Infinite-Dimensional Optimization for Zero-Sum Games via Variational Transport

Lewis Liu [1] **Yufeng Zhang** [2] **Zhuoran Yang** [3] **Reza Babanezhad** [4] **Zhaoran Wang** [2]

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 zero-sum games by a min-max 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 zero-sum games. We then aim to answer the following question:
Will GDA-type algorithms still be provably efficient when extended to infinite-dimensional zero-sum 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 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 that a theoretical version of 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. 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 is of independent interest.

[1]Université de Montréal, Canada [2]Northwestern University, United States [3]Princeton University, United States [4]Samsung SAIT AI Lab, Canada. Correspondence to: Lewis Liu <allis.algo2@gmail.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

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. The optimal strategies define the *pure* NE in the sense that they identify two deterministic strategies. Motivating examples include generative adversarial networks (GANs) (Nowozin et al., 2016; Sanjabi et al., 2018a; Gidel et al., 2018a; Sinha et al., 2017), reinforcement learning (Dai et al., 2017; Ho & Ermon, 2016), distributionally robust optimization (DRO) (Van Parys et al., 2017; Ghosh et al., 2018), and learning exponential families (Dai et al., 2018), among others. Such zero-sum games have been extensively analyzed in convex-concave settings, where a global Nash equilibrium (NE) can be computed by gradient descent-ascent (GDA) type algorithms (Facchinei & Pang, 2007; Hamedani et al., 2018; Monteiro & Svaiter, 2010; Nemirovski, 2004). Nonetheless, in the nonconvex-nonconcave setting, these methods stagger, and a crucial issue arises: *What if the pure NE does not exist (Arora et al., 2017; Jin et al., 2019)?* The finite-dimensional formulation naturally excludes a potentially better or even the only existential mixed NE, and meanwhile is restricted to local convergence in the absence of 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) and smoothness conditions, which cover a range of nonconvex-nonconcave landscapes and the practical training objectives such as GANs with regularization (Arora et al., 2017). A natural approach to distributional optimization problems is the particle-based method (Raginsky et al., 2017; Wibisono, 2018; Zou et al., 2018), where stochastic gradient Langevin dynamics (SGLD) is adopted to draw a sample from the desired distribution via discretization of stochastic differential

equations (Hsieh et al., 2018). 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 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 the theoretical version of VTIG in two scenarios. One is the convergence to the first-order NE under a Riemannian PL condition (Theorem 4.2), and the other is the convergence to the minimax value under a stronger two-sided PL condition (Theorem 4.4). (3) As a technical component, we provide statistical analysis for particle-based gradient estimation by upper bounding 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 (Nemirovski, 2004; Juditsky & Nemirovski, 2016; Hamedani et al., 2018; Monteiro & Svaiter, 2010) are adequately studied with corresponding monotonic variational inequalities (Dang & Lan, 2015; Gidel et al., 2018a) and solved by GDA (Thekumparampil et al., 2019). Meanwhile, primal-dual schemes and negative momentum (Chambolle & Pock, 2016; Daskalakis & Panageas, 2018b; Gidel et al., 2018b) are proposed to help GDA on convergence, which bypasses cyclic dynamics (Mai et al., 2018; Mescheder et al., 2018; Daskalakis & Panageas, 2018a). To tame nonconvexity, (Jin et al., 2019) proves the $\mathcal{O}(\theta^{-4})$ rate in gradient evaluations is required in the convergence to an θ-first order NE with Max-oracle; (Lu et al., 2019a;b) 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 (Sanjabi et al., 2018b) 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 is originally presented in (Goodfellow et al., 2014) without an algorithm to find it. A line of work (Grnarova et al., 2017; Arora et al., 2017; Oliehoek et al., 2018; Hsieh et al., 2018) seeks to further understand and find mixed NEs of GANs. Nonetheless, the existing algorithm in (Hsieh et al., 2018) 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 is studied by Frank-Wolfe (Gaivoronski, 1986) and steepest descent algorithms (Molchanov & Zuyev, 2001) in earlier times. More recently, descent methods in the space of probability measures (Richemond & Maginnis, 2017; Frogner & Poggio, 2018) are getting popular in machine learning, where particle-based methods (Liu et al., 2018; Chen et al., 2018) approximate probability measures for practical implementation. Similarly, two sets of particles in our algorithm also provide the Dirac measure approximation for probability measures.

