From Deterministic Optimal Control Problems to Stochastic Differential Games

A PINN-Based Policy Iteration Approach for Nonconvex HJI Equations

• KIAS CAINS 2026 Winter Workshop

• 2026. 01. 07.

• Minjung Gim (NIMS)

Contents

- Deterministic optimal control problem ⇒ 1st order HJB equation
  - Why value functions satisfy HJB PDEs, and why viscosity solutions appear naturally
- Stochastic differential game ⇒ viscous 2nd order HJI equation
  - Controlled SDE, minimax value function, and the HJI characterization
- Method (Policy iteration + PINN)
  - Linear “policy evaluation” PDE + pointwise minimax “policy improvement”
- Theory & Experiments
  - Convergence under uniform ellipticity + error controlled by PINN residual

Deterministic Optimal Control and Value Function

- Consider a controlled ODE (deterministic)

$$dX(s) = f\big(s, X(s), a(s)\big)ds, \qquad s \in [t, T],$$
$$X(t) = x, \qquad a(s) \in A$$

- with regularity condition of f

- Define a cost functional

$$J(t, x; a) = \int_t^T c\big(s, X(s), a(s)\big)ds + g\big(X(T)\big)$$

- Define the value function

$$v(t, x) = \inf_a J(t, x; a)$$

- $v(t, x)$ is the best achievable future cost starting from state x at time t

Dynamic Programming Principle (DPP)

- Bellman principle
  - An optimal strategy has the property that whatever happens in the first short time interval, the remaining part of the strategy must still be optimal for the new state

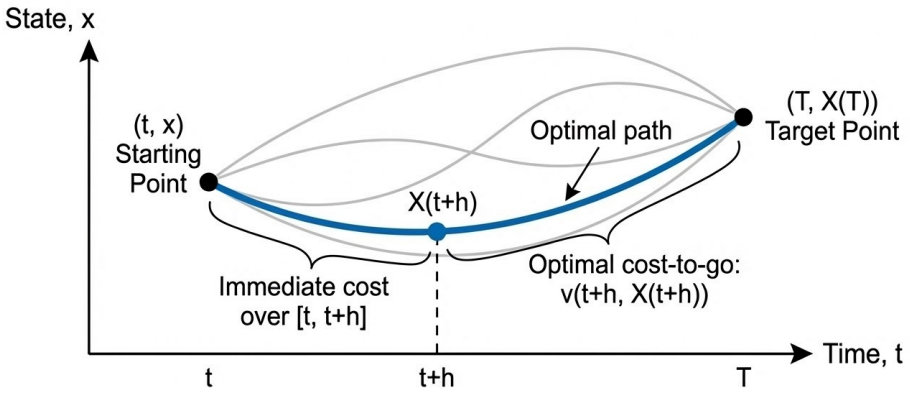
- One-step decomposition (small $h > 0$)
  - Let $X(\cdot)$ be the trajectory under control $a(\cdot)$

$$v(t,x) = \inf_{a} J(t,x;a)$$

$$= \inf_{a}\left[\int_{t}^{t+h} c\big(s, X(s), a(s)\big)ds + v\big(t+h, X(t+h)\big)\right]$$

Dynamic Programming Principle (DPP)

- Intuition in one sentence
  - Optimal cost from (t, x) = immediate cost over $[t, t+h]$ + optimal cost-to-go from the reached state at $t+h$



$$\inf_{a}\left[\int_{t}^{t+h} c\big(s, X(s), a(s)\big)ds + v\big(t+h, X(t+h)\big)\right]$$

- Why this matters
  - DPP is the bridge from optimization over paths to a local PDE condition (via Taylor expansion as $h \to 0$)

First-Order Hamilton—Jacobi—Bellman Equation

- From DPP + Taylor expansion (formal), $v(t,x)$ is a unique viscosity solution of first-order HJB PDE

$$\partial_t v(t,x) + \inf_{a\in A}\{c(t,x,a) + \nabla v(t,x)\cdot f(t,x,a)\} = 0,$$
$$v(T,x) = g(x)$$

- Equivalently,

$$\partial_t v(t,x) + H_{ctrl}(t,x,\nabla v) = 0$$

- where Hamiltonian

$$H_{ctrl}(t,x,p) = \inf_{a\in A}\big(c(t,x,a) + p\cdot f(t,x,a)\big)$$

- **Remark**: Even if f, c, g are smooth, v often develops kinks (nondifferentiable) due to optimization over controls; viscosity solutions capture the correct weak notion

From Control to Differential Games: Why HJI?

