Deep Generalized Schrödinger Bridge

Anonymous Author(s)
Affiliation
Address
email

Abstract

Mean-Field Game (MFG) serves as a crucial mathematical framework in modeling the collective behavior of individual agents interacting stochastically with a large population. In this work, we aim at solving a challenging class of MFGs in which the differentiability of these interacting preferences needs *not* available to the solver, and the population is urged to *converge exactly* to some desired distribution. These setups are, despite being well-motivated for practical purposes, complicated enough to paralyze most (deep) numerical solvers. Nevertheless, we show that Schrödinger Bridge — as an entropy-regularized optimal transport model — can be generalized to accepting mean-field structures, hence solving these MFGs. This is achieved via the application of Forward-Backward Stochastic Differential Equations theory, which, intriguingly, leads to a computational framework with a similar structure to Temporal Difference learning. As such, it opens up novel algorithmic connections to Deep Reinforcement Learning that we leverage for facilitating practical training. Our method, named Deep Generalized Schrödinger Bridge (**DeepGSB**), not only outperforms prior methods in solving classical population navigation MFGs, but is also capable of solving 1000-dimensional *opinion depolarization*, setting a new state-of-the-art numerical solver for high-dimensional MFGs.

1 Introduction

In a scorching morning, you navigated through the crowds toward the office. As you walked through a crosswalk, you were pondering the growing public opinion on a new policy over the past week, and were suddenly interrupted by the honking as the traffic started moving...

From navigation in crowds to propagation of opinion and traffic movement, examples of *individual agents interacting with a large population* are widespread in daily life and, due to their prevalence, appear as an important subject in multidisciplinary scientific areas, including economics [1, 2], opinion modeling [3–5], robotics [6, 7], and more recently machine learning [8–10].

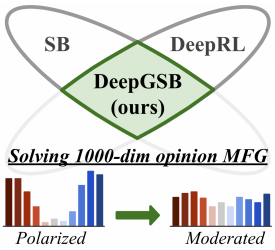

Figure 1: **DeepGSB** paves a new connection between Schrödinger Bridge (SB) and Deep Reinforcement Learning for solving high-dimensional MFGs.

Mathematically, the decision-making processes under these scenarios can be characterized by the **Mean-Field Game** [11–13] **(MFG)**, which models a noncooperative differential game on a finite horizon between a continuum population of rational agents. Let $u(x, t)$ be the *value* function, also known as optimal cost-to-goal, that governs agents' policies at each state $x \in \mathbb{R}^d$ and time $t \in [0, T]$, and denote the resulting population density by $\rho(\cdot, t) \in \mathcal{P}(\mathbb{R}^d)$, where $\mathcal{P}(\mathbb{R}^d)$ is the set of probability measures on $\mathbb{R}^d$. At the Nash equilibrium where no agent has the incentive to change his/her decision, MFG, at its

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Table 1: Comparison to existing methods w.r.t. various desired features in Mean-Field Games (MFGs). Our **DeepGSB** is capable of solving a much wider class of MFGs in higher dimensional state spaces.

	continuous state space	stochastic MF dyn. (2)	converges to exact ρ_{target}	discontinuous MF interaction F	highest dimension
Ruthotto et al. [14]	✓	✗	✗	✗	100
Lin et al. [15]	✓	✗	✗	✗	100
Chen [16]	✗	✓	✓	✗[1]	2
DeepGSB (ours)	✓	✓	✓	✓	**1000**

most general form, solves the following partial differential equations (PDEs):

$$\begin{cases} -\frac{\partial u(x,t)}{\partial t} + H(x, \nabla u, \rho) - \frac{1}{2}\sigma^2 \Delta u = F(x, \rho), & u(x, T) = G(x, \rho(x, T)) \\ \frac{\partial \rho(x,t)}{\partial t} - \nabla \cdot (\rho \, \nabla_p H(x, \nabla u, \rho) - \frac{1}{2}\sigma^2 \Delta \rho = 0, & \rho(x, 0) = \rho_0(x) \end{cases}, \tag{1}$$

where ∇, $\nabla\cdot$, and Δ are respectively the gradient, divergence, and laplacian operators.[2] These two PDEs are respectively known as the Hamilton-Jacobi-Bellman (HJB) and Fokker-Plank (FP) equations, which characterize the evolution of $u(x, t)$ and $\rho(x, t)$. They are coupled with each other through the Hamiltonian $H(x, p, \rho) : \mathbb{R}^d \times \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$, which describes the dynamics of the game, and the mean-field interaction $F(x, \rho) : \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$, which quantifies the agent's preference when interacting with the population. The terminal condition G typically penalizes deviations from some desired target distribution ρ_{target}, *e.g.,* $G \approx D_{\text{KL}}(\rho(\cdot, T)||\rho_{\text{target}}(\cdot))$. Given a solution (u, ρ) to (1), each agent acts accordingly and follows a stochastic differential equation(SDE)

$$\mathrm{d}X_t = -\nabla_p H(X_t, \nabla u(X_t, t), \rho(\cdot, t))\mathrm{d}t + \sigma \mathrm{d}W_t, \quad X_0 \sim \rho_0, \tag{2}$$

where $W_t \in \mathbb{R}^d$ is the Wiener process and $\sigma \in \mathbb{R}$ is some diffusion scaler. At the mean-field limit, *i.e.,* when the number of agents goes to infinity, the collective behavior of (2) yields the density $\rho(\cdot, t)$.

Numerical methods for solving (1) have advanced rapidly with the aid of machine learning. Seminar works such as Ruthotto et al. [14] and Lin et al. [15] approximated (u, ρ) with deep neural networks (DNNs) and directly penalized the violation of PDEs. Despite showing preliminary successes, the underlying dynamics (2) were either degenerate, *e.g.,* $\sigma := 0$ or $H := H(x, \nabla u)$, or completely discarded by instead regressing network outputs on the entire state space, which can scale unfavorably as the dimension d grows. An alternative that avoids both limitations, *i.e.,* it keeps the full stochastic dynamics in (2) while being computationally scalable, is to recast these PDEs to a set of forward-backward SDEs (FBSDEs) by applying the nonlinear Feynman-Kac Lemma [17–19]. The FBSDEs analysis appears extensively in the theoretical study of MFG [20–23], yet development of scalable FBSDEs-based solver has remained, surprisingly, limited. Our work contributes to this direction.

Since ρ_{target} is known in prior, in many cases there are direct interests to seek an optimal policy that guides the agents from an initial distribution ρ_0 to the *exact* ρ_{target}, while respecting the structure of MFG, particularly the MF interaction $F(x, \rho)$. Lifting (1) to this setup, however, is highly nontrivial. Indeed, replacing the *soft* penalty at $u(x, T) = D_{\text{KL}}(\rho||\rho_{\text{target}})$ with a *hard* distributional constraint at $\rho(x, T) = \rho_{\text{target}}$ yields an HJB whose boundary condition can only be defined implicitly through FP, which now contains two distributional constraints and resembles an optimal transport problem. As such, despite being well-motivated, most prior methods have struggled to extend to this setup.

In this work, we show that an emerging machine learning model, called **Schrödinger Bridge** (**SB**) [24–28], provides an elegant recipe for solving this challenging class of *MFGs with distributional boundary constraints* (ρ_0, ρ_{target}). Although SB is traditionally set up as an entropy-regularized optimal transport model [29–31] with $F := 0$, we show that SB-FBSDE [27], an FBSDE-based method for solving SB, can be generalized to accept nontrivial F; hence solving MFG. Interestingly, the new FBSDEs system admits a similar computational structure to temporal difference (TD) learning, leading to a framework that narrows the gap between SB and Deep Reinforcement Learning (DeepRL); see Fig. 1. This connection enables our method to take advantage of DeepRL techniques,

[1] Precisely, Chen [16] considered discontinuous yet non-MF interaction, $F{:=}F(x)$, on a *discrete* state space.

