} \Big\langle (\nabla u)(x), (\nabla v)(x) \Big\rangle \cdot p(x) \ \mathrm{d}x. \tag{C.6}$$

Now combining (C.5) and (C.6) we have that

$$\int_{\mathcal{X}} \left\langle \nabla \left(\frac{\delta F}{\delta p} \right)(x), (\nabla u)(x) \right\rangle \cdot p(x) \ \mathrm{d}x = \big\langle \operatorname{grad} F(p), s \big\rangle_{p} = \int_{\mathcal{X}} \Big\langle (\nabla u)(x), (\nabla v)(x) \Big\rangle \cdot p(x) \ \mathrm{d}x \tag{C.7}$$

holds for any $s \in T_{\mu}\mathcal{M}$. Since s is arbitrarily picked, (C.7) indicates that $\nabla(\delta F/\delta p) = \nabla v$. To conclude, we have

$$\operatorname{grad} F(\mu) = -\operatorname{div}(p \nabla v) = -\operatorname{div}\left[p \cdot \nabla \left(\frac{\delta F}{\delta p} \right) \right], \tag{C.8}$$

which implies the desired result of (3.2).

In what follows, we turn to obtain the explicit form of $\delta F/\delta p$ for F defined in (2.4), while writing F as a functional $F(p)$ in p. Following the definition of the functional derivative by limits with respect to the ℓ_2-Euclidean structure, for any square-integrable function $\varphi \colon \mathcal{X} \to \mathbb{R}$ in $\mathcal{F}$, we can write

$$\int_{\mathcal{X}} \frac{\delta F}{\delta p}(x) \cdot \varphi(x) \ \mathrm{d}x = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \cdot \big[F(p + \epsilon \cdot \varphi) - F(p) \big]. \tag{C.9}$$

For simplicity of notations, we denote by f_{ϵ}^{*} the optimal dual solution to the optimization problem

$$\sup_{f \in \mathcal{F}} \left\{ \int_{\mathcal{X}} f(x) \cdot [p(x) + \epsilon \cdot \varphi(x)] \ \mathrm{d}x - F^{*}(f) \right\} \tag{C.10}$$

for any ϵ sufficiently small. Then by setting f to f_{p}^{*} in the definition of the variational form for $F(p + \epsilon \cdot \varphi)$, the difference in (C.9) satisfies the following lower bound,

$$\begin{aligned} & F(p + \epsilon \cdot \varphi) - F(p) \\ & \quad \geq \left[\int_{\mathcal{X}} f_{p}^{*}(x) \cdot [p(x) + \epsilon \cdot \varphi(x)] \ \mathrm{d}x - F^{*}(f_{p}^{*}) \right] - \left[\int_{\mathcal{X}} f_{p}^{*}(x) \cdot p(x) \ \mathrm{d}x - F^{*}(f_{p}^{*}) \right] \\ & \quad = \epsilon \cdot \int_{\mathcal{X}} f_{p}^{*}(x) \cdot \varphi(x) \ \mathrm{d}x. \end{aligned} \tag{C.11}$$

Meanwhile, we obtain an upper bound of $F(p + \epsilon \cdot \varphi) - F(p)$ by using f_{ϵ}^{*} for both variational maximization objectives as follows,

$$\begin{aligned} & F(p + \epsilon \cdot \varphi) - F(p) \\ & \quad \leq \left[\int_{\mathcal{X}} f_{\epsilon}^{*}(x) \cdot [p(x) + \epsilon \cdot \varphi(x)] \ \mathrm{d}x - F^{*}(f_{\epsilon}^{*}) \right] - \left[\int_{\mathcal{X}} f_{\epsilon}^{*}(x) \cdot p(x) \ \mathrm{d}x - F^{*}(f_{\epsilon}^{*}) \right] \\ & \quad = \epsilon \cdot \int_{\mathcal{X}} f_{\epsilon}^{*}(x) \cdot \varphi(x) \ \mathrm{d}x. \end{aligned} \tag{C.12}$$

Combining (C.11) and (C.12) with the sandwich theorem of limits [59], we obtain that $F(p + \epsilon \cdot \varphi) - F(p)$ tends to zero as ϵ goes to zero. Nevertheless, to derive a characterization for the right hand side of (C.9), it remains to quantify some distance measure between f_{p}^{*} and f_{ϵ}^{*}. Fortunately, since F^{*} is strongly convex, we

are able to set a constant $\gamma > 0$ such that, for any two measurable functions f_1 and f_2, we have

$$\int_{\mathcal{X}} \left[\frac{\partial F^*}{\partial f_1}(x) - \frac{\partial F^*}{\partial f_2}(x) \right] \cdot [f_1(x) - f_2(x)] \, \mathrm{d}x \geq \gamma \cdot \int_{\mathcal{X}} |f_1(x) - f_2(x)|^2 \, \mathrm{d}x. \tag{C.13}$$

Moreover, since f_p^* and f_ϵ^* are maximizers of the optimization problems in (2.4) and (C.10), respectively, we observe that $\partial F^*/\partial f_p^* = p$ and $\partial F^*/\partial f_\epsilon^* = p + \epsilon \cdot \phi$. Hence, by applying Cauchy-Schwarz inequality to the left hand side of (C.13), we have

$$\gamma \cdot \int_{\mathcal{X}} (f_\epsilon^* - f_p^*)^2 \, \mathrm{d}x \leq \epsilon \cdot \int_{\mathcal{X}} \varphi \cdot (f_\epsilon^* - f_p^*) \, \mathrm{d}x \leq \epsilon \cdot \left(\int_{\mathcal{X}} \varphi^2 \, \mathrm{d}x \right)^{1/2} \cdot \left[\int_{\mathcal{X}} (f_\epsilon^* - f_p^*)^2 \, \mathrm{d}x \right], \tag{C.14}$$

which implies that $\|f_\epsilon^* - f_p^*\|_{\ell_2} \leq \epsilon \cdot \|\varphi\|_{\ell_2}$ and consequently f_ϵ^* converges to f_p^* as ϵ tends to zero. Then, the sandwich theorem of limits can be applied to the whole right hand side of (C.9), by plugging (C.11) and (C.12) into (C.9), to obtain that

$$\int_{\mathcal{X}} \frac{\delta F}{\delta p}(x) \cdot \varphi(x) \, \mathrm{d}x = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \cdot \big[F(p + \epsilon \cdot \varphi) - F(p) \big] = \int_{\mathcal{X}} f_p^*(x) \cdot \varphi(x) \, \mathrm{d}x$$

holds for any $\varphi \in \mathcal{F}$, which implies the result $\delta F/\delta p = f_p^*$.

To conclude, by combining (3.2) for F defined in (2.4), we obtain the final explicit form of the Riemannian gradient $\operatorname{grad} F = -\operatorname{div}[p \cdot \nabla(f_p^*)]$. Therefore, the proof of this proposition is completed. □

C.2 Proof of Proposition 3.2

Proof. According to the definition of exponential maps, to prove (3.3), it is sufficient to show the following results for a curve $\gamma \colon [0, 1/h) \to \mathcal{M}$ defined by setting $\gamma(t) = [\operatorname{Exp}_{\mathcal{X}}(t \cdot \nabla u)]_\sharp \mu$ for all $t \in [0, 1/h)$, where $u \in \mathcal{F}$ and h denotes the Lipschitz constant of ∇u.

(i) $\gamma(0) = \mu$.

(ii) $\gamma'(0) = s$.

(iii) $\gamma(t)$ is a geodesic on $\mathcal{M}$ for $t \in [0, 1/h)$.

Please note that $\gamma(0) = \mathrm{id}_\sharp \, \mu = \mu$, where $\mathrm{id} \colon \mathcal{X} \to \mathcal{X}$ is the identity mapping. First of all, we will adopt the following lemma to demonstrate that $\gamma(t)$ is a geodesic on $\mathcal{M}$.

Lemma C.1. Suppose that $\mathcal{X}$ denotes $\mathbb{R}^d$ or a closed convex subset of $\mathbb{R}^d$ with periodic boundary conditions. Let $u : \mathcal{X} \to \mathcal{X}$ be a twice continuously differentiable function over $\mathcal{X}$ with a h-Lipschitz continuous gradient $\nabla u \colon \mathcal{X} \to \mathcal{X}$. Then, for any $\mu \in \mathcal{M}$, a curve $\gamma \colon [0, 1/h) \to \mathcal{M}$ defined by $\gamma(t) = [\operatorname{Exp}_{\mathcal{X}}(t \cdot \nabla u)]_\sharp \mu$ is a geodesic on $\mathcal{M}$.

Proof. We consider separately the two cases where (i) $\mathcal{X}$ is $\mathbb{R}^d$ or (ii) $\mathcal{X}$ denotes a subset of $\mathbb{R}^d$ with periodic boundary condition. The former case is first considered, where $\gamma(t)$ can be formulated as (C.18).

(i) $\mathcal{X}$ is $\mathbb{R}^d$. For completeness, we first need to verify that γ according to the definition is indeed a curve in $\mathcal{M}$, *i.e.,* the pushforward maps admit $\gamma(t) \in \mathcal{M}$ for all $t \in [0, 1/h)$. To proceed in this direction, we define a potential function

$$\varphi_t(x) = \|x\|_2^2/2 + t \cdot u(x) \tag{C.15}$$

for all $t \in [0, 1/h)$. As a result, we have the corresponding vector field $\nabla \varphi_t = \mathrm{id} + t \cdot \nabla u$ exactly induced by our pushforward mapping in the definition, which indicates that $\gamma(t) = [\nabla \varphi_t]_\sharp \mu$. Since ∇u is h-Lipschitz continuous, we observe that φ_t is strongly convex for all $t \in [0, 1/h)$. In addition, φ_t is also twice continuously differentiable as u possesses continuous second-order derivative, which implies the Jacobian of $\nabla \varphi_t$, *i.e.,*, $\nabla^2 \varphi_t$ is continuous and positive definite. Therefore, we conclude that $\nabla \varphi_t \colon \mathbb{R}^d \to \mathbb{R}^d$ is an

invertible mapping. Thus, $[\nabla\varphi_t]_\sharp\mu$ still lies in the distribution family which admits absolute continuity with respect to the Lebesgue measure and positive density everywhere. To summarize, it has been shown that $\gamma(t) \in \mathcal{M}$ for all $t \in [0, 1/h)$.

It follows to prove that the curve γ on $\mathcal{M}$ is a geodesic. To show this, we apply the Brenier's Theorem (see, for example, Theorem 2.12 in [62]) to conclude that there exists a unique optimal transport plan between μ and $\gamma(t)$, which can be written as the gradient of a convex function φ. Meanwhile, the theorem further implies that $\nabla\varphi$ serves as the unique gradient of some convex function such that $\gamma(t)$ can be expressed as $[\nabla\varphi]_\sharp p$. Combining with the definition that $\gamma(t) = [\nabla\varphi_t]_\sharp\mu$ for all $t \in [0, 1/h)$, we assert that $\nabla\varphi_t$ represents the optimal transportation plan between p and $\gamma(t)$. With such useful characterization, we fix any $\bar{t} \in [0, 1/h)$ below and show that γ is a geodesic when confined to $[0, \bar{t}]$. It is observed that

$$\nabla\varphi_t = \mathrm{id} + t \cdot \nabla u = [1 - (t/\bar{t})] \cdot \mathrm{id} + (t/\bar{t}) \cdot \nabla\varphi_{\bar{t}},$$

holds for any $t \in [0, \bar{t}]$, where $\varphi_{\bar{t}}$ is a strongly convex function. Hence, we can write $\gamma(t)$ in terms of $\{[1 - (t/\bar{t})] \cdot \mathrm{id} + (t/\bar{t}) \cdot \nabla\varphi_{\bar{t}}\}_\sharp\mu$. To compute the Wasserstein distance of two points on the curve, for any $0 \le t_1 < t_2 \le \bar{t}$, we have

$$\begin{aligned}
&W_2[\gamma(t_1), \gamma(t_2)] \\
&= \left[\int_{\mathcal{X}} \left\| \left\{[1 - (t_1/\bar{t})]x + (t_1/\bar{t}) \cdot \nabla\varphi_{\bar{t}}(x)\right\} - \left\{[1 - (t_2/\bar{t})]x + (t_2/\bar{t}) \cdot \nabla\varphi_{\bar{t}}(x)\right\} \right\|^2 \mathrm{d}p(x)\right]^{1/2} \\
&= (t_2 - t_1)/\bar{t} \cdot \left[\int_{\mathbb{R}^d} \left\| x - \nabla\varphi_{\bar{t}}(x) \right\|^2 \mathrm{d}p(x)\right]^{1/2} \\
&= (t_2 - t_1)/\bar{t} \cdot W_2[\mu, \gamma(\bar{t})], \qquad \text{(C.16)}
\end{aligned}$$

which implies that $\{\gamma(t)\}_{t\in[0,\bar{t}]}$ is a reparametrized geodesic. Since $\bar{t}$ is arbitrarily chosen within $[0, 1/h)$, it then holds that $\gamma(t)$ is a geodesic for $0 \le t < 1/h$.

(ii) $\mathcal{X}$ denotes a subset of $\mathbb{R}^d$ with periodic boundary condition. We are left to prove the lemma for the case where $\mathcal{X}$ is a closed convex subset of $\mathbb{R}^d$ with periodic boundary condition. In this case, any $x \in \mathcal{X}$ can be identified with an equivalence class of $\mathbb{R}^d$. Moreover, each probability measure $\mu \in \mathcal{P}(\mathcal{X})$ is unique identified with a periodic measure $\bar{\mu} \in \mathcal{P}(\mathbb{R}^d)$ such that $\bar{\mu}$ coincides with μ on $\mathcal{X}$. We say $\bar{\mu}$ is the periodic extension of μ. Since μ is absolutely continuous with respect to the Lebesgue measure and admits positive density, so is $\bar{\mu}$. Moreover, $u\colon \mathcal{X} \to \mathbb{R}$ can also be extended as a periodic function on $\mathbb{R}^d$, and $\varphi_t(x) = \|x\|_2^2/2 + t \cdot u(x)$ is a strongly convex, twice continuously differentiable, and periodic function on $\mathbb{R}^d$. Then it can be shown that $(\mathrm{id} + t \cdot \nabla u)_\sharp\bar{\mu}$ is the periodic extension of $[\mathrm{Exp}_{\mathcal{X}}(t \cdot \nabla u)]_\sharp\mu$ [6]. Thus, these two measures coincide on $\mathcal{X}$, i.e.,

$$[\mathrm{Exp}_{\mathcal{X}}(t \cdot \nabla u)]_\sharp\mu = (\mathrm{id} + t \cdot \nabla u)_\sharp\bar{\mu}\big|_{\mathcal{X}}, \qquad \text{(C.17)}$$

where $\cdot|_{\mathcal{X}}$ denotes the restriction to $\mathcal{X}$. Note that we have shown that $[\nabla\phi_t]_\sharp\bar{\mu}$ is absolutely continuous with respect to the Lebesgue measure and has positive density. Thus, by restricting $[\nabla\phi_t]_\sharp\bar{\mu}$ to $\mathcal{X}$, (C.17) implies that $\gamma(t) \in \mathcal{M}$ for all $t \in [0, 1/h)$, i.e., γ is a curve on $\mathcal{M}$.

Furthermore, to show that γ is a geodesic, we utilize the generalization of Brenier's theorem to probability distributions over a Riemannian manifold [24]. For any $t \in [0, 1/h)$, since $\gamma(t) \in \mathcal{M}$, there exists a unique optimal transportation plan $\Upsilon\colon \mathcal{X} \to \mathcal{X}$ between μ and $\gamma(t)$ such that $\gamma(t) = \Upsilon_\sharp\mu$. Moreover, Υ takes the form of $\Upsilon(x) = \mathrm{Exp}_x[-\nabla\psi(x)]$ for some $\psi\colon \mathcal{X} \to \mathbb{R}$ such that $\|x\|_2^2/2 - \psi(x)$ is convex. Hence, due to the uniqueness and the fact that ∇u is h-Lipschitz, $\mathrm{Exp}_{\mathcal{X}}(-t \cdot \nabla u)$ is the optimal transportation plan between μ and $\gamma(t)$.