- In robust control / reachability / worst-case planning
  - Player I (controller): tries to minimize cost
  - Player II (adversary/disturbance): tries to maximize cost

- This naturally leads to a minimax value function and hence to a Hamilton-Jacobi-Isaacs (HJI) PDE rather than HJB equation

Stochastic Differential Game

- Controlled SDE(two player)

$$dX(s) = f\big(s, X(s), a(s), b(s)\big)ds + \sigma\big(s, X(s)\big)dW_s,$$
$$s \in [t, T], \qquad X(t) = x$$

- Standing assumptions
  - Probability space $\big(\Omega, \mathcal{F}, \{\mathcal{F}_s\}_{\{0 \le s \le T\}}, P\big)$
  - W_s: m-dim standard Brownian motion
  - (f, σ) are Borel measurable, continuous, uniformly Lipschitz in x, with linear growth in x
  - Uniform ellipticity: $\sigma\sigma^{\top} \geqslant \lambda I_d$ for some $\lambda > 0$
  - Admissible controls

$$\mathcal{A}_t = \big\{a: [t, T] \to A \ \{\mathcal{F}_s\}_{s \in [t,T]} - \text{adapted}\big\},$$
$$\mathcal{B}_t = \big\{b: [t, T] \to B \ \{\mathcal{F}_s\}_{s \in [t,T]} - \text{adapted}\big\}$$

Stochastic Differential Game

- Cost functional

$$J(t,x;a,b) = E\left[\int_t^T c\big(s,X(s),a(s),b(s)\big)ds + g\big(X(T)\big)\right]$$

  - Player I minimizes; Player II maximizes

- Nonanticipative strategies for Player II

$$\Gamma_t = \{\beta : \mathcal{A}_t \to \mathcal{B}_t \mid \beta \text{ is nonanticipating}\}$$

- I.e. for $a_1, a_2 \in \mathcal{A}_t$ and $s \in [t,T]$

$$a_1(\cdot) = a_2(\cdot) \text{ on } [t,s) \;\Rightarrow \beta[a_1](\cdot) = \beta[a_2] \text{ on } [t,s)$$

- Value function

$$v(t,x) := \sup_{\beta\in\Gamma_t} \inf_{a\in\mathcal{A}_t} J(t,x;a,\beta[a])$$

Viscous Hamilton—Jacobi—Isaacs equation

- Under standard regularity + uniform ellipticity, the value function satisfies the viscous HJI PDE:

$$\partial_t v(t,x) + H\big(t,x,\nabla v(t,x)\big) = -\frac{1}{2}\mathrm{Tr}\big(\sigma\sigma^\top(t,x)D_{xx}^2 v(t,x)\big),$$
$$v(T,x) = g(x)$$

- Minimax Hamiltonian and Lagrangian

$$H(t,x,p) = \sup_{b\in B}\inf_{a\in A} L(t,x,p)(a,b)$$
$$L(t,x,p)(a,b) = c(t,x,a,b) + p\cdot f(t,x,a,b)$$

Policy Iteration (Idealized Mesh-Free PI: Algorithm 1)

- Policy iteration viewpoint
  - Freeze feedback controls (α_n, β_n), solve a linear PDE for v_n, then update (α, β) from v_n

Policy Iteration (Idealized Mesh-Free PI: Algorithm 1)

1. Policy evaluation (linear PDE with frozen policies)

$$\partial_t v_n(t,x) + L\big(t,x,\nabla v_n(t,x)\big) = -\frac{1}{2}\mathrm{Tr}\big(\sigma\sigma^\top(t,x)D^2_{xx}v_n(t,x)\big),$$
$$v_n(T,x) = g(x)$$
$$L(t,x,p)(\alpha_n,\beta_n) = c(t,x,\alpha_n,\beta_n) + p\cdot f(t,x,\alpha_n,\beta_n)$$

2. Policy improvement
   - Fix (t,x) and set $p = \nabla v_n$
   - Update (α,β) by a pointwise minimax step

$$\alpha_{n+1,b}(t,x) = \operatorname*{argmin}_{a\in A} L(t,x,p)(a,b)$$
$$\beta_{n+1}(t,x) = \operatorname*{argmin}_{b\in B} L(t,x,p)(\alpha_{n+1,b}(t,x),b)$$
$$\alpha_{n+1}(t,x) = \alpha_{n+1,\beta_n(t,x)}(t,x)$$

PINN-Based Policy Iteration (Algorithm 2)

- In practice, we cannot “exactly” solve the linear PDE at each PI step, so we use a PINN for policy evaluation