[2] These operators are taken w.r.t. x unless otherwise noted. See Appendix A.1 for the notational summary.

such as target networks, replay buffer, actor-critic, *etc*, and, more importantly, to handle a wide class of MF interactions that need *not* require to be continuous *nor* differentiable. This is in contrast to most existing works, which require differentiable [14, 15] or quadratic [32] structure on F, or discretize the state space [16]. We validate our method, called **Deep Generalized Schrödinger Bridge (DeepGSB)**, on various challenging MFGs from crowd navigation to *high-dimensional opinion depolarization* (where $d = 1000$), setting a state-of-the-art record in the area of numerical MFG solvers.

In summary, we present the following contributions.

- We present a novel numerical method, rooted in Schrödinger Bridge (SB), for solving a challenging class of Mean-Field Game where the population needs to converge *exactly* to the target distribution.
- The resulting method, **DeepGSB**, generalizes prior SB results to accepting flexible mean-field interaction (*e.g.,* non-differentiable) and enjoys modern training techniques from DeepRL.
- **DeepGSB** achieves promising empirical results in navigating crowd motion and depolarizing 1000-dimensional opinion dynamics, setting a new state-of-the-art numerical MFG solver.

2 Preliminary on Schrödinger Bridge (SB)

The SB problem was originally introduced in the 1930s for quantum mechanics [29, 33] and later draws broader interests with its connection to optimal transport and control [34–37]. Given a pair of boundary distributions (ρ_0, ρ_T), SB seeks an optimal pair of stochastic processes of the forms:

$$\mathrm{d}X_t = [f(X_t, t) + \sigma^2 \nabla \log \Psi(X_t, t)]\mathrm{d}t + \sigma \, \mathrm{d}W_t, \quad X_0 \sim \rho_0, \tag{3a}$$

$$\mathrm{d}\bar{X}_s = [-f(\bar{X}_s, s) + \sigma^2 \nabla \log \widehat{\Psi}(\bar{X}_s, s)]\mathrm{d}s + \sigma \, \mathrm{d}W_s, \quad \bar{X}_0 \sim \rho_T. \tag{3b}$$

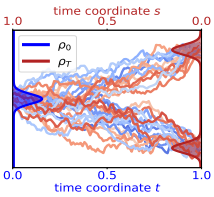

Figure 2: Simulation of the forward (3a) and backward (3b) SDEs in SB, which are minimum-energy solution when $(\Psi, \widehat{\Psi})$ obey the PDEs in (4).

While X_t is a standard stochastic process starting from ρ_0, $\bar{X}_s$ evolves along the "*reversed*" time coordinate $s := T - t$ from ρ_T. The base drift f and diffusion σ are typically known in prior and related to the Hamiltonian H. Suppose $\Psi, \widehat{\Psi} \in C^{2,1}(\mathbb{R}^d, [0, T])$ solve the following coupled PDEs,

$$\begin{cases} \frac{\partial \Psi(x,t)}{\partial t} = -\nabla \Psi^\top f - \frac{1}{2}\sigma^2 \Delta \Psi \\ \frac{\partial \widehat{\Psi}(x,t)}{\partial t} = -\nabla \cdot (\widehat{\Psi} f) + \frac{1}{2}\sigma^2 \Delta \widehat{\Psi} \end{cases} \quad \text{s.t.} \quad \begin{matrix} \Psi(\cdot, 0)\widehat{\Psi}(\cdot, 0) = \rho_0 \\ \Psi(\cdot, T)\widehat{\Psi}(\cdot, T) = \rho_T \end{matrix}, \tag{4}$$

then the theory of SB suggests that the SDEs in (3) are "*optimal*" in the sense that they solve a variational optimization w.r.t. minimum control (*i.e.,* entropy-regularized). Furthermore, the path-wise measure induced by (3a) along $t \in [0, T]$ is equal almost surely to the path-wise measure induced by (3b) along $s := T - t$. In other words, the two SDEs in (3) can be thought of as the "*reversed*" process to each other; and hence we also have $X_T \sim \rho_T$ and $\bar{X}_T \sim \rho_0$ (see Fig. 2).

Due to the coupling constraints at the boundaries, solving (4) is no easier than solving (1). Fortunately, recent advances [27, 28] have demonstrated a computationally scalable numerical method via the application of the nonlinear Feynman-Kac (FK) Lemma — a mathematical tool that recasts certain classes of PDEs into sets of forward-backward SDEs (FBSDEs) via some transformation. These *nonlinear FK transformations* are parametrized in SB-FBSDE [27] by some DNNs with θ and ϕ, *i.e.,*

$$Z_\theta(\cdot, \cdot) \approx \sigma \nabla \log \Psi(\cdot, \cdot) \quad \text{and} \quad \widehat{Z}_\phi(\cdot, \cdot) \approx \sigma \nabla \log \widehat{\Psi}(\cdot, \cdot), \tag{5}$$

and the FBSDEs resulting from (4) and (5) yield the following objectives:

$$\mathcal{L}_{\text{IPF}}(\theta) = \int_0^T \mathbb{E}_{(3b)} \left[\frac{1}{2} \| Z_\theta(\bar{X}_s, s) \|_2^2 + Z_\theta(\bar{X}_s, s)^\top \widehat{Z}_\phi(\bar{X}_s, s) + \nabla \cdot (\sigma Z_\theta(\bar{X}_s, s) + f) \right] \mathrm{d}s, \tag{6a}$$

$$\mathcal{L}_{\text{IPF}}(\phi) = \int_0^T \mathbb{E}_{(3a)} \left[\frac{1}{2} \| \widehat{Z}_\phi(X_t, t) \|_2^2 + \widehat{Z}_\phi(X_t, t)^\top Z_\theta(X_t, t) + \nabla \cdot (\sigma \widehat{Z}_\phi(X_t, t) - f) \right] \mathrm{d}t. \tag{6b}$$

These objectives are alternatively minimized in the same manner as how Iterative Proportional Fitting [38] (IPF) algorithms solve parametrized SBs [24, 25]. Once Z_θ and $\widehat{Z}_\phi$ are learned, the SDEs of SB in (3) can be simulated by substituting (5); see Fig. 2. We refer readers to Appendix A.2 for a detailed review of nonlinear FK Lemma, SB-FBSDE [27], and the derivation of $\mathcal{L}_{\text{IPF}}$ in (6).

3 Deep Generalized Schrödinger Bridge (DeepGSB)

3.1 Connection between the coupled PDEs in MFG and SB

We begin by stating our problem of interest — MFG with hard distributional constraints $(\rho_0, \rho_{\text{target}})$ — in its mathematical form. Similar to prior works [14, 15], we will adopt the control-affine Hamiltonian, $H(x, \nabla u, \rho) := \frac{1}{2}\|\sigma \nabla u\|^2 - \nabla u^\top f(x, \rho)$, given some base drift f. With this, the MFG (1) becomes

$$\begin{cases} -\frac{\partial u(x,t)}{\partial t} + \frac{1}{2}\|\sigma\nabla u\|^2 - \nabla u^\top f - \frac{1}{2}\sigma^2 \Delta u = F(x, \rho), \\ \frac{\partial \rho(x,t)}{\partial t} - \nabla \cdot (\rho\,(\sigma^2 \nabla u - f) - \frac{1}{2}\sigma^2 \Delta \rho = 0, \quad \rho(x, 0) = \rho_0(x),\ \rho(x, T) = \rho_{\text{target}}(x), \end{cases} \tag{7}$$

which, as we briefly discussed in Sec.1, differs from (1) in that the boundary condition of the HJB, $u(x, T)$, is now absorbed into FP and defined implicitly through $\rho(x, T) = \rho_{\text{target}}(x)$. Since analytic conversion between the boundary conditions of (1) and (7) exists only for highly degenerate[3] cases [39], this seemingly innocuous change suffices to paralyze most prior methods. Nevertheless, as (7) now describes a transformation between two distributions (from FP) while obeying some optimality (from HJB), it suggests a deeper connection to optimal transport, and hence the SB.