Similarly in what follows we fix any $\bar{t} \in [0, 1/H)$ and show that γ is a geodesic for $t \in [0, \bar{t}]$. For any

$0 \le t_1 < t_2 \le \bar{t}$, following the derivations in (C.16) and combining (C.17), we have

$$\begin{aligned} W_2[\gamma(t_1), \gamma(t_2)] &= W_2\big\{[\mathrm{Exp}_{\mathcal{X}}(t_1 \cdot \nabla u)]_\sharp \mu, [\mathrm{Exp}_{\mathcal{X}}(t_2 \cdot \nabla u)]_\sharp \mu\big\} \\ &= W_2\Big\{(\mathrm{id} + t_1 \cdot \nabla u)_\sharp \bar{\mu}\big|_{\mathcal{X}}, (\mathrm{id} + t_2 \cdot \nabla u)_\sharp \bar{\mu}\big|_{\mathcal{X}}\Big\} = (t_2 - t_1)/\bar{t} \cdot W_2\Big\{\bar{\mu}\big|_{\mathcal{X}}, (\mathrm{id} + t_2 \cdot \nabla u)_\sharp \bar{\mu}\big|_{\mathcal{X}}\Big\} \\ &= (t_2 - t_1)/\bar{t} \cdot W_2[\mu, \gamma(\bar{t})], \end{aligned}$$

where the second and the last equality follows from (C.17). Thus, we obtain that $\{\gamma(t)\}_{t\in[0,\bar{t}]}$ is a geodesic up to reparametrization, which concludes the proof of Lemma C.1. □

With the required Lemma in place, to finish the proof of Proposition 3.2, it remains to show that $\gamma'(0) = s$. Similar to the proof of the above lemma, we distinguish the two cases where $\mathcal{X}$ is $\mathbb{R}^d$ and $\mathcal{X}$ is a closed convex subset of $\mathbb{R}^d$ with periodic boundary condition. In the former case,

$$[\mathrm{Exp}_{\mathcal{X}}(t \cdot \nabla u)]_\sharp \mu = (\mathrm{id} + t \cdot \nabla u)_\sharp \mu, \qquad \forall t \in [0, 1/h). \tag{C.18}$$

To simplify the notation, we denote $T_t = \mathrm{id} + t \cdot \nabla u$, which is invertible for $t \in [0, 1/h)$. By the definition of pushforward measures, we have

$$\gamma(t)(x) = \big[(T_t)_\sharp \mu\big](x) = \mu\big[T_t^{-1}(x)\big] \cdot \left|\frac{\mathrm{d}}{\mathrm{d}x}\big[T_t^{-1}(x)\big]\right|,$$

where the second equality follows from the change-of-variable formula and $|\frac{\mathrm{d}}{\mathrm{d}x}[T_t^{-1}(x)]|$ is the determinant of the Jacobian. Moreover, when t is sufficiently small, for any $x \in \mathcal{X}$, Taylor expansion in t yields that $T_t^{-1}(x) = x - t \cdot (\nabla u)(x) + o(t)$, which further implies that

$$\begin{aligned} \gamma(t)(x) &= \mu\big[x - t \cdot (\nabla u)(x) + o(t)\big] \cdot \big|I_d - t \cdot (\nabla^2 u)(x) + o(t)\big| \\ &= \big[\mu(x) - t \cdot \big\langle(\nabla u)(x), (\nabla \mu)(x)\big\rangle + o(t)\big] \cdot \big\{1 - t \cdot \mathrm{Tr}[(\nabla^2 u)(x)] + o(t)\big\} \\ &= \mu(x) - t \cdot \big\langle(\nabla u)(x), (\nabla \mu)(x)\big\rangle - t \cdot \mathrm{Tr}[(\nabla^2 u)(x)] \cdot \mu(x) + o(t). \end{aligned} \tag{C.19}$$

where in the second equality we apply Taylor expansion to $\mu[x - t \cdot (\nabla u)(x) + o(t)]$. Moreover, since

$$\big\langle(\nabla u)(x), (\nabla p)(x)\big\rangle + \mathrm{Trace}[(\nabla^2 u)(x)] \cdot p(x) = \mathrm{div}\big[p(x) \cdot (\nabla u)(x)\big],$$

by (C.19) we obtain that

$$\gamma(t)(x) = \mu(x) - t \cdot \mathrm{div}\big[p(x) \cdot (\nabla u)(x)\big] + o(t),$$

which implies that $\gamma'(0) = -\mathrm{div}(p \cdot \nabla u) = s$.

It remains to show $\gamma'(0) = -\mathrm{div}(p \cdot \nabla u) = s$ when $\mathcal{X}$ is a closed compact subset of $\mathbb{R}^d$ with periodic boundary condition. As shown in the proof of Lemma C.1, p can be periodically extended to a measure $\bar{\mu}$ on $\mathbb{R}^n$ and that such an extension is unique. Furthermore, the solution u to the elliptic equation $-\mathrm{div}(-p \cdot \nabla u) = s$ can also also be viewed as a periodic function on $\mathbb{R}^d$. Then it can be shown that $(\mathrm{id} + t \cdot \nabla u)_\sharp \bar{\mu}$ is the periodic extension of $[\mathrm{Exp}_{\mathcal{X}}(t \cdot \nabla u)]_\sharp \mu$ and that (C.17) holds. Note that we have shown that $\widetilde{\gamma}(t) = (\mathrm{id} + t \cdot \nabla u)_\sharp \bar{\mu}$ satisfies that $\widetilde{\gamma}(0) = \bar{\mu}$ and $\widetilde{\gamma}'(0) = s$. Therefore, restricting to $\mathcal{X}$, we conclude that $\gamma'(t) = s$, which completes the proof of the proposition. □

C.3 Examples for functionals with the variational form

The following example shows that if the entropy functional is f-divergence, the conjugate function F^* will be strongly smooth when the link function ψ is strongly convex, with respect to ℓ_2-norm.

Example C.2. Let p, q be two density function over a compact domain $\mathcal{X}$, then the f-divergence

$$I_\psi(p, q) = \int_{\mathcal{X}} p(x) \cdot \psi\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x \tag{C.20}$$

with a strongly convex and smooth function ψ admits the following variational forms (as functionals of p and q respectively),

$$I_\psi(p,q) = I_{\psi,q}(p) = \sup_{f\in\mathcal{F}_p}\left\{\int_{\mathcal{X}} f(x)p(x)\,\mathrm{d}x - F_q^*(f)\right\}, \tag{C.21}$$

$$I_\psi(p,q) = I_{\psi,p}(q) = \sup_{f\in\mathcal{F}_q}\left\{\int_{\mathcal{X}} f(x)q(x)\,\mathrm{d}x - F_p^*(f)\right\}, \tag{C.22}$$

where

$$F_q^*(f) = \int_{\mathcal{X}} -(\psi^*)^{-1}(-f(x))q(x)\,\mathrm{d}x \tag{C.23}$$

and

$$F_p^*(f) = \int \psi^*(f(x))p(x)\,\mathrm{d}x \tag{C.24}$$

are strongly convex and smooth functionals, $\mathcal{F}_p$ and $\mathcal{F}_q$ are the same set of all measurable functions on $\mathcal{X}$. We denote by γ its strong convexity parameter and by L the smoothness parameter.

Proof. Following [45], Fenchel convex duality [54] ensures that we have $I_\psi(p,q)$ in the form of the conjugate of ψ as:

$$I_\psi(p,q) = \sup_g\left\{\int g(x)q(x)\,\mathrm{d}x - \int \psi^*(g(x))p(x)\,\mathrm{d}x\right\}, \tag{C.25}$$

where ψ^* is a strongly convex and smooth function since the convexity and smoothness of ψ, and the superemum is taken over all measurable function $g : \mathcal{X} \to \mathbb{R}$. Hence by replacing g with f formally we obtain (C.22).

Moreover, by letting $f(x) = -\psi^*(g(x))$, we have

$$I_\psi(p,q) = \sup_f\left\{\int f(x)p(x)\,\mathrm{d}x - \int -(\psi^*)^{-1}(-f(x))q(x)\,\mathrm{d}x\right\}. \tag{C.26}$$

By setting $F^*(f) = \int -(\psi^*)^{-1}(-f(x))q(x)\,\mathrm{d}x$, which can be verified to be strongly convex and smooth *w.r.t.* function f, we show the desired result in (C.21). □

Please note that by convex duality and optimality condition in Lemma 1 of [45], the function class $\mathcal{F}_p$ ($\mathcal{F}_q$) over which the superemum is taken can be restricted to a smaller one $\widetilde{\mathcal{F}}$ as long as $\widetilde{\mathcal{F}}$ contains the differential $\partial I_\psi(p,q)/\partial p$ ($\partial I_\psi(p,q)/\partial q$), In this sense, Assumption 4.1 for a smaller function class $\widetilde{\mathcal{F}}$ is feasible since it is natural for the smooth (with continuous first-order gradient and Hessian) function $\partial I_\psi(p,q)/\partial q$ over a compact domain $\mathcal{X}$ to possess a closed, bounded and equicontinuous gradient mapping by Weierstrass theorem [10].

D Proof of Convergence of Variational Form Maximization

We first illustrate our reverse Poincaré inequality with more details. To this end, we restate Lemma 4.2 and prove with a compact domain $\mathcal{X}$ as follows.

D.1 Approximation function class and reverse Poincaré inequality

Recall that $\widetilde{\mathcal{F}}$ is the function class of NNs, defined in (3.5).

Lemma D.1 (Restatement of Lemma 4.2)**.** We consider a function class $\widetilde{\mathcal{F}}$ that is $\nabla\widetilde{\mathcal{F}}$ is closed, bounded, and equicontinuous. More precisely,

(a) "$\nabla\widetilde{\mathcal{F}}$ is bounded" means that there exists a positive constant $M < \infty$ such that $\|\nabla f(x)\| \leq M$ for each $x \in \mathcal{X}$ and each $f \in \widetilde{\mathcal{F}}$.

(b) "$\nabla\widetilde{\mathcal{F}}$ is equicontinuous" implies that for every $\epsilon > 0$ there exists $\delta > 0$ (which depends only on ϵ) such that for any $x, y \in \mathcal{X}$ with the metric $d(\cdot,\cdot)$, if $d(x,y) < \delta$, then it follows that

$$\|\nabla f(x) - \nabla f(y)\| < \epsilon, \qquad \forall f \in \widetilde{\mathcal{F}}. \tag{D.1}$$

Then we have for every $p \geq 1$, there exists a constant $\widetilde{K}$ such that

$$\int_{\mathcal{X}} \|\nabla f(x)\|^p d\mu \leq \widetilde{K} \int_{\mathcal{X}} |f(x)|^p d\mu, \tag{D.2}$$

for any $f \in \widetilde{\mathcal{F}}$, where $\mathcal{X}$ is a compact subset of a metric space, μ is a nonnegative measure over $\mathcal{X}$.

Proof of Lemma 4.2

Proof. With the notations of $L^p_\mu(\mathcal{X})$ norm, (D.2) can be rewritten as

$$\|\nabla f\|_{L^p_\mu(\mathcal{X})} \leq \widetilde{K} \|f\|_{L^p_\mu(\mathcal{X})}, \tag{D.3}$$

On the other hand, Poincaré inequality (See, e.g., Chapter 5 of [17]) claims that there exists some constant K' such that

$$\|f - \frac{1}{\mu(\mathcal{X})} \int_{\mathcal{X}} f d\mu\|_{L^p_\mu(\mathcal{X})} \leq K' \|\nabla f\|_{L^p_\mu(\mathcal{X})}, \tag{D.4}$$

We define an equivalence relation $\sim$ on $L^p_\mu(\mathcal{X})$ such that for any $f, g \in L^p_\mu(\mathcal{X})$, we have

$$f \sim g \qquad \text{if and only if} \qquad f - g = C \tag{D.5}$$

for some constant C. Then we denote by $\widetilde{L}^p_\mu(\mathcal{X}) = L^p_\mu(\mathcal{X})/\sim$ the new function space consisting of equivalence classes $\widetilde{f}$'s of locally summable functions $f : \mathcal{X} \to \mathbb{R}$. Consequently, we define the norm in $\widetilde{L}^p_\mu(\mathcal{X})$ as $\|\widetilde{f}\|_{\widetilde{L}^p_\mu(\mathcal{X})} = \|f\|_{L^p_\mu(\mathcal{X})}$, and a new gradient operator $\widetilde{\nabla} : \widetilde{L}^p_\mu(\mathcal{X}) \to \widetilde{L}^p_\mu(\mathcal{X})$ such that for any $\widetilde{f} \in \widetilde{L}^p_\mu(\mathcal{X})$, $\widetilde{\nabla}$ acts as $\widetilde{\nabla}\widetilde{f} = \nabla f$. Moreover, the inverse of $\widetilde{\nabla}$ exists, denoted by $\widetilde{\nabla}^{-1}$.

Therefore, (D.4) implies that the inverse gradient operator $\widetilde{\nabla}^{-1}$ is continuous. Our goal is reduced to showing that the gradient operator is continuous, *i.e.*, the preimage of open sets in $\widetilde{L}^p_\mu(\mathcal{X})$ under such mapping are also open in $\widetilde{L}^p_\mu(\mathcal{X})$.

According to (D.2) and the properties of continuous mappings between topological spaces (See, e.g., Appendix of [34]), for every compact subset $X \subseteq \widetilde{L}^p_\mu(\mathcal{X})$, the image (preimage of $\widetilde{\nabla}$) under $\widetilde{\nabla}^{-1}$, *i.e.*, $\widetilde{\nabla}^{-1}(X)$, is compact in $\widetilde{L}^p_\mu(\mathcal{X})$, and closed in $\widetilde{L}^p_\mu(\mathcal{X})$ since $\widetilde{L}^p_\mu(\mathcal{X})$ is a Hausdorff space. On the other hand, by Arzela-Ascoli Theorem the set $\widetilde{L}^p_\mu(\mathcal{X}) \cap \nabla\widetilde{\mathcal{F}}$ is compact. Then for every closed subset $Y \subseteq \widetilde{L}^p_\mu(\mathcal{X}) \cap \nabla\widetilde{\mathcal{F}}$, we have Y is also a compact subset, of which the image under $\widetilde{\nabla}^{-1}$ is closed in $\widetilde{L}^p_\mu(\mathcal{X})$. According to the definition the continuity of $\widetilde{\nabla}$ is shown, hence (D.2) is proved. □

Remark D.2. Specifically, we remark that the conditions above which enable the gradient bound (Lemma 4.2) to hold keep consistent with all the other restrictions on function class $\widetilde{\mathcal{F}}$, and can be easily realized by a general class of neural networks.

First, the small function class $\widetilde{\mathcal{F}}$ is designed to solve VFM defined in (2.4) numerically efficiently and meanwhile to provide a decent characterization of Riemannian gradient estimation error bounds. We also observe that to achieve an equivalence between pushing particles and the exponential map, functions in this smaller class are also required to admit uniformly h-Lipschitz continuous gradients. Fortunately, the equicontinuity of $\nabla\widetilde{\mathcal{F}}$ can be implied by h-Lipschitzness of $\nabla f \in \nabla\widetilde{\mathcal{F}}$. In other words, the overall assumptions of function class $\widetilde{\mathcal{F}}$ for our whole analysis are concluded as below.

Assumption D.3 (Overall assumptions of approximation function class)**.** The class of functions $\widetilde{\mathcal{F}}$ over a compact domain $\mathcal{X}$ satisfies the condition that $\nabla\widetilde{\mathcal{F}}$ is closed, bounded, and for each $f \in \widetilde{\mathcal{F}}$, ∇f is h-Lipschitz continuous with $h > 0$.