- PINN residual loss for fixed (α_n, β_n)

$$
\begin{aligned}
&J(\theta_n) \\
&= \frac{1}{N_{int}} \sum_{j=1}^{N_{int}} \Big| \partial_t v_n\big(t^{(j)}, x^{(j)}; \theta_n\big) + L\big(t^{(j)}, x^{(j)}, \nabla v_n\big)(\alpha_n, \beta_n) \\
&+ \frac{1}{2} \mathrm{Tr}\left(\sigma\sigma^{\top}\big(t^{(j)}, x^{(j)}\big) D_{xx}^2 v_n\big(t^{(j)}, x^{(j)}; \theta_n\big)\right) \Big|^2 \\
&+ \frac{1}{N_{bc}} \sum_{k=1}^{N_{bc}} \left| v_n\left(T, x_T^{(k)}; \theta_n\right) - g\left(x_T^{(k)}\right) \right|^2
\end{aligned}
$$

PINN-Based Policy Iteration (Algorithm 2)

- Hard terminal constraint (for training stability)

$$v_n(t,x;\theta_n) = g(x) + (T-t)N_n(t,x;\theta_n)$$

- Algorithm 2 summary
  - Train $v_n(t,x;\theta_n)$ via PINN and compute ∇v_n via automatic differentiation
  - Update $(\alpha_{n+1}, \beta_{n+1})$ pointwise in (t,x)
  - warm-start $\theta_{n+1} \leftarrow \theta_n$

Theory: PI Convergence + Error Control (Practical)

- Exact PI (idealized)
  - Under the regularity assumptions, the exact policy iteration v_n converges (locally uniformly) to v, the unique bounded continuous viscosity solution of the viscous HJI

- PINN-PI: Let Algorithm 2 produce $(\tilde{v}_n, \tilde{\alpha}_n, \tilde{\beta}_n)$. Define the evaluation residual

$$R_n := \partial_t \tilde{v}_n + L(t, x, \nabla \tilde{v}_n) + \frac{1}{2}\mathrm{Tr}(\sigma\sigma^\top D_{xx}^2 \tilde{v}_n),$$
$$r_n := \|R_n\|_2$$

- There exist $C > 0$, $\rho \in (0,1)$ such that

$$\sup_{t\in[0,T]} \|\tilde{v}_n(t,\cdot) - v(t,\cdot)\|_2 \le C(r_n + \rho^n)$$

Experiment: 2D Planning with Moving Obstacle

- Controlled SDE: $X \in \mathbb{R}^2$

$$dX(s) = \big(a(s) + b(s)\big)ds + \sigma dW_s, \qquad |a(s)| \le 1, |b(s)| \le \delta$$

- Cost functional

$$J(t, x; a, b)$$

$$= E\left[\int_t^T \Big(\lambda_1 \|a(s)\|^2 + \lambda_2 \phi\big(s, X(s)\big)\Big) ds + \lambda_3 \left\|X(T) - x_{\text{goal}}\right\|^2\right]$$

- where the obstacle penalty function $\phi(s, x)$ is given by

$$x_{\text{obs}}(s) = (0.5 \cos \pi s\,, 0.5 \sin \pi s)^{\mathsf{T}}$$

$$\phi(s, x) = \exp\left(-\frac{\left\|x - x_{\text{obs}}(s)\right\|^2}{2\varepsilon^2}\right),$$

Experiment: 2D Planning with Moving Obstacle

- Value function

$$v(t,x) = \sup_{\beta \in \Gamma_t} \inf_{a \in \mathcal{A}_t} J(t,x;a,b)$$

- satisfy the viscous HJI equation

$$\partial_t v(t,x) + H\big(t,x,\nabla v(t,x)\big) = -\frac{1}{2}\mathrm{Tr}\big(\sigma\sigma^{\top}(t,x)D_{xx}^2 v(t,x)\big),$$