In addition, with similar settings, (Chizat & Bach, 2018) performs a continuous-time gradient descent on particles' weights and positions. SVGD (Liu, 2017) guarantees to optimally decrease the KL divergence within a function space. For zero-sum games, (Domingo-Enrich et al., 2020) parametrizes mixed strategies as mixtures of particles, whose positions and weights are updated using gradient descent-ascent. More generally, (Lin et al., 2020) aims to solve stochastic mean field games.

Notations. We denote by $[n]$ the set of integers $\{0, 1, ..., n\}$ and by $\mathbb{N}_+$ the set of positive integers. 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 non-negative 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 $d \in \mathbb{N}_+$ are convex compact sets of pure strategies with periodic or zero-flux boundary conditions, and f is the objective function. In nonconvex-nonconcave regimes, as finding local Nash equilibria is NP-hard or even impossible (Jin et al., 2019), 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. Without loss of generality, we set $\mathcal{X}_\mu = \mathcal{X}_\nu = \mathcal{X}$ and write $\mathcal{M} = \mathcal{M}(\mathcal{X})$. Also, $\mathcal{M}(\mathcal{X}) = (\mathcal{P}_2(\mathcal{X}), \mathcal{W}_2)$ is the Wasserstein (W_2-) space, an infinite-dimensional manifold by (Villani, 2008), with the W_2-distance on $\mathcal{P}_2(\mathcal{X})$ defined as

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

where the infimum is taken over the random variables X and Y in $\mathcal{X}$. Here we denote by $\|\cdot\|$ the geodesic distance on $\mathcal{X}$ and by $\mathcal{L}(X)$ the law of a random variable 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 $\mathcal{M}$, we introduce geodesics, tangent vectors and tangent spaces below.

Definition 2.1. Let $\gamma : [0,1] \to \mathcal{P}_2(\mathcal{X})$ be a smooth curve. We call the curve γ a geodesic if there exists a constant $v \geq 0$ such that $\mathcal{W}_2\big(\gamma(t_1) - \gamma(t_2)\big) = v \cdot |t_1 - t_2|$ 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.

The manifold $\mathcal{M}$ is equipped with a weak Riemannian structure in the following sense (Villani, 2008). Given any tangent vectors u, v at $\mu \in \mathcal{M}$ and the vector fields $\widetilde{u}, \widetilde{v}$ of the gradient form 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 $\mathbb{R}^d$. 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 $\gamma(0) = \mu \in \mathcal{M}$ and $\gamma'(0) = u$. 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 $\gamma_u(0) = \mu$ and $\gamma_u'(0) = u$. 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 (Villani, 2003) in the sense that the exponential map is defined on the whole tangent bundle. See §B for more formal definitions.

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

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

where $\mathcal{F}$ is the class of square-integrable functions over $\mathcal{X}$, $F_\mu^*, F_\nu^*\colon \mathcal{F} \to \mathbb{R}$ are strongly convex and smooth functionals *w.r.t.* the ℓ_2-norm. In fact, (2.4) generalize the definition 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

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

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 $\boldsymbol{d}^2(u, v) = \big\langle u - \Gamma_\nu^\mu v, u - \Gamma_\nu^\mu v \big\rangle_\mu$ 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 the Wasserstein space, where the Euclidean distance is replaced by W_2-distance. The following assumption extends the notion of PL condition, also known as gradient domination (Polyak, 1963; Nesterov & Polyak, 2006; Sanjabi et al., 2018b), 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

[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 $x_1, x_2 \in \mathcal{X}$.

for all $\mu \in \mathcal{M}$ we have

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

We call (2.3) a ξ-PL game, or simply a PL game, if $H_\mu(\nu) \triangleq -F(\mu, \nu)$ is ξ-PL *w.r.t.* ν. 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 (Karimi et al., 2016). 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 (Arjovsky et al., 2017) with Kullback-Leibler (KL) divergence regularization,