To bridge these new MFG PDEs (7) to the PDEs appearing in SB (4), we follow standard treatment [30] and apply the Hopf-Cole transform [40, 41]:

$$\begin{aligned} \Psi(x, t) &:= \exp\left(-u(x, t)\right), \\ \widehat{\Psi}(x, t) &:= \rho(x, t) \exp\left(u(x, t)\right), \end{aligned} \tag{8}$$

which, after some algebra (see Appendix A.3.1 for details), yields the following PDEs:

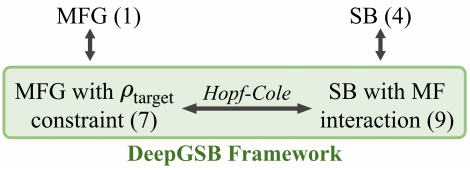

Figure 3: Connection between different coupled PDEs appearing in MFG, SB, and DeepGSB.

$$\begin{cases} \frac{\partial \Psi(x,t)}{\partial t} = -\nabla \Psi^\top f - \frac{1}{2}\sigma^2 \Delta \Psi + F\Psi \\ \frac{\partial \widehat{\Psi}(x,t)}{\partial t} = -\nabla \cdot (\widehat{\Psi} f) + \frac{1}{2}\sigma^2 \Delta \widehat{\Psi} - F\widehat{\Psi} \end{cases} \text{s.t.} \quad \begin{aligned} &\Psi(\cdot, 0)\widehat{\Psi}(\cdot, 0) = \rho_0 \\ &\Psi(\cdot, T)\widehat{\Psi}(\cdot, T) = \rho_{\text{target}} \end{aligned} . \tag{9}$$

It can be readily seen that (9) generalizes (4) by introducing the MF interaction F. Notice that the function $F := F(x, \Psi\widehat{\Psi})$ shall take the product of Ψ and $\widehat{\Psi}$ as its second argument, as the Hopf-Cole transform (8) implies $\Psi(\cdot, t)\widehat{\Psi}(\cdot, t) = \rho(\cdot, t)$. Let $(\Psi, \widehat{\Psi})$ be the solution to these new MF-extended PDEs in (9), and recall the control-affine Hamiltonian we adopted in (7), then one can validate that

$$-\nabla_p H(X_t, \nabla u, \rho) = f - \sigma^2 \nabla u = f + \sigma^2 \nabla \log \Psi.$$

That is, the agent's dynamic (2) coincides with the forward SDE (3a) in SB. Hence, we have connected the MFG (1) and SB (4) frameworks through the PDEs in (7) and (9); see Fig. 3.

3.2 Generalized SB-FBSDEs with mean-field interaction

With (9), we are ready to present our first result that generalizes prior FBSDE for SB to MF interaction. The derivation is standard but somewhat convoluted; we refer readers to Appendix A.3.2 for details.

Theorem 1 (Generalized SB-FBSDEs)**.** *Suppose $\Psi, \widehat{\Psi} \in C^{2,1}$ and let f, F satisfy usual growth and Lipchitz conditions* [42, 43]. *Consider the following nonlinear FK transformations applied to (9):*

$$\begin{aligned} Y_t &\equiv Y(X_t, t) = \log \Psi(X_t, t), \qquad & Z_t &\equiv Z(X_t, t) = \sigma\, \nabla \log \Psi(X_t, t), \\ \widehat{Y}_t &\equiv \widehat{Y}(X_t, t) = \log \widehat{\Psi}(X_t, t), \qquad & \widehat{Z}_t &\equiv \widehat{Z}(X_t, t) = \sigma\, \nabla \log \widehat{\Psi}(X_t, t), \end{aligned} \tag{10}$$

where X_t follows (3a) with $X_0 \sim \rho_0$. Then, the resulting FBSDEs system takes the form:

$$\text{FBSDEs w.r.t. (3a)} : \begin{cases} \mathrm{d}X_t = (f_t + \sigma Z_t)\,\mathrm{d}t + \sigma \mathrm{d}W_t & \text{(11a)} \\ \mathrm{d}Y_t = \left(\frac{1}{2}\|Z_t\|^2 + F_t\right)\mathrm{d}t + Z_t^\top \mathrm{d}W_t & \text{(11b)} \\ \mathrm{d}\widehat{Y}_t = \left(\frac{1}{2}\|\widehat{Z}_t\|^2 + \nabla \cdot (\sigma \widehat{Z}_t - f_t) + \widehat{Z}_t^\top Z_t - F_t\right)\mathrm{d}t + \widehat{Z}_t^\top \mathrm{d}W_t & \text{(11c)} \end{cases}$$

[3] Zhang and Chen [39] suggested $F := 0$, $f := f(x)$, and ρ_0 degenerate to some Dirac delta distribution.

Now, consider a similar transformation in (9) but instead w.r.t. the "reversed" SDE $\bar{X}_s \sim$ (3b) and $\bar{X}_0 \sim \rho_{target}$, i.e., $Y_s \equiv Y(\bar{X}_s, s) = \log \Psi(\bar{X}_s, s)$, *and etc. The resulting FBSDEs system reads*

$$\text{FBSDEs w.r.t. (3b)} : \begin{cases} \mathrm{d}\bar{X}_s = \left(-f_s + \sigma \widehat{Z}_s\right) \mathrm{d}s + \sigma \mathrm{d}W_s & \text{(12a)} \\ \mathrm{d}Y_s = \left(\frac{1}{2}\|Z_s\|^2 + \nabla \cdot (\sigma Z_s + f_s) + Z_s^\top \widehat{Z}_s - F_s\right) \mathrm{d}s + Z_s^\top \mathrm{d}W_s & \text{(12b)} \\ \mathrm{d}\widehat{Y}_s = \left(\frac{1}{2}\|\widehat{Z}_s\|^2 + F_s\right) \mathrm{d}s + \widehat{Z}_s^\top \mathrm{d}W_s & \text{(12c)} \end{cases}$$

Since $Y + \widehat{Y} = \log \rho$ by construction, the functions f and F in (11,12) take the arguments

$$f := f(\cdot, \exp(Y + \widehat{Y})) \quad \textit{and} \quad F := F(\cdot, \exp(Y + \widehat{Y})).$$

Just like how (9) generalizes (4), our results in Theorem 1 also generalize the ones appearing in vanilla SB-FBSDE [27] (see (22) in Appendix A.2) by introducing nontrivial MF interaction F. Despite seemingly complex compared to the original PDEs (9), these FBSDEs systems — namely (11) and (12) — stand as the foundation for developing scalable numerical methods, as they describe precisely how the values of $Y \equiv \log \Psi$ and $\widehat{Y} \equiv \log \widehat{\Psi}$ shall change along the optimal SDEs (notice, *e.g.,* that both Y_t and Z_t are functions of X_t from (10)). Essentially, the nonlinear FK Lemma provides a stochastic representation (in terms of Y and $\widehat{Y}$) of the PDEs in (9) by expanding them w.r.t. the optimal SDEs in (3) using the Itô formula [44]. Consequently, rather than solving the PDEs (9) in the *entire function space* as in the prior work [15], it suffices to solve them *locally around high probability regions characterized by* (3), which leads to computationally scalable methods.