Then we check that our neural network parametrized function class is indeed qualified for such $\widetilde{\mathcal{F}}$. Recall that without loss of generality, we parametrize a function $f : \mathcal{X} \to \mathbb{R}$, *i.e.,* the decision variable for VFM, by the class of two-layer neural networks below, which is denoted as $\mathrm{NN}(\beta; w)$, with each member function as

$$f_\beta(x) = \frac{1}{\sqrt{w}} \sum_{i=1}^{w} b_i \cdot \sigma([\beta]_i^\top x), \tag{D.6}$$

where $x \in \mathcal{X}$ denotes the input data point, w gives the width of the neural network, $b_i \in \{-1, 1\}$ for $i \in [w]$ denotes the output weights, $\sigma(\cdot)$ is a smooth activation function, and $\beta = ([\beta]_1^\top, \ldots, [\beta]_w^\top)^\top \in \mathbb{R}^{wd}$ with $[\beta]_i \in \mathbb{R}^d$ $(i \in [w])$ are the overall input weights. For initialization, we consider the random strategy

$$b_i \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\{-1, 1\}), \quad [\beta(0)]_i \overset{\text{i.i.d.}}{\sim} N(0, I_d/(wd)), \text{ for all } i \in [w]. \tag{D.7}$$

Note that for technical analysis reason, we restrict the input weights β to an bounded ℓ_2-ball centered at the initializer $\beta(0)$ by an additional projection step $\Pi_{\mathcal{B}^0(r_f)}(\bar{\beta}) = \mathrm{argmin}_{\beta \in \mathcal{B}^0(r_f)}\{\|\beta - \bar{\beta}\|_2\}$, where $\mathcal{B}^0(r_f) = \{\beta : \|\beta - \beta(0)\|_2 \le r_f\}$. During the training process, we only backpropagate *w.r.t.* β, while keeping b_i $(i \in [w])$ intact at the initialization which accounts for the feasibility to omit the dependency on b_i $(i \in [w])$ in $\mathrm{NN}(\beta; w)$ and $f_\beta(x)$ in what follows. Therefore, we can directly observe that each $f_\beta \in \mathrm{NN}(\beta; w)$ is closed, and the ℓ_2-norm of the gradient $\nabla_\beta f_\beta$ is always bounded over a compact domain $\mathcal{X}$. Furthermore, we require the Hessian for f_β *w.r.t.* parameter $\beta \in \mathbb{R}^{wd}$ to possess a bounded norm, which is easy to check for neural networks. It turns out that the qualified function class encompasses a wide range of normal neural networks without "sharp points" or "high frequency oscillation" as functions.

D.2 Statistical error of approximating the solution of VFM

In what follows, we proceed to derive the gradient error bound. According to (D.2) of Lemma 4.2, we turn to bound the approximation error of a two-layer neural network for the following variational form maximization (VFM) problem defined in (3.4),

$$\max_{\beta \in \mathbb{R}^{wd}} \Big\{ \frac{1}{N} \sum_{i=1}^{N} f_\beta(x_i) - F^*(f_\beta) \Big\}. \tag{D.8}$$

Note that for notational simplicity and applicability of results to both players, in this section we omit the references to μ and ν. To emphasize we fixed the b_i's throughout the training, we write $f_\beta(x)$ while omitting the dependency on b_i's. We first show that the overparameterization of the NN, f_β parametrized by β, guarantees that it bahaves as its local linearization at the random initialization $\beta(0)$. To this end, we define

$$f_\beta^0(x) = \frac{1}{\sqrt{w}} \sum_{i=1}^{w} b_i \cdot \sigma'([\beta(0)]_i^\top x) \cdot [\beta]_i^\top x, \tag{D.9}$$

whereby the linear structure of $f_\beta^0(x)$ indicates

$$f_\beta^0(x) = \langle \nabla_\beta f_\beta^0(x), \beta \rangle. \tag{D.10}$$

We write $\beta(s)$ as the value of the parameter β at the s-th iteration of VFM (Algorithm 3). For the simplicity of notations, we denote by

$$G_{\beta(s)}(x) = -\nabla_\beta f_{\beta(s)}(x) + \nabla_\beta F^*(f_{\beta(s)}) \tag{D.11}$$

the stochastic gradient vector. Similarly, we also define

$$G_{\beta(s)}^0(x) = -\nabla_\beta f_{\beta(s)}^0(x) + \nabla_\beta F^*(f_{\beta(s)}^0). \tag{D.12}$$

Furthermore, let μ be the probability measure corresponding to input data distribution for the NNs. We denote by

$$\bar{G}_{\beta(s)} = \mathbb{E}_{\mu}\left[G_{\beta(s)}(x)\right] = \int_{\mathcal{X}} G_{\beta(s)}(x) d\mu(x) \tag{D.13}$$

the population mean of the stochastic gradient vector at the s-th iteration, and by $\bar{G}^0_{\beta(s)}$ its localization version. Without loss of generality, we assume $\|x\| \leq 1$ for the data.

We first present the following assumption for bounding the variance of the stochastic gradient vector $G_{\beta(s)}(x)$, which casts the analysis of global convergence into tracking the mean of the gradient vector. We denote by $\mathbb{E}_{\text{init}}[\cdot]$ the expectation over the random initialization for neural network parameter β, and $\mathbb{E}_{\mu}[\cdot]$ the expectation over the input $x \in \mathcal{X}$ conditioned on the random initialization.

Assumption D.4 (Variance of the Stochastic Update Vector)**.** For any $s \leq t$, there exists a constant $C_G^2 = \mathcal{O}(r_f^2)$ independent of s such that

$$\mathbb{E}_{\text{init}} \|G_{\beta(s)} - \bar{G}_{\beta(s)}\|^2_{L^2_{\mu}(\mathcal{X})} \leq C_G^2. \tag{D.14}$$

The following theorem provides the final characterization of the gradient error. Recall that in Algorithm 3, t is the number of iterations as well as the sample size denoted by N for either N_{μ} or N_{ν}, w is the width of neural networks, and r_f is the projection radius.

Furthermore, to formalize the smoothness of the activation function, we have the following assumption.

Assumption D.5 (Smooth Activation Function)**.** There exits an absolute constant $h > 0$ such that for any $x_1, x_2 \in \mathbb{R}^d$ we have

$$|\sigma'(x_1) - \sigma'(x_2)| \leq h\|x_1 - x_2\|. \tag{D.15}$$

For technical consideration, we also assume $\sigma(0) = 0$.

The following lemma quantifies the variance by the introduced function f^0_{β} and the generic function f_{β} in terms of r_f and w.

Lemma D.6. Under Assumption D.4 and D.5, given $\beta \in \mathcal{B}^0(r_f)$, we have

$$\mathbb{E}_{\text{init}} \|f_{\beta} - f^0_{\beta}\|^2_{L^2_{\mu}(\mathcal{X})} = \mathcal{O}(w^{-1} r_f^4). \tag{D.16}$$

Proof. By the definition of f_{β} and f^0_{β}, we have

$$\begin{aligned}
&\left|f_{\beta}(x) - f^0_{\beta}(x)\right| \\
&\leq \frac{1}{\sqrt{w}} \left| \sum_{i=1}^{w} \left(\sigma([\beta]_i^{\top} x) - \sigma'([\beta(0)]_i^{\top} x)[\beta]_i^{\top} x\right) \right| \\
&\leq \frac{1}{\sqrt{w}} \sum_{i=1}^{w} \left| \sigma([\beta]_i^{\top} x) - \sigma'([\beta(0)]_i^{\top} x)[\beta]_i^{\top} x \right| \\
&\leq \frac{1}{\sqrt{w}} \sum_{i=1}^{w} \left| \sigma([\beta(0)]_i^{\top} x) - \sigma'([\beta(0)]_i^{\top} x)[\beta(0)]_i^{\top} x \right| \\
&+ \frac{1}{\sqrt{w}} \sum_{i=1}^{w} \left| \left(\int_0^1 \left(\sigma'\left((1-\eta)[\beta(0)]_i^{\top} x + \eta [\beta]_i^{\top} x\right) - \sigma'\left([\beta(0)]_i^{\top} x\right) \right) \cdot d\eta \right) \cdot x^{\top} \left([\beta]_i - [\beta(0)]_i\right) \right|,
\end{aligned} \tag{D.17}$$

where the last inequality holds by the Taylor expansion, the second last inequality follows from the triangle inequality, and the second equality holds as $|b_i| = 1$. The expectation of the first term on the right-hand side

of (D.17) is bounded by

$$\frac{1}{\sqrt{w}}\mathbb{E}_{\text{init}}\sum_{i=1}^{w}\Big|\sigma([\beta(0)]_i^\top x)-\sigma'([\beta(0)]_i^\top x)[\beta(0)]_i^\top x\Big| = \mathcal{O}\bigg(\frac{1}{\sqrt{w}}\mathbb{E}_{\text{init}}\sum_{i=1}^{w}\Big|\sigma(0)+\big([\beta(0)]_i^\top x\big)^2\Big|\bigg)$$
$$\leq \mathcal{O}\bigg(\frac{1}{\sqrt{w}}\mathbb{E}_{\text{init}}\sum_{i=1}^{w}\|[\beta(0)]_i\|^2\bigg)$$
$$= \mathcal{O}\bigg(\frac{1}{\sqrt{w}}\bigg), \tag{D.18}$$

where we use the Taylor expansion of σ at 0, the fact that $\sigma(0)=0$, $\|x\|\leq 1$, and (D.7).

By squaring both sides of (D.17) and applying Assumption D.5 to the right-hand side of (D.17), we obtain

$$\mathbb{E}_{\text{init},\mu}\big|f_\beta(x)-f_\beta^0(x)\big|^2 = \mathcal{O}\bigg(\frac{1}{w}\bigg)+\mathcal{O}\bigg(\Big(\frac{h}{\sqrt{w}}\mathbb{E}_{\text{init}}\sum_{i=1}^{w}\|[\beta]_i-[\beta(0)]_i\|^2\Big)^2\bigg)$$
$$= \mathcal{O}\bigg(\Big(\frac{h}{\sqrt{w}}\sum_{i=1}^{w}\|[\beta]_i-[\beta(0)]_i\|^2\Big)^2\bigg)$$
$$= \mathcal{O}\bigg(\frac{h}{w}\|\beta-\beta(0)\|^4\bigg),$$
$$= \mathcal{O}\bigg(\frac{r_f^4}{w}\bigg). \tag{D.19}$$

Hence, it follows that

$$\mathbb{E}_{\text{init}}\|f_\beta-f_\beta^0\|_{L_\mu^2(\mathcal{X})}^2 = \mathcal{O}(w^{-1}r_f^4), \tag{D.20}$$

which concludes the proof. □

The following lemma characterizes the difference between the expected gradients of the the original neural network approximator and the locally linearized one.

Lemma D.7. For any $0\leq s\leq t$, we have the following linear approximation error at each iteration:

$$\mathbb{E}_{\text{init}}\|\bar{G}_{\beta(s)}-\bar{G}_{\beta(s)}^0\|_2^2 = \mathcal{O}(w^{-1}r_f^4+w^{-1}r_f^2). \tag{D.21}$$

Proof. By the definition of $\bar{G}_{\beta(s)}$ and $\bar{G}_{\beta(s)}^0$, we have

$$\begin{aligned}
&\mathbb{E}_{\text{init}}\|\bar{G}_{\beta(s)}-\bar{G}_{\beta(s)}^0\|_2^2\\
&\leq \mathbb{E}_{\text{init},\mu}\|-\nabla_\beta f_{\beta(s)}(x)+\nabla_\beta F^*(f_{\beta(s)})+\nabla_\beta f_{\beta(s)}^0(x)-\nabla_\beta F^*(f_{\beta(s)}^0)|_2^2\\
&\leq \underbrace{2\,\mathbb{E}_{\text{init},\mu}\|\nabla_\beta f_{\beta(s)}(x)-\nabla_\beta f_{\beta(s)}^0(x)\|_2^2}_{\text{(i)}}+\underbrace{2\,\mathbb{E}_{\text{init},\mu}\|\nabla_\beta F^*(f_{\beta(s)})-\nabla_\beta F^*(f_{\beta(s)}^0)\|_2^2}_{\text{(ii)}}.
\end{aligned} \tag{D.22}$$

In what follows, we upper bound term (i) and (ii) respectively.

Upper Bounding (i): Recall that

$$\nabla_\beta f_\beta(x) = 1/\sqrt{w}\cdot\big(b_1\cdot\sigma'([\beta]_1^\top x)\cdot x^\top,\ldots,b_w\cdot\sigma'([\beta]_w^\top x)\cdot x^\top\big)^\top,$$

and

$$\nabla_\beta f_\beta^0(x) = 1/\sqrt{w}\cdot\big(b_1\cdot\sigma'([\beta(0)]_1^\top x)\cdot x^\top,\ldots,b_w\cdot\sigma'([\beta(0)]_w^\top x)\cdot x^\top\big)^\top.$$

We have

$$\|\nabla_\beta f_{\beta(s)}(x) - \nabla_\beta f^0_{\beta(s)}(x)\|_2^2 = \frac{1}{w}\sum_{i=1}^{w}\big(\sigma'([\beta]_i^\top x) - \sigma'([\beta(0)]_i^\top x)\big)^2 \cdot \|x\|^2$$
$$\leq \frac{h^2}{w}\sum_{i=1}^{w}\|[\beta]_i - [\beta(0)]_i\|^2$$
$$= \frac{h^2 r_f^2}{w}. \tag{D.23}$$

Here the last inequality follows from the condition $\|x\|^2 \leq 1$, which is also used to derive (D.17). By taking expectation on (D.23) we obtain

$$\mathbb{E}_{\text{init},\mu}\|\nabla_\beta f_{\beta(s)}(x) - \nabla_\beta f^0_{\beta(s)}(x)\|_2^2 = \mathcal{O}(w^{-1}r_f^2). \tag{D.24}$$

Therefore, we have bounded term (i) in (D.22).

Upper Bounding (ii): By the L-smoothness of the dual functional F^*, it follows that

$$\mathbb{E}_{\text{init},\mu}\|\nabla_\beta F^*(f_{\beta(s)}) - \nabla_\beta F^*(f^0_{\beta(s)})\|_2^2$$
$$\leq \mathbb{E}_{\text{init},\mu}\|f_\beta(x) - f^0_\beta(x)\|_2^2, \tag{D.25}$$

where the right-hand side of (D.25) is exactly the left-hand side of (D.16). Hence by Lemma D.6, the term (ii) is bounded by $\mathcal{O}(w^{-1}r_f^4)$. Combining (i) and (ii) concludes the proof for Lemma D.7. □

In what follows, with the explicit linearized learning target $f^0_{\beta^*}$, we are able to characterize the global convergence of Algorithm 3 by the difference between output estimated function and $f^0_{\beta^*}$.

Lemma D.8. Suppose that $t > 64$ iterations of Algorithm 1 are run, and the stepsize is set to be $\eta = t^{-1/2}$. Then, under Assumption D.4 and D.5 we have

$$\mathbb{E}_{\text{init}}\|f_{\widehat{\beta}} - f^0_{\beta^*}\|^2_{L^2_\mu(\mathcal{X})} = \mathcal{O}(r_f^2 t^{-1/2} + w^{-1/2}r_f^3 + w^{-1}r_f^4), \tag{D.26}$$

where β^* is the approximate stationary point such that

$$\beta^* = \Pi_{\mathcal{B}^0(r_f)}(\beta^* - \eta\bar{G}^0_{\beta^*}), \tag{D.27}$$

and $\widehat{\beta} = 1/t \cdot \sum_{s=0}^{t-1}\beta(s)$.