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

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) (Otto & Villani, 2000) 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, $F_\nu(\mu) = \sup_{f \in \mathcal{F}} \Big\{ -\int \exp\big\{ f(x^\mu) + \mathbb{E}_{x^\nu \sim \nu} \mathbb{E}_{\zeta \sim \mathbb{P}_{x^\mu}} [h_{x^\nu}(\zeta)] \big\}\, \mathrm{d}\mu_0(x^\mu) + \int f d\mu + \widehat{F}_\nu \Big\}$, $F_\mu(\nu) = -\sup_{f \in \mathcal{F}} \Big\{ \int f d\nu - \int \exp\big\{ f(x^\nu) + g(x^\nu) - (D\mu)(x^\nu) \big\}\, \mathrm{d}\mu_0(x^\nu) + \widehat{F}_\mu \Big\}$. Here $\widehat{F}_\nu = 1 - \mathrm{KL}(\nu \| \mu_0) + \mathbb{E}_{x^\nu \sim \nu} \mathbb{E}_{\zeta \sim \mathbb{P}_{\text{real}}} [h_{x^\nu}(\zeta)]$ 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

$$\mathcal{J}_\mu(\mu_1, \nu_1) \triangleq -\min_{\mathcal{W}_2(\mu, \mu_1) \leq 1} \left\langle \operatorname{grad}_\mu F(\mu_1, \nu_1), \operatorname{Exp}_{\mu_1}^{-1}(\mu) \right\rangle_{\mu_1},$$

$$\mathcal{J}_\nu(\mu_1, \nu_1) \triangleq \max_{\mathcal{W}_2(\nu, \nu_1) \leq 1} \left\langle \operatorname{grad}_\nu F(\mu_1, \nu_1), \operatorname{Exp}_{\nu_1}^{-1}(\nu) \right\rangle_{\nu_1}$$

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

$$\mathcal{J}_\mu(\mu^*, \nu^*) \leq \theta \quad \text{and} \quad \mathcal{J}_\nu(\mu^*, \nu^*) \leq \theta. \tag{2.8}$$

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 (Sanjabi et al., 2018b) 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 $\operatorname{grad} F_\nu(\mu)$ along the geodesic. Therefore, in the ideal case, the gradient update is given by

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

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 Gra-

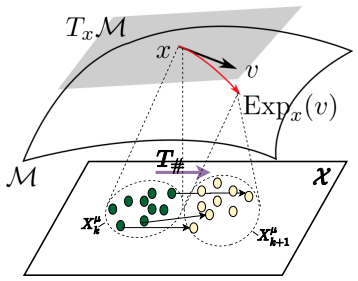

Figure 1. Equivalence between particle pushing in the Euclidean space $\mathcal{X}$ and the exponential map in the Wasserstein space $\mathcal{M}$. The tangent vector $v \in T_x\mathcal{M}$ at x induces the exponential map Exp_x and its correspondence in $\mathcal{X}$, the push-forward map $T_\sharp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

dients). 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

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

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 $\mathrm{grad}\, F_\nu(\mu) = -\,\mathrm{div}(\mu \cdot \nabla f_\mu^*)$.

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 $-\,\mathrm{div}(\mu \cdot \nabla u) = s$ 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

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

where we use $\mathrm{Exp}_{\mathcal{X}}(t \cdot \nabla u)$ to denote the transportation map on $\mathcal{X}$ that sends any $x \in \mathcal{X}$ to a point $\mathrm{Exp}_x(t \cdot \nabla u(x)) \in \mathcal{X}$, 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 $\mathrm{Exp}_\mu[-t \cdot \mathrm{grad}\, F(\mu)] = \big[\mathrm{Exp}_{\mathcal{X}}\big(-t \cdot \nabla f_\mu^*\big)\big]_\sharp \mu$. 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.

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,

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

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 (Csáji et al., 2001; Hofmann et al., 2008).