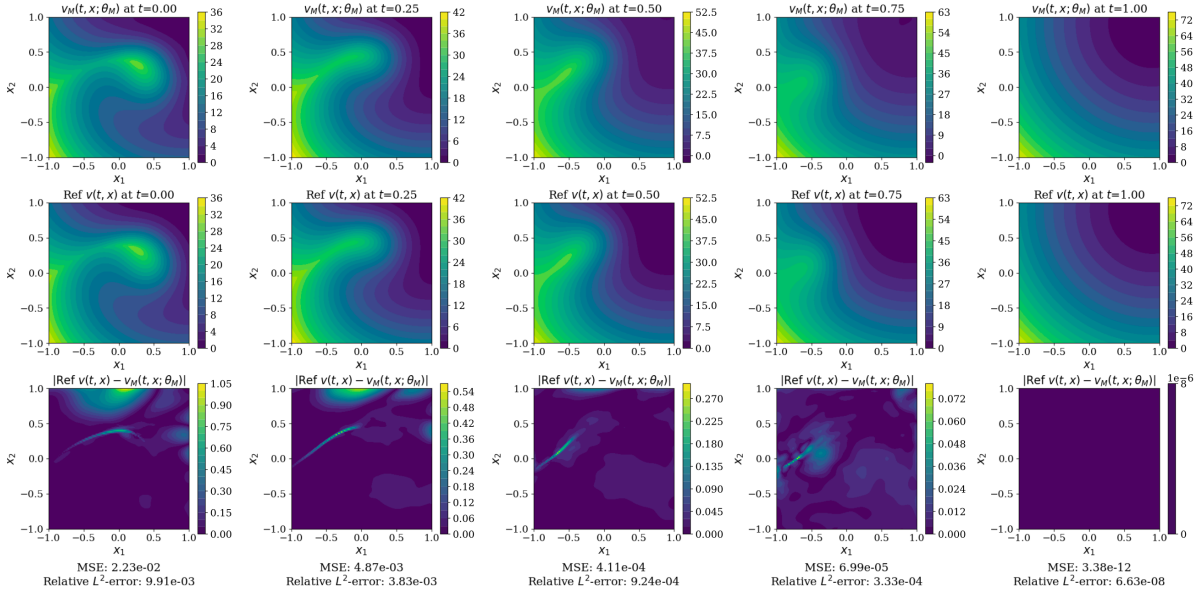
$$v(T,x) = \lambda_3 \left\| x - x_{\mathrm{goal}} \right\|^2$$

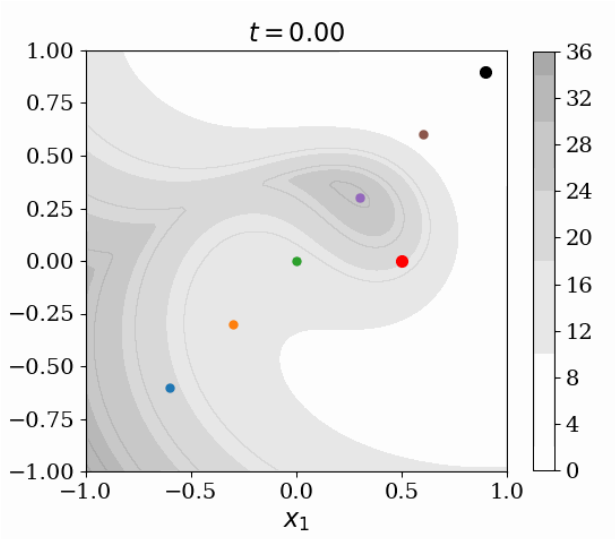
- with

$$H(t,x,p) = \sup_{b \in B} \inf_{a \in A} [\lambda_1 \|a(s)\|^2 + \lambda_2 \phi(t,x) + p \cdot (a+b)]$$

Experiment: 2D Planning with Moving Obstacle

- The simulation domain to $x \in [-1,1]^2$
- The terminal time to $T = 1.0$
- $\delta = 0.1$, $\lambda_1 = 0.1$, $\lambda_2 = 100$, $\lambda_3 = 10$, $\varepsilon = 0.3$ and the diffusion matrix is given by $\sigma = 0.1 I_2$
- The target position is fixed at $x_{\text{goal}} = (0.9,0.9)$





Experiment: 2N+1 Publisher-Subscriber

- Publisher state $x_0(s)$ and subscriber states $x_1(s), \dots, x_{2N}(s)$
- Controlled SDE: $X \in \mathbb{R}^{2N+1}$

$$dX(s) = f\big(X(s), u(s), d(s)\big)ds + \sigma dW_s,$$
$$u(s) \in \mathbb{R}^{2N}, \qquad |u(s)| \le 1,$$
$$d(s) \in \mathbb{R}^{2N}, \qquad |d(s)| \le 1$$

- The drift term is expressed as

$$f(x, u, d) = Ax + Bu + Cd + \psi(x)$$

- where

$$A = e_1 e_1^\top - 1_{2N+1}\, e_1^\top + a\, I_{2N+1}, \qquad B = \begin{bmatrix} 0 \\ bI_{2N} \end{bmatrix}$$

- The rotated disturbance map

$$C = \begin{bmatrix} 0 \\ c\ \mathrm{diag}(R(\theta_1), \dots, R(\theta_N)) \end{bmatrix}, \qquad R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

Experiment: 2N+1 Publisher-Subscriber

- The nonlinear interaction

$$\psi(x) = \begin{bmatrix} \alpha \sin x_0 \\ -\beta x_0 1_{2N} \end{bmatrix} \circ (x \circ x)$$