Proof. First we bound the progress of the one-step update. By the convexity of $\mathcal{B}^0(r_f)$ and the approximate stationary condition (D.27), for each $s < t$ we have

$$\mathbb{E}_\mu[\|\beta(s+1) - \beta^*\|_2^2|\beta(s)]$$
$$= \mathbb{E}_\mu\left[\left\|\Pi_{\mathcal{B}^0(r_f)}\left(\beta(s) - \eta G_{\beta(s)}\right) - \Pi_{\mathcal{B}^0(r_f)}\left(\beta^* - \eta\bar{G}^0_{\beta^*}\right)\right\|_2^2|\beta(s)\right]$$
$$\leq \mathbb{E}_\mu\left[\left\|(\beta(s) - \beta^*) - \eta\left(G_{\beta(s)} - \bar{G}^0_{\beta^*}\right)\right\|_2^2|\beta(s)\right]$$
$$= \|\beta(s) - \beta^*\|_2^2 - 2\eta\langle\beta(s) - \beta^*, \bar{G}_{\beta(s)} - \bar{G}^0_{\beta^*}\rangle + \eta^2\|G_{\beta(s)} - \bar{G}^0_{\beta^*}\|^2_{L^2_\mu(\mathcal{X})}. \tag{D.28}$$

Then our target is reduced to upper bound the last two terms above. For the inner product term, by applying Hölder's inequality we have

$$\langle\beta(s) - \beta^*, \bar{G}_{\beta(s)} - \bar{G}^0_{\beta^*}\rangle = \langle\beta(s) - \beta^*, \bar{G}_{\beta(s)} - \bar{G}^0_{\beta(s)}\rangle + \langle\beta(s) - \beta^*, \bar{G}^s_{\beta(s)} - \bar{G}^0_{\beta^*}\rangle$$
$$\geq -\|\bar{G}_{\beta(s)} - \bar{G}^0_{\beta(s)}\|_2 \cdot \|\beta(s) - \beta^*\|_2 + \langle\beta(s) - \beta^*, \bar{G}^0_{\beta(s)} - \bar{G}^0_{\beta^*}\rangle$$
$$\overset{(a)}{\geq} -r_f\|\bar{G}_{\beta(s)} - \bar{G}^0_{\beta(s)}\|_2 + \langle\beta(s) - \beta^*, \bar{G}^0_{\beta(s)} - \bar{G}^0_{\beta^*}\rangle, \tag{D.29}$$

where (a) uses the fact that $\|\beta(s)-\beta^*\|_2 \le r_f$. For the rest inner product term in (D.29), by plugging in the definition of stochastic gradient vectors we obtain

$$\begin{aligned}\langle \beta(s)-\beta^*, \bar{G}^0_{\beta(s)} - \bar{G}^0_{\beta^*}\rangle &= \mathbb{E}_\mu \left\langle \beta(s)-\beta^*, \nabla_\beta f^0_{\beta^*}(x) - \nabla_\beta f^0_{\beta(s)}(x)\right\rangle \\ &\quad + \mathbb{E}_\mu \left\langle \beta(s)-\beta^*, \nabla_\beta F^*\left(f^0_{\beta(s)}\right) - \nabla_\beta F^*\left(f^0_{\beta^*}\right)\right\rangle \\ &= \mathbb{E}_\mu \left\langle \beta(s)-\beta^*, \nabla_\beta F^*\left(f^0_{\beta(s)}\right) - \nabla_\beta F^*\left(f^0_{\beta^*}\right)\right\rangle, \end{aligned} \tag{D.30}$$

where we use the fact that $\nabla_\beta f^0_\beta(x)$ is independent of the value of β. We proceed to derive the lower bound by the explicit functional gradient form,

$$\begin{aligned}&\mathbb{E}_\mu \left\langle \nabla_\beta F\left(f^0_{\beta(s)}\right) - \nabla_\beta F\left(f^0_{\beta^*}\right), \beta(s)-\beta^*\right\rangle \\ &= \mathbb{E}_\mu \left\langle \left(\frac{\delta F^*}{\delta f}\left(f^0_{\beta(s)}\right) - \frac{\delta F^*}{\delta f}\left(f^0_{\beta^*}\right)\right)\cdot \nabla_\beta f^0_\beta,\ \beta(s)-\beta^*\right\rangle \\ &= \mathbb{E}_\mu \left\langle \frac{\delta F^*}{\delta f}\left(f^0_{\beta(s)}\right) - \frac{\delta F^*}{\delta f}\left(f^0_{\beta^*}\right), f^0_{\beta(s)} - f^0_{\beta^*}\right\rangle \\ &\overset{(b)}{\ge} \gamma \|f^0_{\beta(s)} - f^0_{\beta^*}\|^2_{L^2_\mu(\mathcal{X})}, \end{aligned} \tag{D.31}$$

where (b) is due to the γ-strong convexity of functional F^*.

On the other hand, the third norm term in (D.28) is estimated using the Cauchy-Schwarz inequality as follows,

$$\begin{aligned}\|G_{\beta(s)} - \bar{G}^0_{\beta^*}\|^2_{L^2_\mu(\mathcal{X})} &\le 2\|G_{\beta(s)} - \bar{G}_{\beta(s)}\|^2_{L^2_\mu(\mathcal{X})} + 2\|\bar{G}_{\beta(s)} - \bar{G}^0_{\beta^*}\|^2_{L^2_\mu(\mathcal{X})} \\ &\le 2\mathbb{E}_{\text{init}}\|G_{\beta(s)} - \bar{G}_{\beta(s)}\|^2_{L^2_\mu(\mathcal{X})} + 4\|\bar{G}_{\beta(s)} - \bar{G}^0_{\beta(s)}\|^2_2 + 4\|\bar{G}^0_{\beta(s)} - \bar{G}^0_{\beta^*}\|^2_2 \end{aligned} \tag{D.32}$$

where the first term herein is upper bounded by the variance of $G_{\beta(s)}$ in (D.14) of Assumption D.4, and the second one is controlled by lemma D.6. Thus it suffices to upper bound the last term, *i.e.*, the squared difference between the expected gradient of time-step s and the optimal one of linearization approximation. We have

$$\begin{aligned}\|\bar{G}^0_{\beta(s)} - \bar{G}^0_{\beta^*}\|^2_2 &= \|\nabla_\beta \mathbb{E} f^0_{\beta^*} - \nabla_\beta \mathbb{E} f^0_{\beta(s)} + \nabla_\beta F^*(f^0_{\beta(s)}) - \nabla_\beta F^*(f^0_{\beta^*})\|^2_2 \\ &\le 2\|\nabla_\beta \mathbb{E} f^0_{\beta^*} - \nabla_\beta \mathbb{E} f^0_{\beta(s)}\|^2_2 + 2\|\nabla_\beta F^*(f^0_{\beta(s)}) - \nabla_\beta F^*(f^0_{\beta^*})\|^2_2 \\ &\le 2\|f^0_{\beta^*} - f^0_{\beta(s)}\|^2_{L^2_\mu(\mathcal{X})} + 2L\|f^0_{\beta^*} - f^0_{\beta(s)}\|^2_{L^2_\mu(\mathcal{X})} \qquad \text{(D.33)} \\ &= 2(1+L)\|f^0_{\beta^*} - f^0_{\beta(s)}\|^2_{L^2_\mu(\mathcal{X})}, \qquad \text{(D.34)} \end{aligned}$$

where (D.33) follows from the L-smoothness of the entropy dual functional by Example C.2. To conclude, we have

$$\begin{aligned}&\mathbb{E}_\mu[\|\beta(s+1)-\beta^*\|^2_2 \mid \beta(s)] \\ &\le \|\beta(s)-\beta^*\|^2_2 + 2\eta r_f \|\bar{G}_{\beta(s)} - \bar{G}^0_{\beta(s)}\|^2_2 \\ &+ 2\eta^2 \mathbb{E}_{\text{init}}\|G_{\beta(s)} - \bar{G}_{\beta(s)}\|^2_{L^2_\mu(\mathcal{X})} + 4\eta^2\|\bar{G}_{\beta(s)} - \bar{G}^0_{\beta(s)}\|^2_2 \\ &+ 2\eta(4\eta(1+L)-\gamma)\|f^0_{\beta(s)} - f^0_{\beta^*}\|^2_{L^2_\mu(\mathcal{X})}. \end{aligned} \tag{D.35}$$

By rearranging (D.35), we have

$$\begin{aligned}\|f_{\beta(s)} - f^0_{\beta^*}\|^2_{L^2_\mu(\mathcal{X})} &\le 2\|f_{\beta(s)} - f^0_{\beta(s)}\|^2_{L^2_\mu(\mathcal{X})} + 2\|f^0_{\beta(s)} - f^0_{\beta^*}\|^2_{L^2_\mu(\mathcal{X})} \\ &\le \big(\gamma\eta - 4\eta^2(L+1)\big)^{-1}\Big(\|\beta(s) - \beta^*\|_2^2 - \mathbb{E}_\mu[\|\beta(s+1) - \beta^*\|_2^2 \mid \beta(s)] \\ &\quad + 2C_G^2\eta^2 + \mathcal{O}\Big(w^{-1/2}r_f^3 + w^{-1}r_f^4\Big)\Big), \end{aligned} \tag{D.36}$$

where the second inequality follows from lemma D.6 and D.7, as well as the fact $\eta < \gamma/8(L+1)$ resulted from $t > 64(L+1)^2/\gamma^2$ and $\eta = t^{-1/2}$. We proceed to take total expectation on both sides of (D.36) and telescoping for $s+1 \in [t]$ $(t \ge 1)$ to obtain

$$\begin{aligned}\mathbb{E}_{\text{init}}\|f_{\widehat{\beta}} - f^0_{\beta^*}\|^2_{L^2_\mu(\mathcal{X})} &= \mathbb{E}_{\text{init},\mu}\big[\big(f_{\widehat{\beta}}(x) - f^0_{\beta^*}(x)\big)^2\big] \\ &\le \frac{1}{t}\sum_{s=0}^{t-1}\mathbb{E}_{\text{init},\mu}\big[\big(f_{\beta(s)}(x) - f^0_{\beta^*}(x)\big)^2\big] \\ &\le t^{-1}\cdot\big(\gamma\eta - 4\eta^2(L+1)\big)^{-1}\cdot\Big(\mathbb{E}_{\text{init}}\big[\big(\|\beta(0) - \beta^*\|_2^2\big] + 2tC_G^2\eta^2 \\ &\quad + \mathcal{O}\Big(w^{-1/2}r_f^3 + w^{-1}r_f^4\Big)\Big). \end{aligned} \tag{D.37}$$

By plugging the conditions on t and η, we have

$$\begin{aligned} t^{-1}\cdot\big(\gamma\eta - 4\eta^2(L+1)\big)^{-1} &= t^{-1/2}\cdot\frac{1}{\gamma - 4\eta(L+1)} \\ &\overset{\eta<\gamma/8(L+1)}{\le} t^{-1/2}\cdot\frac{1}{\gamma - \gamma/2} \\ &= \frac{2}{\gamma\sqrt{t}}, \end{aligned} \tag{D.38}$$

Then, we obtain the following bound,

$$\begin{aligned}&\mathbb{E}_{\text{init}}\|f_{\widehat{\beta}} - f^0_{\beta^*}\|^2_{L^2_\mu(\mathcal{X})} \\ &\le \frac{2}{\gamma\sqrt{t}}\big(\mathbb{E}_{\text{init}}[\|\beta(0) - \beta^*\|_2^2] + 2C_G^2\big) + \mathcal{O}(r_f^{5/2}w^{-1/4} + r_f^3 w^{-1/2}) \qquad \text{(D.39)} \\ &\le \frac{2(r_f^2 + 2C_G^2)}{\gamma\sqrt{t}} + \mathcal{O}(r_f^{5/2}w^{-1/4} + r_f^3 w^{-1/2}) \\ &= \mathcal{O}\big(r_f^2 t^{-1/2} + w^{-1/2}r_f^3 + w^{-1}r_f^4\big). \qquad \text{(D.40)}\end{aligned}$$

Hence we complete the proof. □

Now it is straightforward to prove the main theorem for VFM with neural network parametrized function class as follows.

D.3 Gradient Error of Neural Networks

Theorem D.9. Under Assumption 4.1, D.4 and D.5, within the k-th iteration of Algorithm 2, the gradient error $\bar{\varepsilon}_k$ defined in (4.1) satisfies

$$\bar{\varepsilon}_k = \bar{\varepsilon}_k(N) = \mathcal{O}\big(\bar{\varepsilon}(N)\big) = \mathcal{O}\bigg(\widetilde{K}\bigg(\frac{r_f^2}{N^{1/2}} + \frac{r_f^3}{w^{1/2}} + \frac{r_f^4}{w}\bigg)\bigg). \tag{D.41}$$

Here $\bar{\varepsilon}(N)$ is defined as

$$\bar{\varepsilon}(N) = \widetilde{K}\bigg(\frac{r_f^2}{N^{1/2}} + \frac{r_f^3}{w^{1/2}} + \frac{r_f^4}{w}\bigg). \tag{D.42}$$

We remark that the order of the gradient error is independent of iteration k and can be decomposed into the generalization error of $\mathcal{O}(r_f^2/N^{1/2})$ for SGD over the neural tangent kernel and the error of $\mathcal{O}(r_f^3/w^{1/2})$, $\mathcal{O}(r_f^4/w)$ for approximating the neural network by a linear function. In particular, the overall gradient errors decay to zero at the rate of $1/\sqrt{N}$ with a sufficiently large width w of the neural network.

Proof. By Jensen's inequality with respect to squared L_μ^2-norm, we have

$$\begin{aligned}
&\mathbb{E}_{\text{init}}\|f_{\widehat{\beta}} - f_{\beta^*}\|_{L_\mu^2(\mathcal{X})}^2 \\
&\quad \le 2\mathbb{E}_{\text{init}}\|f_{\widehat{\beta}} - f_{\beta^*}^0\|_{L_\mu^2(\mathcal{X})}^2 + 2\mathbb{E}_{\text{init}}\|f_{\beta^*}^0 - f_{\beta^*}\|_{L_\mu^2(\mathcal{X})}^2.
\end{aligned} \tag{D.43}$$

On one hand, by Lemma D.8 the first term above can be bounded with $\mathcal{O}(r_f^2 t^{-1/2} + w^{-1/2} r_f^3 + w^{-1} r_f^4)$. On the other hand, the second term in (D.43) is bounded by $\mathcal{O}(r_f^4 w^{-1})$ through setting $\beta = \beta^*$ in Lemma D.6. To conclude, the total bound for (D.43) turns out to be $\mathcal{O}(r_f^2 t^{-1/2} + w^{-1/2} r_f^3 + w^{-1} r_f^4)$, which completes the proof for the first part of the theorem.

To verify (D.41), we invoke Lemma 4.2 by setting f to $\widetilde{f}_k^* - \widehat{f}_k^*$ with all possible iteration k for both players and p to 2, combining with (4.1) to obtain for any $k \in \mathbb{N}$, we have

$$\begin{aligned}
\bar{\varepsilon}_k &= \mathbb{E}_{X_0}\int_{\mathcal{X}} \big\|\nabla \widetilde{f}_k^*(x) - \nabla \widehat{f}_k^*(x)\big\|_2^2 \,\mathrm{d}\widetilde{\rho}_k \\
&\overset{\text{(D.2)}}{\le} \widetilde{K}\mathbb{E}_{X_0}\int_{\mathcal{X}} \big|\widetilde{f}_k^*(x) - \widehat{f}_k^*(x)\big|_2^2 \,\mathrm{d}\widetilde{\rho}_k \\
&= \widetilde{K}\mathbb{E}_{X_0}\mathbb{E}_{\text{init}}\|f_{\widehat{\beta}} - f_{\beta^*}\|_{L_{\widetilde{\rho}_k}^2(\mathcal{X})}^2 \\
&= \mathcal{O}\bigg(\widetilde{K}\bigg(\frac{r_f^2}{N^{1/2}} + \frac{r_f^3}{w^{1/2}} + \frac{r_f^4}{w}\bigg)\bigg),
\end{aligned} \tag{D.44}$$

where the last equality follows from (D.43) and the measure $\widetilde{\rho}_k$ accounts for either distribution iterates at some iteration k, $\widetilde{\mu}_k$ or $\widetilde{\nu}_k$ in the paper. Therefore, we conclude the whole proof of the theorem for the statistical errors of VFM steps. □

E Proof of Convergence for Distributional Game Optimization

In this section, we layout the complete proof of the convergence rate of the particle-based infinite-dimensional game optimization, for which some intermediate lemmas are also listed below. We define $\nu^*(\mu) \in \mathcal{I}_F(\mu) \triangleq \operatorname{argmax}_\nu F(\mu, \nu)$ for simplicity. Also note that according to the definition in Algorithm 1, we have $\nu_{k+1} = \nu_{K_\nu}(\mu_k)$ as the last iterate of the inner loop for the k-th outer loop. Moreover, in the formal description of VTIG in Algorithm 1, we adopt sets of discrete particles to represent the underlying distribution iterates being updated, where $X_l^\nu(\widetilde{\mu}_k)$ denotes the set of particles for player ν at the l-th iteration given $\widetilde{\mu}_k$ as the current distribution iterate for player μ. Recall that we use notations $\widetilde{\mu}_k$, $\widetilde{\nu}_l(\widetilde{\mu}_k)$ to indicate the distribution iterate derived from constructed transportation maps $\{T_k^\mu\}_{k\in[K_\mu]}$ and $\{T_{kl}^\nu\}_{k\in[K_\mu],l\in[K_\nu]}$ defined in (3.9).