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

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

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,

$\{\widetilde{\mu}_k\}_{k\geq 0}$ and $\{\widetilde{\nu}_l(\widetilde{\mu}_k)\}_{k,l\geq 0}$ are sequences of probability measures of $\{X_k^\mu\}_{k\geq 0}$ and $\{X_l^\nu(\widetilde{\mu}_k)\}_{k,l\geq 0}$ 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 $k \geq 1$. 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),

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

for all $i \in [N_\mu]$. Here $\eta^\mu > 0$ are the stepsize specified in Theorem 4.2. 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 $x^\mu_{i,k} \in \mathbb{R}^d$.

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

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

for all $i \in [N_\nu]$ in Line 7 of Algorithm 1, with fixed $\widetilde{\mu}_k$. In particular, the sequences of probability measures $\{\widetilde{\mu}_k\}_{k\geq 0}$ and $\{\widetilde{\nu}_l(\widetilde{\mu}_k)\}_{k,l\geq 0}$ 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

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

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 $k \geq 1$ 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, $x^\mu_{i,k} \overset{\text{i.i.d.}}{\sim} \widetilde{\mu}_k$ and $x^\nu_{i,l}(\widetilde{\mu}_k) \overset{\text{i.i.d.}}{\sim} \widetilde{\nu}_l(\widetilde{\mu}_k)$ 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 $k \geq 0$.

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

1: **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 $\eta^\mu \in \big(0, \min\{1/h, 2/\widetilde{L}\}\big)$, $\eta^\nu \in \big(0, \min\{1/(4L_2), 1/h\}\big)$.
2: Initialize N_μ (N_ν) particles $X_0^\mu = \{x^\mu_{i,0}\}_{i\in[N_\mu]}$ $\big(X_0^\nu\big)$ by drawing N_μ (N_ν) i.i.d. samples from $\widetilde{\mu}_0$ ($\widetilde{\nu}_0$).
3: **for** $k = 0, 1, 2, \ldots, K_\mu - 1$ **do**
4: Set $X_0^\nu(\widetilde{\mu}_k) = X_k^\nu$
5: **for** $l = 0, 1, 2, \ldots, K_\nu - 1$ **do**
6: $\widetilde{f}^*_{l,k} \leftarrow \mathbf{VFM}\big(X_l^\nu(\widetilde{\mu}_k), F^*_{\widetilde{\mu}_k}, N_\nu\big)$
7: Push ν-particles: $x^\nu_{i,l+1}(\widetilde{\mu}_k) \leftarrow \mathrm{Exp}_{x^\nu_{i,l}(\widetilde{\mu}_k)}\big[-\eta^\nu\cdot\nabla\widetilde{f}^*_{l,k}\big(x^\nu_{i,l}(\widetilde{\mu}_k)\big)\big]$ for all $i \in [N_\nu]$
8: **end for**
9: Set $X^\nu_{k+1} = X^\nu_{K_\nu}(\widetilde{\mu}_k)$
10: $\widetilde{f}_k^* \leftarrow \mathbf{VFM}\big(X_k^\mu, F^*_{\widetilde{\nu}_{k+1}}, N_\mu\big)$
11: Push μ-particles: $x^\mu_{i,k+1} \leftarrow \mathrm{Exp}_{x^\mu_{i,k}}[-\eta^\mu\cdot\nabla\widetilde{f}_k^*(x^\mu_{i,k})]$ for all $i \in [N_\mu]$
12: Set $X^\mu_{k+1} = \{x^\mu_{i,k+1}\}_{i\in[N_\mu]}$
13: **end for**
14: **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 $\mathrm{grad}\, F(\widetilde{\mu}_k) = -\mathrm{div}(\widetilde{\mu}_k\cdot\nabla f_k^*)$. However, with only finite samples, we obtain an estimator $\widetilde{f}_k^*$ of f_k^*. Hence, the gradient estimate at $\widetilde{\mu}_k$ is $-\mathrm{div}(\widetilde{\mu}_k\cdot\nabla\widetilde{f}_k^*)$, and the difference between $\mathrm{grad}\, F(\widetilde{\mu}_k)$ and its estimate is denoted by $\delta_k = -\mathrm{div}[\widetilde{\mu}_k\cdot(\nabla\widetilde{f}_k^* - \nabla f_k^*)]$. 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. Moreover, based on Assumption 4.1 we identify a new type of reverse Poincaré inequality (Baudoin & Bonnefont, 2016) in §D.1. Due to the use of fundamental topology and analysis property of $\mathcal{X}$ and $\widetilde{\mathcal{F}}$, our analysis can also 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 D.1, 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