Parametrization. Looking from Theorem 1, it suffices to approximate $Y \approx Y_\theta$ and $\widehat{Y} \approx \widehat{Y}_\phi$ with two DNNs, since one can then infer $Z \approx \sigma \nabla Y_\theta$ and $\widehat{Z} \approx \sigma \nabla \widehat{Y}_\phi$, as suggested by (10), and solve for $(X_t, \bar{X}_s)$ via (11a, 12a). We refer to this choice as the *critic* parametrization as they resemble the critic networks [45]. An alternative parametrization, which offers extra robustness in practice, is to approximate $(Y, Z, \widehat{Y}, \widehat{Z}) \approx (Y_{\theta_1}, Z_{\theta_2}, \widehat{Y}_{\phi_1}, \widehat{Z}_{\phi_2})$[4] with four DNNs. We call this the *actor-critic* parametrization since $(Z, \widehat{Z})$ share the same role as the actor networks [46] in propagating dynamics.

Challenges in solving (11, 12) with IPF. SB-FBSDE [27] proposed to solve their FBSDE with IPF algorithm, which alternates optimizing the log-likelihood objectives $\log \rho_0 \propto \int \mathbb{E}[\mathrm{d}Y_t + \mathrm{d}\widehat{Y}_t]$ and $\log \rho_T \propto \int \mathbb{E}[\mathrm{d}Y_s + \mathrm{d}\widehat{Y}_s]$. Taking $F := 0$ in Theorem 1, these integrations will indeed recover the IPF objectives in (6), up to some independent constant. However, upon closer inspection, these integrations do *not* depend on the choice of F, since the two Fs appearing in (11b,11c) or (12b,12c) will cancel each other out in the sum of $\mathrm{d}Y + \mathrm{d}\widehat{Y}$ when computing the log-likelihoods. Mathematically, this makes sense when $(Y, \widehat{Y})$ are the exact solutions to the FBSDE systems (11, 12), yet it can be problematic when the parametrized $(Y_\theta, \widehat{Y}_\phi)$ have not yet satisfied these equations. This implies that

The IPF algorithm, which underlies modern numerical methods for solving parametrized SBs [24, 25, 27, 28]*, is insufficient in solving the FBSDEs with MF interaction* (11, 12).

Hence, we must seek additional training objectives, if any, in order to develop methods that fully respect the MF interacting structure in these new FBSDEs systems. As we will demonstrate in the following subsection, this can be achieved by drawing an intriguing connection to DeepRL.

3.3 Design of computational framework via Deep Reinforcement Learning (DeepRL)

Let us revisit the relation between the FBSDEs (11, 12) and their PDEs counterparts — but this time the HJB in (7). Take (X_t, Y_t) for example: The fact that $Y_t = \log \Psi(X_t, t) = -u(X_t, t)$ suggests an alternative interpretation of Y_t as the stochastic representation of the HJB, which, crucially, can be seen as the continuous-time analogue of the celebrated Bellman equation [47] for finite-horizon stochastic systems [48]. Indeed, discretizing (11b) with some fixed step size δt yields

$$Y_{t+\delta t} = Y_t + \left(\frac{1}{2}\|Z_t\|^2 + F_t\right) \delta t + Z_t^\top \delta W_t, \quad \delta W_t \sim \mathcal{N}(\mathbf{0}, \delta t \boldsymbol{I}), \tag{13}$$

[4] Hereafter, we drop the subscripts $\theta_{1,2}$ and $\phi_{1,2}$ for brevity, *i.e.,* $(Y_{\theta_1}, Z_{\theta_2}, \widehat{Y}_{\phi_1}, \widehat{Z}_{\phi_2}) \equiv (Y_\theta, Z_\theta, \widehat{Y}_\phi, \widehat{Z}_\phi)$.

Algorithm 1 Deep Generalized Schrödinger Bridge (DeepGSB)

Input: $(Y_\theta, \widehat{Y}_\phi)$ for critic parametrization or $(Y_\theta, \widehat{Y}_\phi, Z_\theta, \widehat{Z}_\phi)$ for actor-critic parametrization.
repeat
  Sample trajectories $\boldsymbol{X} \equiv \{X_t, Z_t, \delta W_t\}_{t\in[0,T]}$ from (11a); add $\boldsymbol{X}$ to replay buffer $\mathcal{B}$.
  for $k = 1$ **to** K **do**
    Sample on-policy $\boldsymbol{X}_{\text{on}}$ and off-policy $\boldsymbol{X}_{\text{off}}$ samples respectively from $\boldsymbol{X}$ and $\mathcal{B}$.
    Compute $\mathcal{L}(\phi) = \mathcal{L}_{\text{IPF}}(\phi; \boldsymbol{X}_{\text{on}}) + \mathcal{L}_{\text{TD}}(\phi; \boldsymbol{X}_{\text{on}}) + \mathcal{L}_{\text{TD}}(\phi; \boldsymbol{X}_{\text{off}}) + \|\widehat{Z}_\phi - \sigma\nabla\widehat{Y}\|$.
    Update ϕ with the gradient $\nabla_\phi\mathcal{L}(\phi)$.
  end for
  Sample trajectories $\bar{\boldsymbol{X}} \equiv \{\bar{X}_s, \widehat{Z}_s, \delta W_s\}_{s\in[0,T]}$ from (12a); add $\bar{\boldsymbol{X}}$ to replay buffer $\bar{\mathcal{B}}$.
  for $k = 1$ **to** K **do**
    Sample on-policy $\bar{\boldsymbol{X}}_{\text{on}}$ and off-policy $\bar{\boldsymbol{X}}_{\text{off}}$ samples respectively from $\bar{\boldsymbol{X}}$ and $\bar{\mathcal{B}}$.
    Compute $\mathcal{L}(\theta) = \mathcal{L}_{\text{IPF}}(\theta; \bar{\boldsymbol{X}}_{\text{on}}) + \mathcal{L}_{\text{TD}}(\theta; \bar{\boldsymbol{X}}_{\text{on}}) + \mathcal{L}_{\text{TD}}(\theta; \bar{\boldsymbol{X}}_{\text{off}}) + \|Z_\theta - \sigma\nabla Y\|$.
    Update θ with the gradient $\nabla_\theta\mathcal{L}(\theta)$.
  end for
until converges

which resembles a (non-discounted) Temporal Difference [49, 50] (TD) except that, in addition to the standard "*rewards*" (in terms of control and state costs), we also have a stochastic term. This stochastic term, which vanishes in the vanilla Bellman equation upon taking expectations, plays a crucial role in characterizing the inherited stochasticity of the value function Y_t. With this interpretation in mind, we aim to construct suitable TD targets for our FBSDEs systems. The results are summarized below.