For simplicity we define $H_\mu(\nu) \triangleq -F(\mu, \nu)$. The following proposition identifies that $\mathcal{M}(\mathcal{X})$ is well defined.

Proposition E.1. The manifold $\mathcal{M}(\mathcal{X})$ is compact, that is, there exists $R > 0$ such that for any $\mu_1, \mu_2 \in \mathcal{M}(\mathcal{X})$ we have $W_2(\mu_1, \mu_2) \le 2R$.

Proof. Since $\mathcal{X}$ is defined as a compact domain, according to the topological properties in Theorem 6.16 of [63], the manifold $\mathcal{M}(\mathcal{X})$ is also compact. Also, as the underlying variable space $\mathcal{X}$ is bounded, the Wasserstein distance defined over $\mathcal{M}(\mathcal{X})$ over $\mathcal{X}$ is bounded, by setting $2R = \sup_{\mu_1,\mu_2\in\mathcal{M}(\mathcal{X}_\mu)} \mathcal{W}_2(\mu_1,\mu_2)$ we arrive at the conclusion. □

Before proceeding, we first state the following additional definitions and lemmas.

Definition E.2. We say that a function $G : \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ satisfies the Quadratic Grouth (QG) condition *w.r.t.* $\mathcal{W}_2$ metric with constant $\gamma > 0$ if

$$G(\mu) - G(\mu^*) \geq \frac{\gamma}{2} W_2^2(\mu,\mu^*), \quad \forall \mu \in \mathcal{M}(\mathcal{X}), \tag{E.1}$$

where μ^* is the minimizer of the function. Note that if function G is PL with constant ξ, then G satisfies the QG condition with constant $\gamma = 4\xi$ [33].

The next lemma demonstrates the stability of $\nu^*(\mu)$ controlled by the variation of μ.

Lemma E.3. Given that $\mathcal{I}_F(\mu)$ is closed, then for any $\mu_1, \mu_2 \in \mathcal{M}$ and $\nu_1 \in \mathcal{I}_F(\mu_1)$, there exists a $\nu_2 \in \mathcal{I}_F(\mu_2)$ such that

$$\mathcal{W}_2(\nu_1,\nu_2) \leq \frac{L_0}{2\xi}\mathcal{W}_2(\mu_1,\mu_2). \tag{E.2}$$

Proof. By Assumption 2.2 and the PL condition of F, we have

$$\begin{aligned} 2\xi\big(G(\mu_2) - F(\mu_2,\nu_1)\big) &\leq \|\operatorname{grad} F_{\mu_2}(\nu_1)\|_{\nu_1}^2 \\ &= \boldsymbol{d}^2\big(\operatorname{grad} F_{\mu_2}(\nu_1), \operatorname{grad} F_{\mu_1}(\nu_1)\big) \\ &\leq L_0^2 \mathcal{W}_2^2(\mu_1,\mu_2). \end{aligned} \tag{E.3}$$

Here the equality holds by $\operatorname{grad} F_{\mu_1}(\nu_1) = 0$ a.e. ν_1. Moreover, by the QG property of $H_\mu(\nu)$ it follows that there exists a $\nu_2 \in \mathcal{I}_F(\mu_2)$ such that

$$\mathcal{W}_2^2(\nu_1,\nu_2) \leq \frac{1}{2\xi}\cdot\big(G(\mu_2) - F(\mu_2,\nu_1)\big) \leq \frac{L_0^2}{4\xi^2}\mathcal{W}_2^2(\mu_1,\mu_2), \tag{E.4}$$

where we use (E.3) for the second inequality. This concludes the result. □

At the two-player game optimization scale (VFM not involved), considering the interaction of the two players, our algorithm runs multiple gradient ascent steps in the inner loop to estimate the *inner maximization value functional* defined as $G(\mu) \triangleq \max_{\nu\in\mathcal{M}(\mathcal{X}_\nu)} F(\mu,\nu)$, of which the Riemannian gradient *w.r.t.* μ at the optimum $\nu^*(\mu)$ is adopted to estimate the Riemannian gradient of $G(\mu)$. Inspired by this, we rewrite (2.3) as $\min_{\mu\in\mathcal{M}(\mathcal{X}_\mu)} G(\mu)$ and observe that the overall VTIG behaves as a gradient descent-like algorithm over the inner maximization value functional G. Hence we can make use of nonconvex optimization techniques to analyze the overall convergence properties of such zero-sum two-player games.

It is not straightforward to see $\operatorname{grad} G(\mu) = \operatorname{grad}_\mu F(\mu,\nu^*(\mu))$ as there may be multiple $\nu^*(\mu)$'s under Assumption 2.3, breaking the condition of Danskin's theorem [4] even for finite cases. Fortunately, we show in the following with Riemannian PL condition, we can still prove a Danskin-type result. In a nutshell, our proposed algorithm is a gradient descent-like algorithm on the inner maximization value functional. To proceed with two-player games of a higher hierarchy for Problem (2.3), we follow the assumptions of smoothness in different aspect of F and provide a Danskin-type lemma in the following section.

E.1 Danskin-type Lemma

Lemma E.4. Under Assumption 2.2 and 2.3, we have

$$\operatorname{grad} G(\mu) = \operatorname{grad}_\mu F\big(\mu, \nu^*(\mu)\big) \text{ a.e. } \mu, \quad \forall\, \nu^*(\mu) \in \operatorname*{argmin}_{\nu \in \mathcal{M}(\mathcal{X}_\nu)} H_\mu(\nu).$$

In addition, G is $\widetilde{L}_1$-Lipschitz and $\widetilde{L}$-smooth with $\widetilde{L}_1 = L_\mu + L_\nu L_0/(2\xi)$ and $\widetilde{L} = L_1 + L_0^2/(2\xi)$.

Proof. By the definition of the directional derivative and gradient in Section B.2, we take the curve γ, which starts with $\gamma(0) = \mu \in \mathcal{M}(\mathcal{X})$ and $\gamma'(0) = u \in T_\mu \mathcal{M}(\mathcal{X})$, as the geodesic. Then by Taylor series expansion at $\gamma(0)$, we have for any scalar τ

$$\begin{aligned} G[\gamma(\tau)] - G(\mu) &= F(\gamma(\tau), \nu^*(\gamma(\tau))) - F(\mu, \nu^*(\mu)) \\ &= \big\langle \operatorname{grad}_\mu F(\mu, \nu^*(\mu)), \tau u \big\rangle_\mu + \Big\langle \operatorname{grad}_\nu F(\mu, \nu^*(\mu)), \operatorname{Exp}^{-1}_{\nu^*(\mu)}\big(\nu^*\big(\gamma(\tau)\big)\big) \Big\rangle_{\nu^*(\mu)} + \mathcal{O}(\tau^2) \\ &= \big\langle \operatorname{grad}_\mu F(\mu, \nu^*(\mu)), \tau u \big\rangle_\mu + \mathcal{O}(\tau^2). \end{aligned} \tag{E.5}$$

Here the last equality follows from the optimal condition on ν: $\operatorname{grad}_\nu F(\mu, \nu^*(\mu)) = 0$, and the additional higher order term in τ integrates terms in $\mathcal{W}_2(\nu^*(\gamma(\tau)), \nu^*(\mu))$ since by Lemma E.3 there exists $\nu^*(\gamma(\tau)) \in \mathcal{I}_F(\gamma(\tau))$ such that

$$\mathcal{W}_2(\nu^*(\gamma(\tau)), \nu^*(\mu)) \le \frac{L_0}{2\xi} \tau \|u\|. \tag{E.6}$$

In the meanwhile, by the limit form of directional derivative we have

$$\begin{aligned} \langle \operatorname{grad} G(\mu), u \rangle_\mu &= \nabla_u G(\mu) \\ &= \frac{d}{d\tau} G\left[\gamma(\tau)\right] \big|_{t=0} \\ &= \lim_{\tau \to 0^+} \frac{G[\gamma(\tau)] - G[\gamma(0)]}{\tau} \\ &= \big\langle \operatorname{grad}_\mu F(\mu, \nu^*(\mu)), u \big\rangle_\mu . \end{aligned} \tag{E.7}$$

Due to the arbitrarity of u, we conclude that $\operatorname{grad} G(\mu) = \operatorname{grad}_\mu F(\mu, \nu^*(\mu))$ a.e. μ.

In what follows we show the Lipschitzness and smoothness of G. For $\mu_1, \mu_2 \in \mathcal{M}(\mathcal{X})$, let $\nu^*(\mu_1) \in \mathcal{I}_F(\mu_1)$ and $\nu^*(\mu_2) \in \operatorname{argmin}_{\nu \in \mathcal{I}_F(\mu_2)} \mathcal{W}_2^2(\nu, \nu^*(\mu_1))$, then

$$\begin{aligned} &|G(\mu_1) - G(\mu_2)| \\ &= |F(\mu_1, \nu^*(\mu_1)) - F(\mu_2, \nu^*(\mu_1)) + F(\mu_2, \nu^*(\mu_1)) - F(\mu_2, \nu^*(\mu_2))| \\ &\le L_\mu \mathcal{W}_2(\mu_1, \mu_2) + L_\nu \mathcal{W}_2(\nu^*(\mu_1), \nu^*(\mu_2)) \\ &\le \left(L_\mu + \frac{L_\nu L_0}{2\xi}\right) \mathcal{W}_2(\mu_1, \mu_2), \end{aligned} \tag{E.8}$$

where the last inequality holds by Lemma E.3. Hence by setting $\widetilde{L}_1 = L_\mu + L_\nu L_0/(2\xi)$ we obtain that function G is $\widetilde{L}_1$-Lipschitz. In addition, for the difference between gradients we have

$$\begin{aligned} \boldsymbol{d}\left(\operatorname{grad} G(\mu_1), \operatorname{grad} G(\mu_2)\right) &= \boldsymbol{d}\left(\operatorname{grad}_\mu F(\mu_1, \nu^*(\mu_1)), \operatorname{grad}_\mu F(\mu_2, \nu^*(\mu_2))\right) \\ &\le \boldsymbol{d}\left(\operatorname{grad}_\mu F(\mu_1, \nu^*(\mu_1)), \operatorname{grad}_\mu F(\mu_2, \nu^*(\mu_1))\right) \\ &+ \boldsymbol{d}\left(\operatorname{grad}_\mu F(\mu_2, \nu^*(\mu_1)), \operatorname{grad}_\mu F(\mu_2, \nu^*(\mu_2))\right) \\ &\le \widetilde{L}_1 \mathcal{W}_2(\mu_1, \mu_2) + L_0 \mathcal{W}_2(\nu^*(\mu_1), \nu^*(\mu_2)) \\ &\le \left(\widetilde{L}_1 + \frac{L_0^2}{2\xi}\right) \mathcal{W}_2(\mu_1, \mu_2), \end{aligned} \tag{E.9}$$

where the second inequality holds by Assumption 2.2 and the last inequality still follows from Lemma E.3.

Therefore, by setting $\widetilde{L} = \widetilde{L}_1 + L_0^2/(2\xi)$ we conclude the proof. $\square$

E.2 Convergence Results for Player ν

Considering the nested loops in Algorithm 2, we first present the following linear convergence rate with statistical errors for player ν. Recall that $\widetilde{\mathcal{F}}$ defined in (3.5) is the function class over which we solve (2.4). We define $\nu^*(\mu) = \operatorname{argmax}_{\nu\in\mathcal{M}} F(\mu,\nu)$ given $\mu \in \mathcal{M}$. Also, we write $\widetilde{\nu}_l(\widetilde{\mu}_k)$ as $\widetilde{\nu}_l$ given any fixed $k \geq 0$ for notational simplicity.

Lemma E.5. Let F admit the variational form under Assumptions 2.2 and 2.3. Suppose that $\widetilde{\mathcal{F}}$ satisfies Assumption 4.1. Also, we set the stepsize $\eta^\nu \in \big(0, \min\{1/(4L_\nu), 1/h\}\big)$. Then, for any $l \geq 1$ and a fixed $\mu \in \mathcal{M}$, we have

$$F\big(\mu,\nu^*(\mu)\big) - \mathbb{E}\big[F(\mu,\widetilde{\nu}_l)\big] \leq \sigma^l \cdot \big[F(\mu,\nu^*(\mu)) - F(\mu,\widetilde{\nu}_0)\big] + \eta^\nu \sigma^l \sum_{m=0}^{l-1} \bar{\varepsilon}_m \sigma^{-(m+1)}, \tag{E.10}$$

where $\sigma = 1 - \xi\eta^\nu/2 \in (0,1)$ is the contraction coefficient, the expectation is taken *w.r.t.* the initial samples $X_0^\nu = \{x_{i,0}^\nu\}_{i\in[N_\nu]}$, and $\bar{\varepsilon}_m$ ($m \in [l-1]$) is the (expected) gradient error at timestep m defined in (4.1).