$$\bar{\varepsilon}_\mu = \mathcal{O}(N_\mu^{-1/2}), \quad \bar{\varepsilon}_\nu = \mathcal{O}(N_\nu^{-1/2}) \tag{4.2}$$

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 4.2 (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

$$K_\nu \geq K_\nu(\theta) = \mathcal{O}\Big(\log\frac{(1-\sigma)\widehat{M}_H - \eta^\nu\bar{\varepsilon}_\nu}{\theta}\Big/\log\frac{1}{\sigma}\Big), \quad \text{where } \widehat{M}_H = \max\Big\{M_H, \frac{\eta^\nu\bar{\varepsilon}_\nu + 1}{1-\sigma}\Big\}, \tag{4.3}$$

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

$$\mathbb{E}_{X_0}\big[\mathcal{J}_\mu^2(\widetilde{\mu}_k, \widetilde{\nu}_{k+1})\big] = \mathcal{O}\Big(\big(\Delta + \sqrt{\bar{\varepsilon}_\mu}\big)^2 \cdot \Big(\big(\Delta + \sqrt{\bar{\varepsilon}_\mu}\big) + \frac{M_G}{K_\mu}\Big)\Big),$$
$$\mathbb{E}_{X_0}\big[\mathcal{J}_\nu(\widetilde{\mu}_k, \widetilde{\nu}_{k+1})\big] = \mathcal{O}\Big(\frac{L_2\Delta}{L_0}\Big). \tag{4.4}$$

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 (4.2).

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

The proof of Theorem 4.2 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 (4.2) 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 (2.8) 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 4.3 (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 4.3 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 4.4 (Convergence to the Minimax Value of Two-Sided PL Games)**.** Let the objective F satisfy (2.4), Assumption 2.2 and 4.3. 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

$$\begin{aligned}\mathbb{E}_{X_0}\Big[F\big(\widetilde{\mu}_k, \nu^*(\widetilde{\mu}_k)\big)\Big] - F(\mu^*, \nu^*) \\ \leq \underbrace{\widetilde{\sigma}^k \cdot \Big(\mathbb{E}_{X_0}\big[F(\widetilde{\mu}_0, \widetilde{\nu}_1)\big] - F(\mu^*, \nu^*)\Big)}_{\text{(i)}} \\ + \underbrace{\frac{1-\widetilde{\sigma}^k}{1-\widetilde{\sigma}} \cdot \eta^\mu\big(\bar{\varepsilon}_\mu + \widetilde{\Delta}^2\big)}_{\text{(ii)}}, \end{aligned} \tag{4.5}$$

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 (4.2). 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 4.4 differs from that of Theorem 4.2 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 4.3, Theorem 4.4 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 $\bar{\varepsilon}_\mu$ 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 $\bar{\varepsilon}_\nu$ of player ν scaled by $(1-\sigma^{K_\nu})/(1-\sigma)$. Specifically, in the total bound (4.5) $\bar{\epsilon}_\mu$ 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 4.2 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.

5. 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 (Hsieh et al., 2018) 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, and the sampling procedure is faster than the SGLD-based methd.

6. Conclusion

In this paper, we lift finite-dimensional zero-sum games to infinite-dimensional distributional zero-sum games over a space of probability measures, in order to find mixed NEs for finite-dimensional games. We then propose a particle-based variational transport algorithm in the functional space to solve such games, by analogy with the gradient descent-ascent algorithm in finite-dimensional spaces. Furthermore, we provide the first complete statistical and convergence guarantees for such particle-based method. Our analysis applies to problems with different assumptions on non-convexity (PL games and two-sided PL games). Toy experiments show promising empirical results.