Proposition 2 (TD targets for Generalized SB-FBSDEs)**.** *The single-step TD targets take the forms:*

$$\widehat{\text{TD}}^{\textit{single}}_{t+\delta t} := \widehat{Y}_t + \left(\frac{1}{2}\|\widehat{Z}_t\|^2 + \nabla\cdot(\sigma\widehat{Z}_t - f_t) + \widehat{Z}_t^\top Z_t - F_t\right)\delta t + \widehat{Z}_t^\top \delta W_t, \tag{14a}$$

$$\text{TD}^{\textit{single}}_{s+\delta s} := Y_s + \left(\frac{1}{2}\|Z_s\|^2 + \nabla\cdot(\sigma Z_s + f_s) + Z_s^\top \widehat{Z}_s - F_s\right)\delta s + Z_s^\top \delta W_s. \tag{14b}$$

Further, denote the incremental updates in (14) by $\delta\widehat{Y}_t := \widehat{\text{TD}}^{\textit{single}}_{t+\delta t} - \widehat{Y}_t$ *and* $\delta Y_s := \text{TD}^{\textit{single}}_{s+\delta s} - Y_s$. *Then, the multi-step TD targets take the forms:*

$$\widehat{\text{TD}}^{\textit{multi}}_{t+\delta t} := \widehat{\text{TD}}_0 + \sum_{\tau=\delta t}^{t} \delta\widehat{Y}_\tau, \qquad \text{TD}^{\textit{multi}}_{s+\delta s} := \text{TD}_0 + \sum_{\tau=\delta s}^{s} \delta Y_\tau, \tag{15}$$

where $\widehat{\text{TD}}_0 := \log\rho_0 - Y_0$ *and* $\text{TD}_0 := \log\rho_{\textit{target}} - \widehat{Y}_0$ *ensure the boundary conditions.*

It can be readily seen that the single-step TD targets in (14) obey a similar structure to (13), except deriving from different SDEs (11c, 12b); we will discuss these choices of SDEs shortly. Given (14), one can construct their multi-step TD targets (15) straightforwardly. In practice, we find that these multi-step variants often yield better performance, as consistently observed in the DeepRL literature [51–53]. Additionally, common practices such as computing $\widehat{\text{TD}}_t$ and TD_s using the exponential moving averaging (*i.e.,* target values) and replay buffers also help stabilize training. With Proposition 2, we can construct our TD objectives as

$$\mathcal{L}_{\text{TD}}(\theta) = \sum_{s=0}^{T} \mathbb{E}_{(3b)}\left[\|Y_\theta(\bar{X}_s, s) - \text{TD}_s\|\right]\delta s, \quad \mathcal{L}_{\text{TD}}(\phi) = \sum_{t=0}^{T} \mathbb{E}_{(3a)}\left[\|\widehat{Y}_\phi(X_t, t) - \widehat{\text{TD}}_t\|\right]\delta t \tag{16}$$

From (16), it should now be clear that these SDEs (11c, 12b) are deliberately chosen in Proposition 2 so that the related objectives for each parameter, *e.g.,* $\mathcal{L}_{\text{IPF}}(\theta)$ and $\mathcal{L}_{\text{TD}}(\theta)$, can be evaluated from the *same* sampling expectation, *e.g.,* $\mathbb{E}_{\bar{X}_s\sim(3b)}[\cdots]$. Doing so reduces the computational overhead. The fact that the TD targets in (16) appear as the regressing targets, *i.e., labels*, implies that from a computational standpoint, the MF interaction F needs *not* to be continuous or differentiable. We highlight this distinction, together with the results in Theorem 1 and Proposition 2, as they not only facilitate scalable methods for solving the challenging class of MFGs in (7), but also open up new algorithmic opportunities in the intersection of SB and DeepRL.

Our **DeepGSB** is summarized in Alg. 1, which alternately optimizes ϕ and θ as in IPF, yet with the new training objectives that account for the MF structure. Note that for the actor-critic parametrization, we add a *consistency loss*, *e.g.,* $\mathcal{L}_{\text{FK}}(\theta) := \|Z_\theta - \sigma\nabla Y\|$, to ensure that the nonlinear FK relations in (10) approximately hold. For the critic parametrization, these conditions are satisfied by construction.

4 Experiment

Instantiation of MFGs. We testify our DeepGSB on two classes of MFGs, including classical crowd navigation (d=2) and high-dimensional (d=1000) opinion depolarization. For crowd navigation, we consider three MFGs appearing in prior methods [14, 15], including *(i)* asymmetric obstacle avoidance, *(ii)* entropy interaction with a V-shape bottleneck, and *(iii)* congestion interaction on an S-shape tunnel. We will refer to them respectively as GMM, V-neck, and S-tunnel. The obstacles and the initial/target Gaussian distributions $(\rho_0, \rho_{\text{target}})$ are shown in Fig. 4. For opinion depolarization, we set ρ_0 and ρ_{target} to two zero-mean Gaussians with varying variances for representing the initially polarized and desired moderated opinion distributions. Finally, we consider zero and constant base drift f respectively for GMM and V-neck/S-tunnel, and adopt the polarized MF dynamics [4] for opinion MFG; see Sec. 4.2 for a detailed discussion.

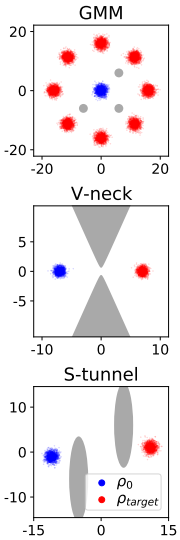

Figure 4: Crowd navigation MFGs.

MF interactions F. We follow standard treatments from the MFG theory [11] by noting that given a functional $\mathcal{F}(\rho)$ that quantifies the MF cost w.r.t. the population ρ, *e.g.,* $\mathcal{F}_{\text{congestion}} := \mathbb{E}_{x,y\sim\rho}[\frac{1}{\|x-y\|^2+1}]$ or $\mathcal{F}_{\text{entropy}} := \mathbb{E}_\rho[\log\rho]$, one can derive its associated MF interaction function $F(x,\rho)$ by taking the functional derivative, *i.e.,* $\frac{\delta\mathcal{F}(\rho)}{\delta\rho}(x) = F(x,\rho)$. Hence, the congestion and entropy MF interactions, together with the obstacle cost, follow:

$$F_{\text{congestion}} := \mathbb{E}_{y\sim\rho}\left[\frac{2}{\|x-y\|^2+1}\right], \qquad (17)$$
$$F_{\text{entropy}} := \log\rho(x,t)+1, \quad F_{\text{obstacle}} := 1500\cdot\mathbb{1}_{\text{obs}}(x),$$

where $\mathbb{1}_{\text{obs}}(\cdot)$ is the (discontinuous) indicator of the problem-dependent obstacle set. We leave the derivation of (17) to Appendix A.3.3 and summarize the MF interaction in Table 2.

Table 2: Setup of MF interaction F for each MFG.

GMM	F_{obstacle}
V-neck	$F_{\text{obstacle}} + F_{\text{entropy}}$
S-tunnel	$F_{\text{obstacle}} + F_{\text{congestion}}$
Opinion	F_{entropy}

Architecture & Hyperparameters. Recall the two parametrizations considered in Sec.3.2, *i.e.,* the critic $(Y_\theta, \sigma\nabla Y_\theta, \widehat{Y}_\phi, \sigma\nabla\widehat{Y}_\phi)$, **DeepGSB-c**, and actor-critic $(Y_\theta, Z_\theta, \widehat{Y}_\phi, \widehat{Z}_\phi)$, **DeepGSB-ac**. In both cases, we parametrize them with standard fully-connected DNNs for crowd navigation, and deep residual networks for high-dimensional opinion MFGs. All networks adopt sinusoidal time embeddings and are trained with AdamW [54]. All SDEs in (11, 12) are simulated with the Euler-Maruyama method. Due to space constraints, we will focus mostly on the results of **DeepGSB-ac** and leave the discussion of **DeepGSB-c**, alone with additional experimental details, to Appendix A.4.