Proof. Our proof is based on quantifying some contraction between errors of adjacent iterates. Recall that we denote $H_\mu(\nu) = -F(\mu,\nu)$. For fixed $\mu \in \mathcal{M}(\mathcal{X})$, we write $H_\mu(\nu)$ as $H(\nu)$ below. To begin with, by Proposition 3.2, we can equivalently write the iteration of VTIG for $\widetilde{\nu}_l$ as

$$\widetilde{\nu}_{l+1} = \operatorname{Exp}_{\widetilde{\nu}_l}\big\{-\eta^\nu \cdot [\operatorname{grad} H(\widetilde{\nu}_l) + \delta_l]\big\}, \qquad \delta_l = -\operatorname{div}\big[\widetilde{\nu}_l \cdot (\nabla \widetilde{f}_l^* - \nabla f_l^*)\big]. \tag{E.11}$$

Notice that $\delta_l \in T_{\widetilde{\nu}_l}\mathcal{M}$ is a tangent vector at point $\widetilde{\nu}_l$. Moreover, by Assumption 2.2 functional H is L_2-smooth. Combining this property with (E.11), we have

$$\begin{aligned}
H(\widetilde{\nu}_{l+1}) &\leq H(\widetilde{\nu}_l) - \eta^\nu \cdot \big\langle \operatorname{grad} H(\widetilde{\nu}_l), \operatorname{grad} H(\widetilde{\nu}_l)\big\rangle_{\widetilde{\nu}_l} - \eta^\nu \big\langle \operatorname{grad} H(\widetilde{\nu}_l), \delta_l\big\rangle_{\widetilde{\nu}_l} \\
&\quad + \frac{(\eta^\nu)^2 L_2}{2} \big\langle \operatorname{grad} H(\widetilde{\nu}_l) + \delta_l, \operatorname{grad} H(\widetilde{\nu}_l) + \delta_l\big\rangle_{\widetilde{\nu}_l} \\
&= H(\widetilde{\nu}_l) - \left(\eta^\nu - \frac{(\eta^\nu)^2 L_2}{2}\right) \cdot \big\langle \operatorname{grad} H(\widetilde{\nu}_l), \operatorname{grad} H(\widetilde{\nu}_l)\big\rangle_{\widetilde{\nu}_l} \\
&\quad + \big(\eta^\nu + (\eta^\nu)^2 \cdot L_2\big) \cdot \big|\big\langle \operatorname{grad} H(\widetilde{\nu}_l), \delta_l\big\rangle_{\widetilde{\nu}_l}\big| + \frac{(\eta^\nu)^2 L_2}{2} \big\langle \delta_l, \delta_l\big\rangle_{\widetilde{\nu}_l},
\end{aligned} \tag{E.12}$$

where $\langle\cdot,\cdot\rangle_{\widetilde{\nu}_l}$ is the Riemannain metric of $\mathcal{M}$ at $\widetilde{\nu}_l$. By basic inequality $2ab \leq a^2 + b^2$, we have

$$\big|\big\langle \operatorname{grad} H(\widetilde{\nu}_l), \delta_l\big\rangle_{\widetilde{\nu}_l}\big| \leq \frac{1}{2}\big\langle \operatorname{grad} H(\widetilde{\nu}_l), \operatorname{grad} H(\widetilde{\nu}_l)\big\rangle_{\widetilde{\nu}_l} + \frac{1}{2}\big\langle \delta_l, \delta_l\big\rangle_{\widetilde{\nu}_l}. \tag{E.13}$$

Thus, plugging (E.13) into (E.12), we obtain

$$\begin{aligned}
H(\widetilde{\nu}_{l+1}) &\leq H(\widetilde{\nu}_l) - \frac{\eta^\nu(1 - 2\eta^\nu L_2)}{2} \cdot \big\langle \operatorname{grad} H(\widetilde{\nu}_l), \operatorname{grad} H(\widetilde{\nu}_l)\big\rangle_{\widetilde{\nu}_l} \\
&\quad + \frac{\eta^\nu(1 + 2\eta^\nu L_2)}{2} \cdot \big\langle \delta_l, \delta_l\big\rangle_{\widetilde{\nu}_l}.
\end{aligned} \tag{E.14}$$

Furthermore, since H is ξ-gradient dominated under Assumption 2.3, based on (E.14) we have

$$H(\widetilde{\nu}_{l+1}) \leq H(\widetilde{\nu}_l) - \xi\eta^\nu \cdot (1 - 2\eta^\nu L_2) \cdot \big[H(\widetilde{\nu}_l) - \inf_{\nu\in\mathcal{M}(\mathcal{X})} H(\nu)\big] + \frac{\eta^\nu(1 + 2\eta^\nu L_2)}{2} \cdot \big\langle \delta_l, \delta_l\big\rangle_{\widetilde{\nu}_l}. \tag{E.15}$$

As we have introduced in Section 2.1 and 4.1, $\langle \delta_l, \delta_l\rangle_{\widetilde{\nu}_l}$ is equal to ε_l defined in (4.1). Thus, (E.15) can be equivalently written as

$$\begin{aligned}
H(\widetilde{\nu}_{l+1}) - \inf_{\nu\in\mathcal{M}(\mathcal{X})} H(\nu) &\leq [1 - \xi\eta^\nu \cdot (1 - 2\eta^\nu L_2)] \cdot \big[H(\widetilde{\nu}_l) - \inf_{\nu\in\mathcal{M}(\mathcal{X})} H(\nu)\big] \\
&\quad + \frac{\eta^\nu(1 + 2\eta^\nu L_2)}{2} \cdot \bar{\varepsilon}_l.
\end{aligned} \tag{E.16}$$

Hence, we have derived the performance of a single step of variational transport with regard to the objective value.

Moreover, recall that we set the stepsize for player ν to be a constant $\eta^\nu \le 1/(4L_2)$, which guarantees

$$1 - \xi \cdot \eta^\nu \cdot (1 - 2\eta^\nu L_2) \le 1 - \eta^\nu \cdot \xi/2 \in (0, 1), \qquad (1 + 2\eta^\nu L_2)/2 \le 1. \tag{E.17}$$

For simplicity of the notation, we define $\sigma = 1 - \eta^\nu \cdot \xi/2$. Thus, by (E.17), it follows that

$$H(\widetilde{\nu}_{l+1}) - \inf_{\nu \in \mathcal{M}(\mathcal{X})} H(\nu) \le \sigma \cdot \big[H(\widetilde{\nu}_l) - \inf_{\nu \in \mathcal{M}(\mathcal{X})} H(\nu)\big] + \eta^\nu \varepsilon_l. \tag{E.18}$$

By multiplying $\sigma^{-(l+1)}$ to both sides of (E.18), we obtain

$$\sigma^{-(l+1)} \cdot \big[F(\widetilde{\nu}_{l+1}) - \inf_{\nu \in \mathcal{M}(\mathcal{X})} H(\nu)\big] \le \sigma^{-l} \cdot \big[H(\widetilde{\nu}_l) - \inf_{\nu \in \mathcal{M}(\mathcal{X})} H(\nu)\big] + \sigma^{-(l+1)} \cdot \eta^\nu \varepsilon_l. \tag{E.19}$$

Therefore, $\{\sigma^{-l} \cdot [H(\widetilde{\nu}_l) - \inf_{\nu \in \mathcal{M}(\mathcal{X})} H(\nu)]\}_{l \ge 0}$ admits a telescoping sequence and thus by summing (E.19) over l it holds that

$$\sigma^{-l} \cdot \big[H(\widetilde{\nu}_l) - \inf_{\nu \in \mathcal{M}(\mathcal{X})} H(\nu)\big] \le \sum_{m=0}^{l-1} \sigma^{-(m+1)} \cdot \eta^\nu \cdot \varepsilon_m + \big[H(\widetilde{\nu}_0) - \inf_{\nu \in \mathcal{M}(\mathcal{X})} H(\nu)\big], \tag{E.20}$$

which is equivalent to

$$\sigma^{-l} \cdot \big[F(\mu, \widetilde{\nu}^*(\mu)) - F(\mu, \widetilde{\nu}_l)\big] \le \sum_{m=0}^{l-1} \sigma^{-(m+1)} \cdot \eta^\nu \cdot \varepsilon_m + \big[F(\mu, \widetilde{\nu}^*(\mu)) - F(\mu, \widetilde{\nu}_0)\big] \tag{E.21}$$

for all $l \ge 1$. Thus, we obtain the desired result by taking expectation with respect to the intial particle sampling X_0^ν for player ν. □

Lemma E.5 characterizes the expected error of the inner loop upper bounded by the sum of an optimization error decaying at a linear rate and a statistical error term of the order $\mathcal{O}(N_\nu^{-1/2})$ according to §4.1. Moreover, (E.10) justifies that the particle-based functional gradient descent in W_2-space for a Riemannian PL objective behaves similarly as in finite-dimensional spaces [57] up to a scaled statistical error.

By Integrating the convergence result for one-player variational transport processes, together with the properties of relation between the inner loop and outer loop as well as objective landscape, we are now ready to demonstrate the proof of main theorem for infinite-dimensional distributional game optimizaiton.

E.3 Proof of Theorem 4.3

Theorem E.6 (Convergence of Infinite-Dimensional PL Games, Formal)**.** Suppose that the objective F admits a variational form under Assumption 2.2 and 2.3. Also, the function class $\widetilde{\mathcal{F}}$ satisfies Assumption 4.1. We set the stepsizes to be $\eta^\mu \in \big[0, \min\{1/h, 2/\widetilde{L}\}\big)$ and $\eta^\nu \in \big(0, \min\{1/(4L_\nu), 1/h\}\big)$, where $\widetilde{L} = L_1 + L_0^2/\xi$. Then, for any $\theta > 0$, if

$$K_\nu \ge K_\nu(\theta) = \mathcal{O}\Big(\log \frac{(1-\sigma)\widehat{M}_H - \eta^\nu \bar{\varepsilon}_\nu}{\theta} \Big/ \log \frac{1}{\sigma}\Big), \quad \text{where } \widehat{M}_H = \max\Big\{M_H, \frac{\eta^\nu \bar{\varepsilon}_\nu + 1}{1-\sigma}\Big\}, \tag{E.22}$$

there exists an iteration $k \in [K_\mu]$ such that

$$\mathbb{E}_{X_0}\big[\mathcal{J}_\mu^2(\widetilde{\mu}_k, \widetilde{\nu}_{k+1})\big] = \mathcal{O}\bigg(\frac{2}{\widehat{L}}\big(\Delta + \sqrt{\bar{\varepsilon}_\mu} + 2\widetilde{L}R + L_G\big)^2 \cdot \Big(2R(\Delta + \sqrt{\bar{\varepsilon}_\mu}) + \frac{M_G}{K_\mu}\Big)\bigg), \tag{E.23}$$

$$\mathbb{E}_{X_0}\big[\mathcal{J}_\nu(\widetilde{\mu}_k, \widetilde{\nu}_{k+1})\big] = \mathcal{O}\Big(\frac{L_2 \Delta}{L_0}\Big). \tag{E.24}$$

Here $\Delta = L_0\sqrt{\dfrac{\eta_\nu \bar{\varepsilon}_\nu + \theta}{2\xi(1-\sigma)}}$, $\widehat{L} = 1/\eta^\mu - \widetilde{L}/2$, $R = \sup_{\mu_1, \mu_2 \in \mathcal{M}} \mathcal{W}_2(\mu_1, \mu_2)/2$, and the gradient error terms $\bar{\varepsilon}_\mu$ and $\bar{\varepsilon}_\nu$ are characterized in (4.3).

Proof of Theorem 4.3. We write $\widehat{\operatorname{grad}}_\mu F(\widetilde{\mu}_k, \widetilde{\nu}_{k+1}) = -\operatorname{div}(\widetilde{\mu}_k \cdot \nabla \widetilde{f}_k^*)$, where $\nabla \widetilde{f}_k^*$ is the solution to the VFM problem in timestep k, as the gradient estimate *w.r.t.* μ at timestep k. By the particle pushing step in Algorithm 2 we have

$$\big\langle \widehat{\operatorname{grad}}_\mu F(\widetilde{\mu}_k, \widetilde{\nu}_{k+1}), \operatorname{Exp}_{\widetilde{\mu}_k}^{-1}(\widetilde{\mu}_{k+1}) \big\rangle_{\widetilde{\mu}_k} = -\frac{1}{\eta^\mu} \big\langle \operatorname{Exp}_{\widetilde{\mu}_k}^{-1}(\widetilde{\mu}_{k+1}), \operatorname{Exp}_{\widetilde{\mu}_k}^{-1}(\widetilde{\mu}_{k+1}) \big\rangle_{\widetilde{\mu}_k}. \tag{E.25}$$

It follows that

$$\begin{aligned}
&\big\langle \operatorname{grad}_\mu F(\widetilde{\mu}_k, \nu^*(\widetilde{\mu}_k)), \operatorname{Exp}_{\widetilde{\mu}_k}^{-1}(\widetilde{\mu}_{k+1}) \big\rangle_{\widetilde{\mu}_k} \\
&= \big\langle \operatorname{grad}_\mu F(\widetilde{\mu}_k, \widetilde{\nu}^*(\widetilde{\mu}_k)) - \widehat{\operatorname{grad}}_\mu F(\widetilde{\mu}_k, \widetilde{\nu}_{k+1}), \operatorname{Exp}_{\widetilde{\mu}_k}^{-1}(\widetilde{\mu}_{k+1}) \big\rangle_{\widetilde{\mu}_k} \\
&- \frac{1}{\eta^\mu} \big\langle \operatorname{Exp}_{\widetilde{\mu}_k}^{-1}(\widetilde{\mu}_{k+1}), \operatorname{Exp}_{\widetilde{\mu}_k}^{-1}(\widetilde{\mu}_{k+1}) \big\rangle_{\widetilde{\mu}_k} \\
&= \big\langle \Delta_k - \delta_k, \operatorname{Exp}_{\widetilde{\mu}_k}^{-1}(\widetilde{\mu}_{k+1}) \big\rangle_{\widetilde{\mu}_k} - \frac{1}{\eta^\mu} \big\langle \operatorname{Exp}_{\widetilde{\mu}_k}^{-1}(\widetilde{\mu}_{k+1}), \operatorname{Exp}_{\widetilde{\mu}_k}^{-1}(\widetilde{\mu}_{k+1}) \big\rangle_{\widetilde{\mu}_k},
\end{aligned} \tag{E.26}$$

where $\nu^*(\widetilde{\mu}_k) \in \mathcal{I}_F(\widetilde{\mu}_k)$ and

$$\delta_k = -\operatorname{div}\big[\widetilde{\mu}_k \cdot (\nabla \widetilde{f}_k^* - \nabla f_k^*)\big], \tag{E.27}$$

$$\Delta_k = \operatorname{grad} G(\widetilde{\mu}_k) - \operatorname{grad}_\mu F(\widetilde{\mu}_k, \widetilde{\nu}_{k+1}). \tag{E.28}$$

Hence, we have $\Delta_k - \delta_k = \operatorname{grad}_\mu F(\widetilde{\mu}_k, \nu^*(\widetilde{\mu}_k)) - \widehat{\operatorname{grad}}_\mu F(\widetilde{\mu}_k, \widetilde{\nu}_{k+1}) = \operatorname{grad} G(\widetilde{\mu}_k) - \widehat{\operatorname{grad}}_\mu F(\widetilde{\mu}_k, \widetilde{\nu}_{k+1})$ for any $k \in [K_\mu]$. Note that the error term Δ_k is incurred by using $\widetilde{\nu}_{k+1}$ to approximate $\nu^*(\widetilde{\mu}_k)$.

By Assumption 4.1, ∇f is h-Lipschitz continuous on $\mathcal{X}$ for all $f \in \widetilde{\mathcal{F}}$. Since $\eta^\mu < 1/h$, by Proposition 3.2, we can equivalently write one timestep of the outer loop in VTIG as

$$\begin{aligned}
\widetilde{\mu}_{k+1} &= \operatorname{Exp}_{\widetilde{\mu}_k}\big\{-\eta_k^\mu \cdot \widehat{\operatorname{grad}}_\mu F(\widetilde{\mu}_k, \widetilde{\nu}_{k+1})\big\} \\
&= \operatorname{Exp}_{\widetilde{\mu}_k}\big\{-\eta_k^\mu \cdot [\operatorname{grad}_\mu F(\widetilde{\mu}_k, \widetilde{\nu}_{k+1}) + \delta_k]\big\} &\text{(E.29)} \\
&= \operatorname{Exp}_{\widetilde{\mu}_k}\big\{-\eta^\mu \cdot [\operatorname{grad} G(\widetilde{\mu}_k) - \Delta_k + \delta_k]\big\}, &\text{(E.30)}
\end{aligned}$$

Note that $\delta_k \in T_{\widetilde{\mu}_k}\mathcal{M}$ is a tangent vector at point $\widetilde{\mu}_k$. Since G is $\widetilde{L}$-smooth under Lemma E.4, combining (E.26) and (E.29), we obtain

$$G(\widetilde{\mu}_{k+1}) \le G(\widetilde{\mu}_k) + \big\langle \operatorname{grad}_\mu F(\widetilde{\mu}_k, \nu^*(\widetilde{\mu}_k)), \operatorname{Exp}_{\widetilde{\mu}_k}^{-1}(\widetilde{\mu}_{k+1}) \big\rangle_{\widetilde{\mu}_k} + \frac{\widetilde{L}}{2} \big\langle \operatorname{Exp}_{\widetilde{\mu}_k}^{-1}(\widetilde{\mu}_{k+1}), \operatorname{Exp}_{\widetilde{\mu}_k}^{-1}(\widetilde{\mu}_{k+1}) \big\rangle_{\widetilde{\mu}_k} \tag{E.31}$$

$$\overset{\text{(E.26)}}{=} G(\widetilde{\mu}_k) + \big\langle \Delta_k - \delta_k, \operatorname{Exp}_{\widetilde{\mu}_k}^{-1}(\widetilde{\mu}_{k+1}) \big\rangle_{\widetilde{\mu}_k} - \left(\frac{1}{\eta^\mu} - \frac{\widetilde{L}}{2}\right) \big\langle \operatorname{Exp}_{\widetilde{\mu}_k}^{-1}(\widetilde{\mu}_{k+1}), \operatorname{Exp}_{\widetilde{\mu}_k}^{-1}(\widetilde{\mu}_{k+1}) \big\rangle_{\widetilde{\mu}_k}.$$