Acknowledgments

We thank Baptiste Goujaud and Damien Scieur for enlightening discussion. Zhaoran Wang acknowledges National Science Foundation (Awards 2048075, 2008827, 2015568, 1934931), Simons Institute (Theory of Reinforcement Learning), Amazon, J.P. Morgan, and Two Sigma for their sup-

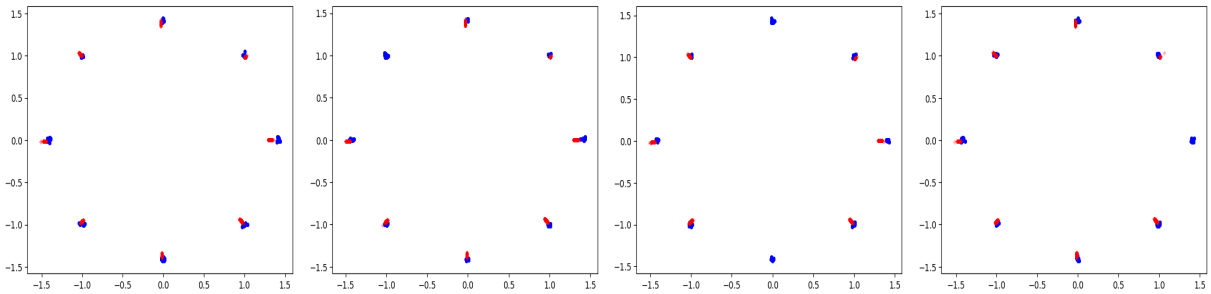

Figure 2. Toy experiment results: blue dots represent samples from the gaussian mixture and the red dots represent the samples generated from generators.

ports. Zhuoran Yang acknowledges Simons Institute (Theory of Reinforcement Learning).

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. Generalization and equilibrium in generative adversarial nets (GANs). In *International Conference on Machine Learning*, pp. 224–232. JMLR. org, 2017.

Baudoin, F. and Bonnefont, M. Reverse Poincaré inequalities, isoperimetry, and riesz transforms in carnot groups. *Nonlinear Analysis*, 131:48–59, 2016.

Bernhard, P. and Rapaport, A. On a theorem of Danskin with an application to a theorem of Von Neumann-Sion. *Nonlinear Analysis: Theory, Methods & Applications*, 24 (8):1163–1181, 1995.

Burago, D., Ivanov, S., and Burago, Y. Course in metric geometry. 2001.

Carlen, E. A. and Gangbo, W. Constrained steepest descent in the 2-Wasserstein metric. *Annals of Mathematics*, pp. 807–846, 2003.

Chambolle, A. and Pock, T. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1-2):253–287, 2016.

Chen, C., Zhang, R., Wang, W., Li, B., and Chen, L. A unified particle-optimization framework for scalable Bayesian sampling. *arXiv preprint arXiv:1805.11659*, 2018.

Chern, S.-S., Chen, W.-h., and Lam, K. S. *Lectures on differential geometry*, volume 1. World Scientific Publishing Company, 1999.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. *arXiv preprint arXiv:1805.09545*, 2018.

Cotter, N. E. The Stone-Weierstrass theorem and its application to neural networks. *IEEE Transactions on Neural Networks*, 1(4):290–295, 1990.

Csáji, B. C. et al. Approximation with artificial neural networks. *Faculty of Sciences, Etvs Lornd University, Hungary*, 24(48):7, 2001.

Dai, B., Shaw, A., Li, L., Xiao, L., He, N., Liu, Z., Chen, J., and Song, L. Sbeed: Convergent reinforcement learning with nonlinear function approximation. *arXiv preprint arXiv:1712.10285*, 2017.

Dai, B., Dai, H., Gretton, A., Song, L., Schuurmans, D., and He, N. Kernel exponential family estimation via doubly dual embedding. *arXiv preprint arXiv:1811.02228*, 2018.