4.1 Two-dimensional crowd navigation

Figure 5 shows the simulation results of our **DeepGSB-ac** on three crowd navigation MFGs. We also report existing numerical methods [14–16] that are best-tuned on each MFG (see Appendix A.4.1 for details) but note that in practice, they either require softening F to be differentiable [14, 15] to yield reasonable results, or discretizing the state space [16], which can lead to prohibitive complexity.[5]

We first compare to Chen [16] on GMM (see Fig. 5a) as their method only applies to non-MF interaction, *i.e.,* $F := F(x)$. While **DeepGSB-ac** guides the population to smoothly avoid all obstacles (notice the sharp contours of Y around them), [16] struggles to escape due to the discretization of the state space (hence the policy). To better examine the effect of ρ in $F(x,\rho)$, we next simulate

[5] As stated in [16], the complexity scales *quadratically* w.r.t. the number of discretized grid points.

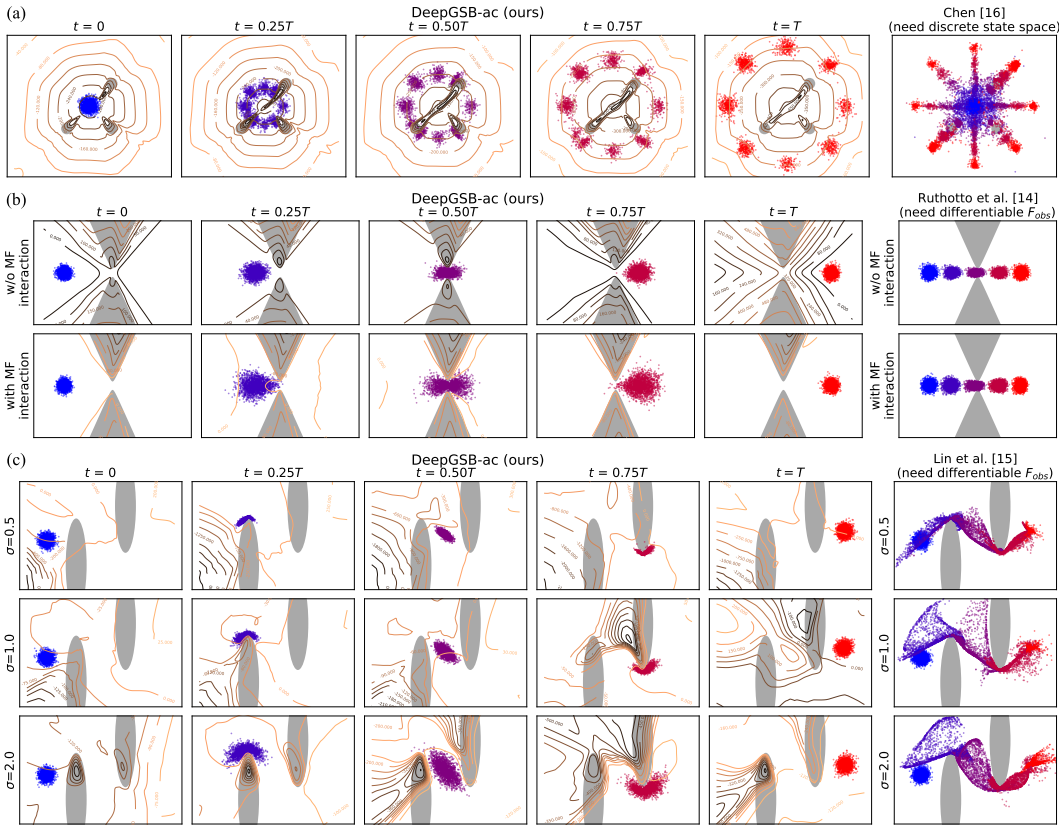

Figure 5: Simulation of the three crowd navigation MFGs, including (a) GMM, (b) V-neck, and (c) S-tunnel, from $t = 0$ to T. The first five columns show the population snapshots, each with a different color, guided by our **DeepGSB-ac**, whereas the sixth (rightmost) column overlays the same population snapshots generated by existing methods [14–16]. The time-varying contours represent $Y_\theta \approx \log \Psi$ whose gradient relates to the policy via $Z = \sigma \nabla Y$. This figure is best viewed in color.

the dynamics on V-neck (see Fig. 5b) with and without the MF interaction. It is clear that our **DeepGSB-ac** encourages the population to spread out once the entropy interaction F_{entropy} is enabled, yet a similar effect is barely observed in [14]. We observed difficulties in balancing the MF interaction F and the terminal penalty $D_{\text{KL}}(\rho_T || \rho_{\text{target}})$ for [14], yet this problem is alleviated in DeepGSB by construction. Lastly, we testify the robustness of our method on S-tunnel (see Fig. 5c) w.r.t. varying diffusions $\sigma = \{0.5, 1, 2\}$. Again, our **DeepGSB-ac** reaches the same ρ_{target} despite being subject to different levels of stochasticity. This is in contrast to [15], which, due to discarding the SDE dynamics, necessitates solving PDEs on the entire state space that may be sensitive to hyperparameters. In short, our DeepGSB outperforms prior methods [14–16] by better respecting obstacles and MF interactions yet without losing convergence to ρ_{target}, and further enjoying robustness to hyperparameters.

Table 3 compares DeepGSB-ac with DeepGSB-c. While DeepGSB-ac typically achieves smaller Wasserstein distance to ρ_{target} and TD errors, it seldom closes the consistency gap of the nonlinear FK $\mathcal{L}_{\text{FK}}(\theta) := \|Z_\theta - \sigma \nabla Y\|$ in (10), as opposed to DeepGSB-c. In practice, the results from DeepGSB-c are visually indistinguishable from DeepGSB-ac, despite the different contours of Y; we leave them to Appendix A.4.2 due to space constraint.

Table 3: Comparison of DeepGSB-ac *vs.* DeepGSB-c w.r.t Wasserstein distance to ρ_{target} and FBSDEs violation, in terms of TD errors and nonlinear FK, averaged over 3 runs.

MFGs	DeepGSB	$\mathcal{W}_2 \downarrow$	FBSDEs Violation $\downarrow$		
			$\mathcal{L}_{\text{TD}}(\phi)$	$\mathcal{L}_{\text{TD}}(\theta)$	$\mathcal{L}_{\text{FK}}(\theta)$
GMM	-ac	**.27**$\pm$.16	9.5$\pm$2.5	**7.1**$\pm$0.6	5.2$\pm$1.1
	-c	.61$\pm$.91	**7.0**$\pm$1.3	10.1$\pm$1.6	**0.0**$\pm$0.0
V-neck	-ac	**.00**$\pm$.00	**4.9**$\pm$1.5	**4.1**$\pm$0.5	0.6$\pm$0.2
	-c	.01$\pm$.00	8.2$\pm$0.8	8.7$\pm$1.6	**0.0**$\pm$0.0
S-tunnel	-ac	**.01**$\pm$.00	**25.5**$\pm$2.3	28.6$\pm$3.6	2.1$\pm$0.1
	-c	.03$\pm$.01	30.9$\pm$6.9	**26.4**$\pm$5.5	**0.0**$\pm$0.0

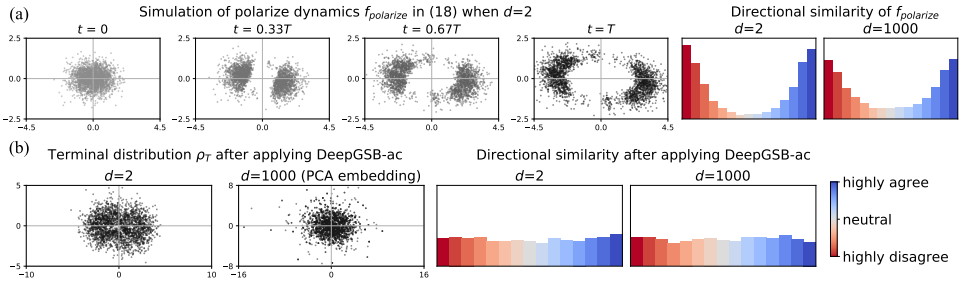

Figure 6: (a) Visualization of polarized dynamics $\bar{f}_{\text{polarize}}$ in 2- and 1000-dimensional opinion space, where the *directional similarity* [3] counts the histogram of cosine angle between pairwise opinions at the terminal distribution ρ_T. (b) **DeepGSB-ac** guides ρ_T toward moderated distributions, hence *depolarizes* the opinion dynamics. We use the first two principal components to visualize d=1000.