Since $\eta^\mu < 2/\widetilde{L}$, a lower bound of $\|\operatorname{Exp}_{\widetilde{\mu}_k}^{-1}(\widetilde{\mu}_{k+1})\|$ is needed. On the other hand, we have for any $\mu \in \mathcal{M}$,

$$\begin{aligned}
\big\langle \widehat{\operatorname{grad}}_\mu F(\widetilde{\mu}_k, \widetilde{\nu}_{k+1}), \operatorname{Exp}_{\widetilde{\mu}_k}^{-1}(\mu) \big\rangle_{\widetilde{\mu}_k} &\ge -\|\widehat{\operatorname{grad}}_\mu F(\widetilde{\mu}_k, \widetilde{\nu}_{k+1})\|_{\widetilde{\mu}_k} \cdot \|\operatorname{Exp}_{\widetilde{\mu}_k}^{-1}(\mu)\|_\mu \\
&\ge -\|\widehat{\operatorname{grad}}_\mu F(\widetilde{\mu}_k, \widetilde{\nu}_{k+1})\|_{\widetilde{\mu}_k} \big(\|\operatorname{Exp}_{\widetilde{\mu}_{k+1}}^{-1}(\mu)\|_\mu + \|\operatorname{Exp}_{\widetilde{\mu}_k}^{-1}(\widetilde{\mu}_{k+1})\|_{\widetilde{\mu}_{k+1}}\big) \\
&\ge -\frac{1}{\eta^\mu} \|\operatorname{Exp}_{\widetilde{\mu}_k}^{-1}(\widetilde{\mu}_{k+1})\|_{\widetilde{\mu}_{k+1}} \|\operatorname{Exp}_{\widetilde{\mu}_{k+1}}^{-1}(\mu)\|_\mu \\
&- \big(\|\Delta_k\| + \|\delta_k\| + \|\operatorname{grad} G(\widetilde{\mu}_k)\|\big) \|\operatorname{Exp}_{\widetilde{\mu}_k}^{-1}(\widetilde{\mu}_{k+1})\|_{\widetilde{\mu}_{k+1}},
\end{aligned} \tag{E.32}$$

where the first inequality follows from Cauchy-Schwartz inequality (E.32) is implied by triangle inequality *w.r.t.* Riemannian metric, and the last one follows from (E.25) and (E.26). Then by definitions $\|\operatorname{grad} G(\widetilde{\mu}_k)\| \le G_{\max}$ and $\|\operatorname{Exp}_{\widetilde{\mu}_k}^{-1}(\widetilde{\mu}_{k+1})\| \le 2R$. Additionally, by Theorem E.5 using fixed $\mathbb{E}\epsilon_m = \bar{\epsilon}_\nu$

with expectation *w.r.t.* initial sampled particles, we have

$$\begin{aligned}\mathbb{E}\|\Delta_k\| &\le L_0 \mathbb{E} W_2(\widetilde{\nu}_{k+1}, \nu^*(\widetilde{\mu}_k)) \\ &\le L_0 \sqrt{\frac{\sigma^{K_\nu}[F(\widetilde{\mu}_k, \nu^*(\widetilde{\mu}_k)) - F(\widetilde{\mu}_k, \widetilde{\nu}_0)] + \eta^\nu \bar{\varepsilon}_\nu \cdot \frac{1-\sigma^{K_\nu}}{1-\sigma}}{2\xi}} \\ &\le L_0 \sqrt{\frac{\eta^\nu \bar{\varepsilon}_\nu + \theta}{2\xi(1-\sigma)}}, \end{aligned} \tag{E.33}$$

where the second inequality holds by applying Definition E.2 to the gradient dominated function $H_\mu(\nu) = -F(\mu, \nu)$ together with Theorem E.5, and the last inequality follows from the choice of K_ν in the theorem. Then by $\|\operatorname{grad} G(\widetilde{\mu}_k)\| \le G_{\max}$ and $\|\mathrm{Exp}_{\widetilde{\mu}_k}^{-1}(\widetilde{\mu}_{k+1})\| \le 2R$, combining (E.32) we obtain

$$-\mathcal{J}_\mu(\widetilde{\mu}_k, \widetilde{\nu}_{k+1}) \ge -\left(\|\Delta_k\| + \|\delta_k\| + G_{\max} + 2R/\eta^\mu\right) \|\mathrm{Exp}_{\widetilde{\mu}_k}^{-1}(\widetilde{\mu}_{k+1})\|, \tag{E.34}$$

that is,

$$\|\mathrm{Exp}_{\widetilde{\mu}_k}^{-1}(\widetilde{\mu}_{k+1})\| \ge \frac{\mathcal{J}_\mu(\widetilde{\mu}_k, \widetilde{\nu}_{k+1})}{\|\Delta_k\| + \|\delta_k\| + G_{\max} + 2\widetilde{L}R}. \tag{E.35}$$

Plugging (E.35) into (E.31), we obtain the following progress made by pushing μ-particles in the outer loop.

$$\begin{aligned} G(\mu_{k+1}) &\le G(\widetilde{\mu}_k) + \left\langle \Delta_k - \delta_k, \mathrm{Exp}_{\widetilde{\mu}_k}^{-1}(\widetilde{\mu}_{k+1}) \right\rangle_{\widetilde{\mu}_k} - \frac{(1/\eta^\mu - \widetilde{L}/2)\mathcal{J}_\mu^2(\widetilde{\mu}_k, \widetilde{\nu}_{k+1})}{\left(\|\Delta_k\| + \|\delta_k\| + G_{\max} + 2\widetilde{L}R\right)^2} \\ &\le G(\widetilde{\mu}_k) + 2R\left(\|\Delta_k\| + \|\delta_k\|\right) - \frac{(1/\eta^\mu - \widetilde{L}/2)\mathcal{J}_\mu^2(\widetilde{\mu}_k, \widetilde{\nu}_{k+1})}{\left(\|\Delta_k\| + \|\delta_k\| + G_{\max} + 2\widetilde{L}R\right)^2}, \end{aligned} \tag{E.36}$$

where the last inequality holds by applying the Cauchy-Schwartz inequality to the inner product term $\left\langle \Delta_k - \delta_k, \mathrm{Exp}_{\widetilde{\mu}_k}^{-1}(\widetilde{\mu}_{k+1}) \right\rangle_{\widetilde{\mu}_k}$ and $\widetilde{\mu}_k$'s lie in a ball of radius R.

Therefore, we have a telescoping sequence and by summing (E.36) over k for the whole loop while taking expection with respect to initial sampled particles on both sides we get

$$\begin{aligned} &\frac{1}{K_\mu}\sum_{k=0}^{K_\mu} \mathbb{E}\mathcal{J}_\mu^2(\widetilde{\mu}_k, \widetilde{\nu}_{k+1}) \\ &\le \frac{2}{1/\eta^\mu - \widetilde{L}/2} \mathbb{E}\left(2\widetilde{L} + G_{\max} + \|\Delta_k\| + \|\delta_k\|\right)^2 \cdot \left(2R\left(\|\Delta_k\| + \|\delta_k\|\right) + \frac{G(\widetilde{\mu}_0) - G(\widetilde{\mu}_{K_\mu})}{K_\mu}\right) \\ &\le \mathcal{O}\left(\frac{2}{1/\eta^\mu - \widetilde{L}/2}\left(\Delta + \sqrt{\bar{\varepsilon}_\mu} + 2\widetilde{L}R + G_{\max}\right)^2 \cdot \left(2R(\Delta + \sqrt{\bar{\varepsilon}_\mu}) + \frac{M_G}{K_\mu}\right)\right), \end{aligned} \tag{E.37}$$

$$\tag{E.38}$$

where the last inequality follows from (E.33). Here

$$\Delta = L_0 \sqrt{\frac{\eta_\nu \bar{\varepsilon}_\nu + \theta}{2\xi(1-\sigma)}}, \tag{E.39}$$

and

$$\sqrt{\bar{\varepsilon}_k} = \mathcal{O}\left(\sqrt{\widetilde{K}\left(\frac{r_f^2}{N^{1/2}} + \frac{r_f^3}{w^{1/2}} + \frac{r_f^4}{w}\right)}\right).$$

Therefore, by the Pigeonhole principle there must exists $k \in [K_\mu]$ such that

$$\mathbb{E}\mathcal{J}_\mu^2(\widetilde{\mu}_k, \widetilde{\nu}_{k+1}) = \mathcal{O}\left(\frac{2}{1/\eta^\mu - \widetilde{L}/2} \left(\Delta + \sqrt{\bar{\varepsilon}_\mu} + 2\widetilde{L}R + G_{\max} \right)^2 \cdot \left(2R(\Delta + \sqrt{\bar{\varepsilon}_\mu}) + \frac{M_G}{K_\mu} \right) \right).$$

Hence we complete the proof of the first term in (4.5) of Theorem 4.3.

For the second result on $\mathcal{J}_\nu(\mu_k, \nu_{k+1})$, we follow the similar technique by the Cauchy-Schwartz inequality and smoothnes of function F with respect to ν in Assumption 2.2 to obtain

$$\begin{aligned}
\mathbb{E}\mathcal{J}_\nu(\widetilde{\mu}_k, \widetilde{\nu}_{k+1}) &\leq \mathbb{E}\| \operatorname{grad}_\nu F(\widetilde{\mu}_k, \widetilde{\nu}_{k+1})\|_{\widetilde{\nu}_{k+1}} \\
&= \mathbb{E}\boldsymbol{d}\big(\operatorname{grad}_\nu F(\widetilde{\mu}_k, \nu^*(\widetilde{\mu}_k)), \operatorname{grad}_\nu F(\widetilde{\mu}_k, \widetilde{\nu}_{k+1}) \big) \\
&\leq L_2 \mathbb{E}\mathcal{W}_2\big(\widetilde{\nu}_{k+1}, \widetilde{\nu}^*(\mu_k)\big) \\
&\overset{\text{(E.1)}}{\leq} L_2 \sqrt{\frac{F(\widetilde{\mu}_k, \widetilde{\nu}^*(\mu_k)) - \mathbb{E}F(\widetilde{\mu}_k, \widetilde{\nu}_{k+1})}{2\xi}} \\
&\leq L_2 \sqrt{\frac{\sigma^{K_\nu}[F(\widetilde{\mu}_k, \nu^*(\widetilde{\mu}_k)) - F(\widetilde{\mu}_k, \widetilde{\nu}_0)] + \eta^\nu \bar{\varepsilon}_\nu \cdot \frac{1-\sigma^{K_\nu}}{1-\sigma}}{2\xi}} \\
&\leq \frac{L_2}{L_0} \cdot \Delta, \qquad \text{(E.40)}
\end{aligned}$$

where (E.1) relates Wasserstein distance to the objective difference, the last but one inequality follows from Theorem E.5, the last inequality holds by the definition of Δ in (E.39) and the same choice of K_μ in the proof for a desired $\mathcal{J}_\mu(\widetilde{\mu}_k, \widetilde{\nu}_{k+1})$ ($k \in [K_\mu]$). Therefore, the second term in (4.5) of Theorem 4.3 is proved.

To sum up, we conclude the proof of our main theorem and the sketch of proof ideas. □

We additionally remark that for our convergence rate to the approximate IFNE coincides with the finite-dimensional case regardless of the order of numbers of particles adopted. More clearly, we resort to writing our result in terms of the θ-IFNE for some $\theta > 0$. First, we can alternatively claim that there exists $k \in [K_\mu]$ such that the following equivalent result holds,

$$\mathbb{E}\mathcal{J}_\mu(\widetilde{\mu}_k, \widetilde{\nu}_{k+1}) = \mathcal{O}\left(\left(\Delta + \sqrt{\bar{\varepsilon}_\mu} + 2\widetilde{L}R + G_{\max} \right) \cdot \sqrt{\frac{2}{1/\eta^\mu - \widetilde{L}/2} \left(2R(\Delta + \sqrt{\bar{\varepsilon}_\mu}) + \frac{M_G}{K_\mu} \right)} \right), \qquad \text{(E.41)}$$

$$\mathbb{E}\mathcal{J}_\nu(\widetilde{\mu}_k, \widetilde{\nu}_{k+1}) = \mathcal{O}\left(\frac{L_2}{L_0} \cdot \Delta \right). \qquad \text{(E.42)}$$

By fixing N_μ and N_ν for any $\theta > 0$, we observe that when we set $K_\mu \geq K_\mu(\theta) = \mathcal{O}(\theta^{-2})$ and $K_\nu \geq K_\nu(\theta) = \mathcal{O}(2\log(\theta^{-1}))$ simultaneously, we are able to obtain a θ-IFNE in expectation since

$$\mathbb{E}\mathcal{J}_\mu(\widetilde{\mu}_k, \widetilde{\nu}_{k+1}) \leq \theta, \qquad \text{(E.43)}$$

$$\mathbb{E}\mathcal{J}_\nu(\widetilde{\mu}_k, \widetilde{\nu}_{k+1}) \leq \theta, \qquad \text{(E.44)}$$

which is exactly the order in [57], implying the PL-condition and smoothness in infinite-dimensional settings work similarly as in finite-dimensional problems.

On the other hand, when the order of N_μ and N_ν dominate, we can compute that when $N_\mu \geq N_\mu(\theta) = \mathcal{O}(\theta^{-4})$ and $N_\nu \geq N_\nu(\theta) = \mathcal{O}(\theta^{-4})$ for any $\theta > 0$, we obtain the θ-IFNE in expectation. Such a result implies that we need more particles than timesteps to guarantee a given precision of the solution. This also implies that the statistical error induced by measure approximation in infinite-dimensional spaces is more prominent than the optimization error.

F Proof of Convergence to the Minimax Value of Two-Sided PL Games

F.1 PL Condition on the Inner Maximization Functional

Lemma F.1. For a two-sided PL-game depicted by Assumption 4.4, the functional $G(\mu) = \max_{\nu\in\mathcal{M}} F(\mu,\nu)$ satisfies the ξ_1-PL condition.