Dang, C. D. and Lan, G. On the convergence properties of non-euclidean extragradient methods for variational inequalities with generalized monotone operators. *Computational Optimization and Applications*, 60(2):277–310, 2015.

Daskalakis, C. and Panageas, I. Last-iterate convergence: Zero-sum games and constrained min-max optimization. *arXiv preprint arXiv:1807.04252*, 2018a.

Daskalakis, C. and Panageas, I. The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in Neural Information Processing Systems*, pp. 9236–9246, 2018b.

Domingo-Enrich, C., Jelassi, S., Mensch, A., Rotskoff, G., and Bruna, J. A mean-field analysis of two-player zero-sum games. *arXiv preprint arXiv:2002.06277*, 2020.

Evans, L. *Partial Differential Equations*. American Mathematical Society, 2010.

Facchinei, F. and Pang, J.-S. *Finite-dimensional variational inequalities and complementarity problems*. Springer Science & Business Media, 2007.

Frogner, C. and Poggio, T. Approximate inference with Wasserstein gradient flows. *arXiv preprint arXiv:1806.04542*, 2018.

Gaivoronski, A. Linearization methods for optimization of functionals which depend on probability measures. In *Stochastic Programming 84 Part II*, pp. 157–181. Springer, 1986.

Ghosh, S., Squillante, M., and Wollega, E. Efficient stochastic gradient descent for distributionally robust learning. *arXiv preprint arXiv:1805.08728*, 2018.

Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*, 2018a.

Gidel, G., Hemmat, R. A., Pezeshki, M., Lepriol, R., Huang, G., Lacoste-Julien, S., and Mitliagkas, I. Negative momentum for improved game dynamics. *arXiv preprint arXiv:1807.04740*, 2018b.

Gigli, N. On the inverse implication of Brenier-McCann theorems and the structure of $(p_2(\mathcal{M}), w_2)$. *Methods and Applications of Analysis*, 18(2):127–158, 2011.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

Grnarova, P., Levy, K. Y., Lucchi, A., Hofmann, T., and Krause, A. An online learning approach to generative adversarial networks. *arXiv preprint arXiv:1706.03269*, 2017.

Hamedani, E. Y., Jalilzadeh, A., Aybat, N., and Shanbhag, U. Iteration complexity of randomized primal-dual methods for convex-concave saddle point problems. *arXiv preprint arXiv:1806.04118*, 2018.

Ho, J. and Ermon, S. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, pp. 4565–4573, 2016.

Hofmann, T., Schölkopf, B., and Smola, A. J. Kernel methods in machine learning. *Annals of Statistics*, pp. 1171–1220, 2008.

Hsieh, Y.-P., Liu, C., and Cevher, V. Finding mixed nash equilibria of generative adversarial networks. *arXiv preprint arXiv:1811.02002*, 2018.

Jin, C., Netrapalli, P., and Jordan, M. I. Minmax optimization: Stable limit points of gradient descent ascent are locally optimal. *arXiv preprint arXiv:1902.00618*, 2019.

Juditsky, A. and Nemirovski, A. Solving variational inequalities with monotone operators on domains given by linear minimization oracles. *Mathematical Programming*, 156 (1-2):221–256, 2016.

Karimi, H., Nutini, J., and Schmidt, M. 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*, pp. 795–811. Springer, 2016.

Lee, J. M. *Introduction to smooth manifolds*. Springer, 2001.

Lin, A. T., Fung, S. W., Li, W., Nurbekyan, L., and Osher, S. J. Alternating the population and control neural networks to solve high-dimensional stochastic mean-field games. *arXiv preprint arXiv:2002.10113*, 2020.

Liu, C., Zhuo, J., Cheng, P., Zhang, R., Zhu, J., and Carin, L. Understanding and accelerating particle-based variational inference. *arXiv preprint arXiv:1807.01750*, 2018.

Liu, Q. Stein variational gradient descent as gradient flow. *arXiv preprint arXiv:1704.07520*, 2017.

Liu, Y., Shang, F., Cheng, J., Cheng, H., and Jiao, L. Accelerated first-order methods for geodesically convex optimization on Riemannian manifolds. In *Advances in Neural Information Processing Systems*, pp. 4868–4877, 2017.