4.2 High-dimensional opinion depolarization

Finally, we showcase our DeepGSB in solving high-dimensional MFGs in the application of opinion dynamics [3–5], where each agent now possesses a d-dimensional opinion $x \in \mathbb{R}^d$ that evolves through the interactions with the population. In light of increasing recent attention, we consider a particular class of opinion dynamics known to yield strong *polarization* [4], *i.e.,* the agents' opinions tend to partition into groups holding diametric views. Take the *party model* [4] for instance: Given a random information $\xi \in \mathbb{R}^d$ sampled from some distribution independent of ρ, each agent updates his/her opinion following a normalized polarize dynamic $\bar{f}_{\text{polarize}} = {}^{f_{\text{polarize}}}\!/_{\|f_{\text{polarize}}\|^{\frac{1}{2}}}$, where

$$f_{\text{polarize}}(x, \rho; \xi) := \mathbb{E}_{y\sim\rho}\left[a(x, y; \xi)\bar{y}\right], \quad a(x, y; \xi) := \begin{cases} 1 & \text{if } \operatorname{sign}(\langle x, \xi\rangle) = \operatorname{sign}(\langle y, \xi\rangle) \\ -1 & \text{otherwise} \end{cases}, \tag{18}$$

and $\bar{y} = {}^{y}\!/_{\|y\|^{\frac{1}{2}}}$. The *agreement* function $a(x, y; \xi)$ indicates whether the two opinions x and y agree on the information ξ. Intuitively, the dynamic in (18) suggests that the agents tend to be receptive to opinions they agree with, and antagonistic to opinions they disagree with. As shown in Fig. 6a, this behavioral assumption, also known as biased assimilation [55, 56], can easily lead to polarization.

We can apply our MFG framework (7) to this polarized base drift (18), where, starting from some weakly polarized ρ_0, we seek a policy that compensates the polarization tendency and helps guide the population towards a moderated opinion distribution ρ_{target} (assuming as Gaussian for simplicity). We consider the entropy MF interaction F_{entropy} as it encourages opinions diversity before reaching consensus. As shown in Fig. 6b, in both lower- (d=2) and higher- (d=1000) dimensions, our **DeepGSB-ac** successfully guides the population towards the desired distribution centered symmetrically at $\mathbf{0} \in \mathbb{R}^d$, thereby mitigates the polarization. Results of DeepGSB-c remain similar despite more sensitive to hyperparameters; see Appendix A.4.2. We highlight these state-of-the-art results for solving these challenging MFGs that, comparing to existing methods [14, 15], consider a more difficult mean-field dynamic ($f(x, \rho)$ *vs.* $f(x)$) in an order of magnitude higher dimension (d=1000 *vs.* d=100).

5 Conclusion, Limitation, and Broader Impact

We present **DeepGSB**, a new numerical method for solving a challenging class of MFGs with distributional boundary constraints. By generalizing prior FBSDE theory for Schrödinger Bridge to accepting mean-field interactions, we show that practical training can be achieved via an intriguing algorithmic connection to DeepRL. Our DeepGSB outperforms prior methods in population navigation MFGs and sets a new state-of-the-art record in depolarizing 1000-dimensional opinion MFGs.

Our DeepGSB is primarily developed for solving MFGs in *unconstrained* state spaces such as $\mathbb{R}^d$. Yet, for applications in, *e.g.,* biomedicines and robotics, it may be necessary to consider *constrained* state spaces or *manifolds*, so that the resulting methods further respect domain-specific structures.

Study of MFGs possesses its own societal influence. Thus, as a numerical MFG solver, DeepGSB may pose a potential impact in offering solutions to previously unsolvable MFGs under more practical settings, thereby facilitating new understanding of population behavior with interacting agents.

References

[1] Yves Achdou, Francisco J Buera, Jean-Michel Lasry, Pierre-Louis Lions, and Benjamin Moll. Partial differential equation models in macroeconomics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 372(2028):20130397, 2014.

[2] Yves Achdou, Jiequn Han, Jean-Michel Lasry, Pierre-Louis Lions, and Benjamin Moll. Income and wealth distribution in macroeconomics: A continuous-time approach. *The review of economic studies*, 89(1):45–86, 2022.

[3] Simon Schweighofer, David Garcia, and Frank Schweitzer. An agent-based model of multi-dimensional opinion dynamics and opinion alignment. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(9):093139, 2020.

[4] Jason Gaitonde, Jon Kleinberg, and Éva Tardos. Polarization in geometric opinion dynamics. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 499–519, 2021.

[5] Jan Hązła, Yan Jin, Elchanan Mossel, and Govind Ramnarayan. A geometric model of opinion polarization. *arXiv preprint arXiv:1910.05274*, 2019.

[6] Zhiyu Liu, Bo Wu, and Hai Lin. A mean field game approach to swarming robots control. In *2018 Annual American Control Conference (ACC)*, pages 4293–4298. IEEE, 2018.

[7] Karthik Elamvazhuthi and Spring Berman. Mean-field models in swarm robotics: A survey. *Bioinspiration & Biomimetics*, 15(1):015001, 2019.

[8] Yiping Lu, Chao Ma, Yulong Lu, Jianfeng Lu, and Lexing Ying. A mean-field analysis of deep resnet and beyond: Towards provable optimization via overparameterization from depth. *arXiv preprint arXiv:2003.05508*, 2020.

[9] Kaitong Hu, Anna Kazeykina, and Zhenjie Ren. Mean-field langevin system, optimal control and deep neural networks. *arXiv preprint arXiv:1909.07278*, 2019.

[10] E Weinan, Jiequn Han, and Qianxiao Li. A mean-field optimal control formulation of deep learning. *arXiv preprint arXiv:1807.01083*, 2018.

[11] Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese journal of mathematics*, 2(1):229–260, 2007.

[12] Olivier Guéant, Jean-Michel Lasry, and Pierre-Louis Lions. Mean field games and applications. In *Paris-Princeton lectures on mathematical finance 2010*, pages 205–266. Springer, 2011.

[13] Alain Bensoussan, Jens Frehse, Phillip Yam, et al. *Mean field games and mean field type control theory*, volume 101. Springer, 2013.

[14] Lars Ruthotto, Stanley J Osher, Wuchen Li, Levon Nurbekyan, and Samy Wu Fung. A machine learning framework for solving high-dimensional mean field game and mean field control problems. *Proceedings of the National Academy of Sciences*, 117(17):9183–9193, 2020.