Proof. Since $F_\nu(\mu) = F(\mu,\nu)$ satisfies the ξ_1-PL condition for any given $\nu \in \mathcal{M}(\mathcal{X}_\nu)$, we have

$$2\xi_1 \cdot \left(F(\mu,\nu^*(\mu)) - \min_{\widetilde{\mu}\in\mathcal{M}(\mathcal{X}_\mu)} F\big(\widetilde{\mu},\nu^*(\mu)\big)\right) \le \big\langle \operatorname{grad} F(\mu,\nu^*(\mu)), \operatorname{grad} F(\mu,\nu^*(\mu))\big\rangle_\mu$$
$$= \big\langle \operatorname{grad} G(\mu), \operatorname{grad} G(\mu)\big\rangle_\mu, \tag{F.1}$$

where the last equality follows from Lemma E.4. On the other hand, we have by definition

$$\min_{\widetilde{\mu}\in\mathcal{M}(\mathcal{X}_\mu)} F(\widetilde{\mu},\nu^*(\mu)) \le \min_{\widetilde{\mu}\in\mathcal{M}(\mathcal{X}_\mu)} \max_{\nu\in\mathcal{M}(\mathcal{X}_\nu)} F(\widetilde{\mu},\nu) = \min_{\mu\in\mathcal{M}(\mathcal{X}_\mu)} G(\mu). \tag{F.2}$$

By plugging (F.2) into (F.1), we arrive at

$$2\xi_1 \cdot \left(G(\mu) - \min_{\mu\in\mathcal{M}(\mathcal{X}_\mu)} G(\mu)\right) \le \big\langle \operatorname{grad} G(\mu), \operatorname{grad} G(\mu)\big\rangle_\mu, \tag{F.3}$$

which implies that $G(\mu)$ satisfies ξ_1-PL condition. □

F.2 Proof of Theorem 4.5

Proof. The proof is based on Lemma E.5 while different from the proof of Theorem 4.3 by the fact that we are now able to bound the norm of the Riemannian gradient of $G(\mu)$ from below. By the smoothness of G under Lemma E.4, we have

$$G(\widetilde{\mu}_{k+1}) \le G(\widetilde{\mu}_k) + \big\langle \operatorname{grad}_\mu F(\widetilde{\mu}_k,\nu^*(\widetilde{\mu}_k)), \operatorname{Exp}^{-1}_{\widetilde{\mu}_k}(\widetilde{\mu}_{k+1})\big\rangle_{\widetilde{\mu}_k} + \frac{\widetilde{L}}{2}\big\langle \operatorname{Exp}^{-1}_{\widetilde{\mu}_k}(\mu_{k+1}), \operatorname{Exp}^{-1}_{\widetilde{\mu}_k}(\mu_{k+1})\big\rangle_{\widetilde{\mu}_k} \tag{F.4}$$

$$\overset{(E.29)}{=} G(\widetilde{\mu}_k) - \eta^\mu \big\langle \operatorname{grad} G(\widetilde{\mu}_k), \operatorname{grad} G(\widetilde{\mu}_k)\big\rangle_{\widetilde{\mu}_k} - \eta^\mu \big\langle \operatorname{grad} G(\widetilde{\mu}_k), \delta_k - \Delta_k\big\rangle_{\widetilde{\mu}_k}$$
$$+ \frac{\widetilde{L}}{2}\big\langle \operatorname{grad} G(\widetilde{\mu}_k) - \Delta_k + \delta_k, \operatorname{grad} G(\widetilde{\mu}_k) - \Delta_k + \delta_k\big\rangle_{\widetilde{\mu}_k}$$
$$\le G(\widetilde{\mu}_k) - \left(\eta^\mu - (\eta^\mu)^2 \cdot \frac{\widetilde{L}}{2}\right) \cdot \big\langle \operatorname{grad} G(\widetilde{\mu}_k), \operatorname{grad} G(\widetilde{\mu}_k)\big\rangle_{\widetilde{\mu}_k}$$
$$+ \left(\eta^\mu + \widetilde{L}(\eta^\mu)^2\right) \big|\big\langle \operatorname{grad} G(\widetilde{\mu}_k), \delta_k - \Delta_k\big\rangle_{\widetilde{\mu}_k}\big| + \frac{(\eta^\mu)^2\widetilde{L}}{2}\big\langle \delta_k - \Delta_k, \delta_k - \Delta_k\big\rangle_{\widetilde{\mu}_k}. \tag{F.5}$$

Here $\delta_k = \widehat{\operatorname{grad}}_\mu F(\widetilde{\mu}_k,\widetilde{\nu}_{k+1}) - \operatorname{grad}_\mu F(\widetilde{\mu}_k,\widetilde{\nu}_{k+1}) = -\operatorname{div}\big[\widetilde{p}_k\cdot(\nabla \widetilde{f}_k^* - \nabla f_k^*)\big]$ and $\Delta_k = \operatorname{grad} G(\widetilde{\mu}_k) - \operatorname{grad}_\mu F(\widetilde{\mu}_k,\widetilde{\nu}_{k+1})$ are defined in (E.27). Since $\langle\cdot,\cdot\rangle_{\widetilde{\mu}_k}$ serves as the Riemannian metric of $\mathcal{M}(\mathcal{X}_\mu)$ at $\widetilde{\mu}_k$, then by the Cauchy-Schwartz inequality we obtain

$$2\big|\big\langle \operatorname{grad} G(\widetilde{\mu}_k), \delta_k - \Delta_k\big\rangle_{\widetilde{\mu}_k}\big| \le \big\langle \operatorname{grad} G(\widetilde{\mu}_k), \operatorname{grad} G(\widetilde{\mu}_k)\big\rangle_{\widetilde{\mu}_k} + \big\langle \delta_k - \Delta_k, \delta_k - \Delta_k\big\rangle_{\widetilde{\mu}_k}$$
$$\le \big\langle \operatorname{grad} G(\widetilde{\mu}_k), \operatorname{grad} G(\widetilde{\mu}_k)\big\rangle_{\widetilde{\mu}_k} + 2\big\langle \delta_k, \delta_k\big\rangle_{\widetilde{\mu}_k} + 2\big\langle \Delta_k, \Delta_k\big\rangle_{\widetilde{\mu}_k}. \tag{F.6}$$

Hence by plugging (F.6) into (F.4), we have

$$G(\widetilde{\mu}_{k+1}) \le G(\widetilde{\mu}_k) + \frac{\eta^\mu(2\eta^\mu\widetilde{L} - 1)}{2} \cdot \big\langle \operatorname{grad} G(\widetilde{\mu}_k), \operatorname{grad} G(\widetilde{\mu}_k)\big\rangle_{\widetilde{\mu}_k}$$
$$+ \eta^\mu\left(1 + 2(\eta^\mu)^2\widetilde{L}\right) \cdot \left(\big\langle \delta_k, \delta_k\big\rangle_{\widetilde{\mu}_k} + \big\langle \Delta_k, \Delta_k\big\rangle_{\widetilde{\mu}_k}\right). \tag{F.7}$$

Since $\eta^\mu \le 1/(2\widetilde{L})$, we require a lower bound on $\langle \operatorname{grad} G(\widetilde{\mu}_k), \operatorname{grad} G(\widetilde{\mu}_k)\rangle_{\widetilde{\mu}_k}$. Following from the ξ_1-PL condition of G in Lemma F.1, we can further obtain

$$G(\widetilde{\mu}_{k+1}) \le G(\widetilde{\mu}_k) + \xi_1\eta^\mu(2\eta^\mu\widetilde{L}-1)\cdot\left(G(\widetilde{\mu}_k) - \min_{\mu\in\mathcal{M}(\mathcal{X}_\mu)} G(\mu)\right) + \eta^\mu\left(1+2(\eta^\mu)^2\widetilde{L}\right)\cdot\left(\langle\delta_k,\delta_k\rangle_{\widetilde{\mu}_k} + \langle\Delta_k,\Delta_k\rangle_{\widetilde{\mu}_k}\right), \tag{F.8}$$

which can be rewritten as

$$G(\widetilde{\mu}_{k+1}) - \min_{\mu\in\mathcal{M}(\mathcal{X}_\mu)} G(\mu) \le [1-\xi_1\eta^\mu(1-2\eta^\mu\widetilde{L})]\cdot\left(G(\widetilde{\mu}_k) - \min_{\mu\in\mathcal{M}(\mathcal{X}_\mu)} G(\mu)\right) + \eta^\mu\left(1+2(\eta^\mu)^2\widetilde{L}\right)\cdot\left(\langle\delta_k,\delta_k\rangle_{\widetilde{\mu}_k} + \langle\Delta_k,\Delta_k\rangle_{\widetilde{\mu}_k}\right). \tag{F.9}$$

Now we invoke the upper bounds on $\|\delta_k\|_{\widetilde{\mu}_k}$ in (D.41) and $\|\Delta_k\|_{\widetilde{\mu}_k}$ in (E.33) to further develop a contraction for the value of G at each timestep,

$$\mathbb{E}G(\widetilde{\mu}_{k+1}) - \min_{\mu\in\mathcal{M}(\mathcal{X}_\mu)} G(\mu) \le [1-\xi_1\eta^\mu(1-2\eta^\mu\widetilde{L})]\cdot\left(\mathbb{E}G(\widetilde{\mu}_k) - \min_{\mu\in\mathcal{M}(\mathcal{X}_\mu)} G(\mu)\right) + \eta^\mu\left(1+2\eta^\mu\widetilde{L}\right)\cdot\left(\mathbb{E}\langle\delta_k,\delta_k\rangle_{\widetilde{\mu}_k} + \widetilde{\Delta}^2\right), \tag{F.10}$$

where we define

$$\widetilde{\Delta}^2 = L_0^2\cdot\frac{\sigma^{K_\nu}\cdot M_H + \eta^\nu\bar{\varepsilon}_\nu\cdot\frac{1-\sigma^{K_\nu}}{1-\sigma}}{2\xi_2}, \tag{F.11}$$

with ξ_2 as the parameter of PL condition for the inner-loop player, K_ν as the number of timesteps for the inner-loop player defined in Algorithm 1, and the expectation is taken with respect to the randomness of initial particles. Also, by the definitions of $\widetilde{L}$, ξ_1, and $\eta^\mu \in \left(0, 1/(4\widetilde{L})\right)$, we have $1-\xi_1\eta^\mu(1-2\eta^\mu\widetilde{L}) < 1-\xi_1\eta^\mu/2 < 1$. For simplicity, let $\widetilde{\sigma} = 1-\xi_1\eta^\mu/2$, by multiplying $\widetilde{\sigma}^{-(k+1)}$ to (F.10) we have

$$\widetilde{\sigma}^{-(k+1)}\cdot[\mathbb{E}G(\widetilde{\mu}_{k+1}) - \min_{\mu\in\mathcal{M}(\mathcal{X}_\mu)} G(\mu)] \le \widetilde{\sigma}^{-k}\cdot[\mathbb{E}G(\widetilde{\mu}_k) - \min_{\mu\in\mathcal{M}(\mathcal{X}_\mu)} G(\mu)] + \widetilde{\sigma}^{-(k+1)}\cdot\eta^\mu\cdot\left(\mathbb{E}\langle\delta_k,\delta_k\rangle_{\widetilde{\mu}_k} + L_0^2\cdot\frac{\eta^\nu\bar{\varepsilon}_\nu+\theta}{2\xi(1-\sigma)}\right), \tag{F.12}$$

Note that $\mathbb{E}\langle\delta_k,\delta_k\rangle_{\widetilde{\mu}_k} = \bar{\varepsilon}_k$, which is bounded in (D.41).. Then by summing up over the telescoping sequence $\{\widetilde{\sigma}^{-k}\cdot[\mathbb{E}G(\widetilde{\mu}_k) - \min_{\mu\in\mathcal{M}(\mathcal{X}_\mu)} G(\mu)]\}_{k\ge 0}$ in (F.12), we obtain the following optimization error bound at iteration k for $k \ge 1$,

$$\begin{aligned}\mathbb{E}G(\widetilde{\mu}_k) - \min_{\mu\in\mathcal{M}(\mathcal{X}_\mu)} G(\mu) &\le \widetilde{\sigma}^k\cdot[G(\widetilde{\mu}_0) - \min_{\mu\in\mathcal{M}(\mathcal{X}_\mu)} G(\mu)] + \sum_{m=0}^{k-1}\widetilde{\sigma}^{k-(m+1)}\cdot\eta^\mu\cdot\left(\bar{\varepsilon}_m + L_0^2\cdot\frac{\eta^\nu\bar{\varepsilon}_\nu+\theta}{2\xi(1-\sigma)}\right)\\ &\overset{\text{(D.41)}}{\le} \widetilde{\sigma}^k\cdot[G(\widetilde{\mu}_0) - \min_{\mu\in\mathcal{M}(\mathcal{X}_\mu)} G(\mu)] + \frac{1-\widetilde{\sigma}^k}{1-\widetilde{\sigma}}\cdot\eta^\mu\cdot\left(\bar{\varepsilon}_\mu + L_0^2\cdot\frac{\eta^\nu\bar{\varepsilon}_\nu+\theta}{2\xi(1-\sigma)}\right),\end{aligned} \tag{F.13}$$

where

$$\bar{\varepsilon}_\mu = \mathcal{O}\left(\widetilde{K}\left(\frac{r_f^2}{N_\mu^{1/2}} + \frac{r_f^3}{w^{1/2}} + \frac{r_f^4}{w}\right)\right),$$
$$\widetilde{\Delta}^2 = L_0^2\cdot\frac{\sigma^{K_\nu}\cdot M_H + \eta^\nu\bar{\varepsilon}_\nu\cdot\frac{1-\sigma^{K_\nu}}{1-\sigma}}{2\xi_2}. \tag{F.14}$$

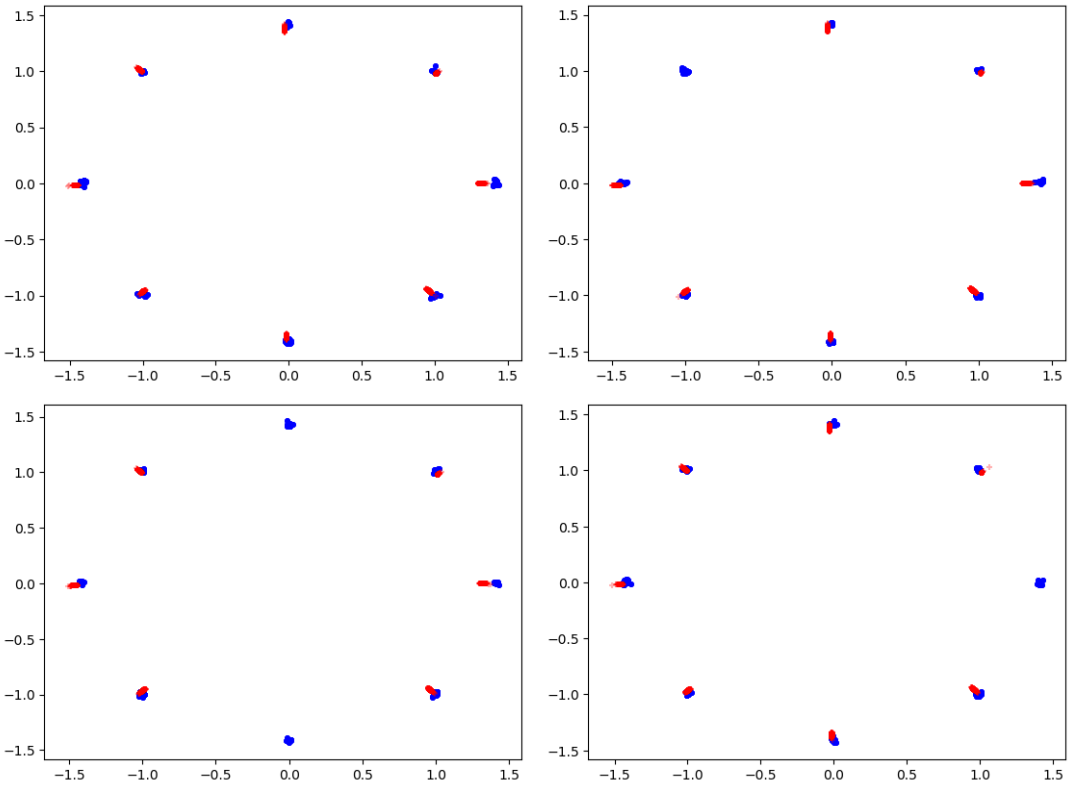

Figure 2: Toy experiment results: blue dots represent samples from the gaussian mixture and the red dots represent the samples generated from generators.

By the definition of G, we have $\min_{\mu \in \mathcal{M}(\mathcal{X})} G(\mu) = F(\mu^*, \nu^*)$. As a result, (F.13) gives the error bound in objective value F at iteration k. Hence we conclude the proof of Theorem 4.5. □

G Toy Experiments

In this section we report some results for a toy experiment with a Gaussian mixture model with 8 Gaussian distributions. For simplicity, we drop the regularizer terms from WGAN loss and consider a mixture of 8 generators and discriminators corresponding to the particles for parameters of the generator and the discriminator of WGAN. Both generators and discriminators are MLP with 3 layers. We also don't tune the learning rate and set it to be 10^{-4}. We run the model for 20000 iterations which is small compared to the typical number of iterations used in practice to train a WGAN model. In our experiment we reused the code provided by [30] with some simple modification. We present some samples generated from trained generators in Figure 2. The blue dots are generated from real mixture models and the red ones are generated from generators. We observe that the distribution generated by our generator matches the groundtruth after a short training period.