Lu, S., Tsaknakis, I., and Hong, M. Block alternating optimization for non-convex min-max problems: algorithms and applications in signal processing and communications. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4754–4758. IEEE, 2019a.

Lu, S., Tsaknakis, I., Hong, M., and Chen, Y. Hybrid block successive approximation for one-sided non-convex min-max problems: algorithms and applications. *arXiv preprint arXiv:1902.08294*, 2019b.

Mai, T., Mihail, M., Panageas, I., Ratcliff, W., Vazirani, V., and Yunker, P. Cycles in zero-sum differential games and biological diversity. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 339–350, 2018.

Mescheder, L., Geiger, A., and Nowozin, S. Which training methods for GANs do actually converge? *arXiv preprint arXiv:1801.04406*, 2018.

Molchanov, I. and Zuyev, S. Variational calculus in the space of measures and optimal design. In *Optimum design 2000*, pp. 79–90. Springer, 2001.

Monteiro, R. D. and Svaiter, B. F. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM Journal on Optimization*, 20(6): 2755–2787, 2010.

Nemirovski, A. 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(1): 229–251, 2004.

Nesterov, Y. and Polyak, B. T. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Nowozin, S., Cseke, B., and Tomioka, R. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pp. 271–279, 2016.

Oliehoek, F. A., Savani, R., Gallego, J., van der Pol, E., and Groß, R. Beyond local nash equilibria for adversarial networks. In *Benelux Conference on Artificial Intelligence*, pp. 73–89. Springer, 2018.

Otto, F. and Villani, C. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

Petersen, P., Axler, S., and Ribet, K. *Riemannian geometry*, volume 171. Springer, 2006.

Polyak, B. T. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

Raginsky, M., Rakhlin, A., and Telgarsky, M. Nonconvex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. *arXiv preprint arXiv:1702.03849*, 2017.

Richemond, P. H. and Maginnis, B. On Wasserstein reinforcement learning and the Fokker-Planck equation. *arXiv preprint arXiv:1712.07185*, 2017.

Rockafellar, R. T. *Convex analysis*. Number 28. Princeton university press, 1970.

Rudin, W. Principles of mathematical analysis. 1976.

Sanjabi, M., Ba, J., Razaviyayn, M., and Lee, J. D. On the convergence and robustness of training GANs with regularized optimal transport. In *Advances in Neural Information Processing Systems*, pp. 7091–7101, 2018a.

Sanjabi, M., Razaviyayn, M., and Lee, J. D. Solving non-convex non-concave min-max games under Polyak-Łojasiewicz condition. *arXiv preprint arXiv:1812.02878*, 2018b.

Sinha, A., Namkoong, H., and Duchi, J. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

Stewart, J. Chapter 15.2 limits and continuity, multivariable calculus, 2008.

Thekumparampil, K. K., Jain, P., Netrapalli, P., and Oh, S. Efficient algorithms for smooth minimax optimization. In *Advances in Neural Information Processing Systems*, pp. 12659–12670, 2019.

Van Parys, B. P., Esfahani, P. M., and Kuhn, D. From data to decisions: Distributionally robust optimization is optimal. *arXiv preprint arXiv:1704.04118*, 2017.

Villani, C. *Topics in optimal transportation*. Number 58. American Mathematical Society, 2003.

Villani, C. *Optimal transport: old and new*, volume 338. Springer, 2008.

Wibisono, A. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. *arXiv preprint arXiv:1802.08089*, 2018.

Zhang, H. and Sra, S. An estimate sequence for geodesically convex optimization. In *Conference On Learning Theory*, pp. 1703–1723, 2018.

Zhang, H., Reddi, S. J., and Sra, S. Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. In *Advances in Neural Information Processing Systems*, pp. 4592–4600, 2016.

Zou, D., Xu, P., and Gu, Q. Subsampled stochastic variance-reduced gradient Langevin dynamics. In *International Conference on Uncertainty in Artificial Intelligence*, 2018.