[15] Alex Tong Lin, Samy Wu Fung, Wuchen Li, Levon Nurbekyan, and Stanley J Osher. Alternating the population and control neural networks to solve high-dimensional stochastic mean-field games. *Proceedings of the National Academy of Sciences*, 118(31), 2021.

[16] Yongxin Chen. Density control of interacting agent systems. *arXiv preprint arXiv:2108.07342*, 2021.

[17] Jiequn Han, Arnulf Jentzen, and E Weinan. Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, 115(34): 8505–8510, 2018.

[18] Ioannis Exarchos and Evangelos A Theodorou. Stochastic optimal control via forward and backward stochastic differential equations and importance sampling. *Automatica*, 87:159–165, 2018.

[19] Marcus Pereira, Ziyi Wang, Ioannis Exarchos, and Evangelos A Theodorou. Neural network architectures for stochastic control using the nonlinear feynman-kac lemma. *arXiv preprint arXiv:1902.03986*, 2019.

[20] René Carmona, François Delarue, and Aimé Lachapelle. Control of mckean–vlasov dynamics versus mean field games. *Mathematics and Financial Economics*, 7(2):131–166, 2013.

[21] René Carmona and François Delarue. Mean field forward-backward stochastic differential equations. *Electronic Communications in Probability*, 18:1–15, 2013.

[22] René Carmona and Mathieu Laurière. Convergence analysis of machine learning algorithms for the numerical solution of mean field control and games: Ii–the finite horizon case. *arXiv preprint arXiv:1908.01613*, 2019.

[23] René Carmona and Mathieu Laurière. Convergence analysis of machine learning algorithms for the numerical solution of mean field control and games i: the ergodic case. *SIAM Journal on Numerical Analysis*, 59(3):1455–1485, 2021.

[24] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *arXiv preprint arXiv:2106.01357*, 2021.

[25] Francisco Vargas, Pierre Thodoroff, Neil D Lawrence, and Austen Lamacraft. Solving schrödinger bridges via maximum likelihood. *arXiv preprint arXiv:2106.02081*, 2021.

[26] Gefei Wang, Yuling Jiao, Qian Xu, Yang Wang, and Can Yang. Deep generative learning via schrödinger bridge. *arXiv preprint arXiv:2106.10410*, 2021.

[27] Tianrong Chen, Guan-Horng Liu, and Evangelos A Theodorou. Likelihood training of schrödinger bridge using forward-backward sdes theory. *arXiv preprint arXiv:2110.11291*, 2021.

[28] Charlotte Bunne, Ya-Ping Hsieh, Marco Cuturi, and Andreas Krause. Recovering stochastic dynamics via gaussian schr\" odinger bridges. *arXiv preprint arXiv:2202.05722*, 2022.

[29] Erwin Schrödinger. Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. In *Annales de l'institut Henri Poincaré*, volume 2, pages 269–310, 1932.

[30] Kenneth Caluya and Abhishek Halder. Wasserstein proximal algorithms for the schrödinger bridge problem: Density control with nonlinear drift. *IEEE Transactions on Automatic Control*, 2021.

[31] Julio Backhoff, Giovanni Conforti, Ivan Gentil, and Christian Léonard. The mean field schrödinger problem: ergodic behavior, entropy estimates and functional inequalities. *Probability Theory and Related Fields*, 178(1):475–530, 2020.

[32] Yongxin Chen, Tryphon Georgiou, and Michele Pavon. Optimal steering of inertial particles diffusing anisotropically with losses. In *2015 American Control Conference (ACC)*, pages 1252–1257. IEEE, 2015.

[33] Erwin Schrödinger. *Über die umkehrung der naturgesetze*. Verlag der Akademie der Wissenschaften in Kommission bei Walter De Gruyter u . . . , 1931.

[34] Christian Léonard. From the schrödinger problem to the monge–kantorovich problem. *Journal of Functional Analysis*, 262(4):1879–1920, 2012.

[35] Christian Léonard. A survey of the schr\" odinger problem and some of its connections with optimal transport. *arXiv preprint arXiv:1308.0215*, 2013.

[36] Michele Pavon and Anton Wakolbinger. On free energy, stochastic control, and schrödinger processes. In *Modeling, Estimation and Control of Systems with Uncertainty*, pages 334–348. Springer, 1991.

[37] Paolo Dai Pra. A stochastic control approach to reciprocal diffusion processes. *Applied mathematics and Optimization*, 23(1):313–329, 1991.

[38] Solomon Kullback. Probability densities with given marginals. *The Annals of Mathematical Statistics*, 39(4):1236–1243, 1968.

[39] Qinsheng Zhang and Yongxin Chen. Path integral sampler: a stochastic control approach for sampling. *arXiv preprint arXiv:2111.15141*, 2021.

[40] Eberhard Hopf. The partial differential equation ut+ uux= μxx. *Communications on Pure and Applied mathematics*, 3(3):201–230, 1950.

[41] Julian D Cole. On a quasi-linear parabolic equation occurring in aerodynamics. *Quarterly of applied mathematics*, 9(3):225–236, 1951.

[42] Jiongmin Yong and Xun Yu Zhou. *Stochastic controls: Hamiltonian systems and HJB equations*, volume 43. Springer Science & Business Media, 1999.

[43] Magdalena Kobylanski. Backward stochastic differential equations and partial differential equations with quadratic growth. *Annals of probability*, pages 558–602, 2000.

[44] Kiyosi Itô. *On stochastic differential equations*, volume 4. American Mathematical Soc., 1951.

[45] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[46] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[47] Richard Bellman. The theory of dynamic programming. Technical report, Rand corp santa monica ca, 1954.

[48] Hilbert J Kappen. Path integrals and symmetry breaking for optimal control theory. *Journal of statistical mechanics: theory and experiment*, 2005(11):P11011, 2005.

[49] Emanuel Todorov. Efficient computation of optimal actions. *Proceedings of the national academy of sciences*, 106(28):11478–11483, 2009.

[50] Michael Lutter, Shie Mannor, Jan Peters, Dieter Fox, and Animesh Garg. Value iteration in continuous actions, states and time. *arXiv preprint arXiv:2105.04682*, 2021.

[51] Lingheng Meng, Rob Gorbet, and Dana Kulić. The effect of multi-step methods on overestimation in deep reinforcement learning. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 347–353. IEEE, 2021.

[52] Harm van Seijen. Effective multi-step temporal-difference learning for non-linear function approximation. *arXiv preprint arXiv:1608.05151*, 2016.

[53] Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[54] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[55] Charles G Lord, Lee Ross, and Mark R Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology*, 37(11):2098, 1979.

[56] Pranav Dandekar, Ashish Goel, and David T Lee. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences*, 110(15):5791–5796, 2013.

[57] Etienne Pardoux and Shige Peng. Backward stochastic differential equations and quasilinear parabolic partial differential equations. In *Stochastic partial differential equations and their applications*, pages 200–217. Springer, 1992.

[58] Balint Negyesi, Kristoffer Andersson, and Cornelis W Oosterlee. The one step malliavin scheme: new discretization of bsdes implemented with deep learning regressions. *arXiv preprint arXiv:2110.05421*, 2021.

[59] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[60] Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

[61] Grigorios A Pavliotis. *Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations*, volume 60. Springer, 2014.

[62] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018.

[63] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Sec.5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Sec.5.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Theorem 1.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix A.3.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Alg. 1.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix A.4.1.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Table 3.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix A.4.1.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] See Appendix A.4.1.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] Not applicable.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] Not applicable.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] Not applicable.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] Not applicable.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] Not applicable.