- where $\circ$ is the Hadamard product
- Terminal Cost function

$$g(x) = \frac{1}{2}\left(N x_0^2 + \sum_{j=1}^{2N} x_j^2 - N r^2 \right)$$

- Value function can be decomposable

$$v(t,x) = \sum_{i=1}^{N} v_i(t, x_0, x_{2i-1}, x_{2i})$$

Experiment: 2N+1 Publisher-Subscriber

- The Hamiltonian

$$H(x,p) = p^{\top} q(x) - \|B^{\top} p\|_1 + \|C^{\top} p\|_1$$

- where $q(x) = \dot{x} = Ax + \psi(x)$

- For $N = 1$ (publisher, 2 subscribers), let $p = (p_1, p_2, p_3)^{\top}$ and

$$B = \begin{bmatrix} 0 \\ bI_2 \end{bmatrix}, \qquad C = \begin{bmatrix} 0 \\ c\, R(\theta) \end{bmatrix}$$

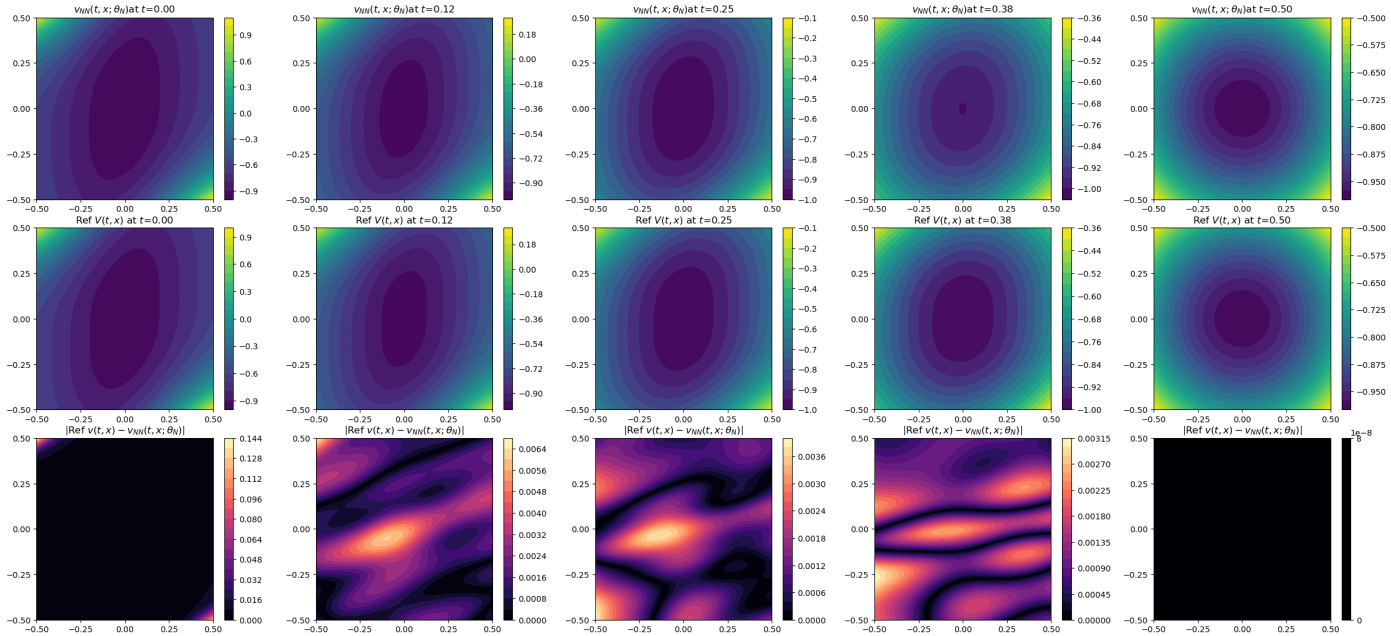
- Then generally $H(x,p)$ is neither convex nor concave in p, i.e.,

$$-\|B^{\top} p\|_1 + \|C^{\top} p\|_1 = -b\|(p_2, p_3)\|_1 + c\|R(\theta)^{\top}(p_2, p_3)\|_1$$

Experiment: 2N+1 Publisher-Subscriber

- 3D results

	Mean	std
PINN PI	1.03e-02	1.16e-03
Direct PINN	1.41e-02	2.95e-03



MSE: 5.41e-05
Relative L^2-error: 9.58e-03

MSE: 4.79e-06
Relative L^2-error: 2.75e-03

MSE: 1.98e-06
Relative L^2-error: 1.72e-03

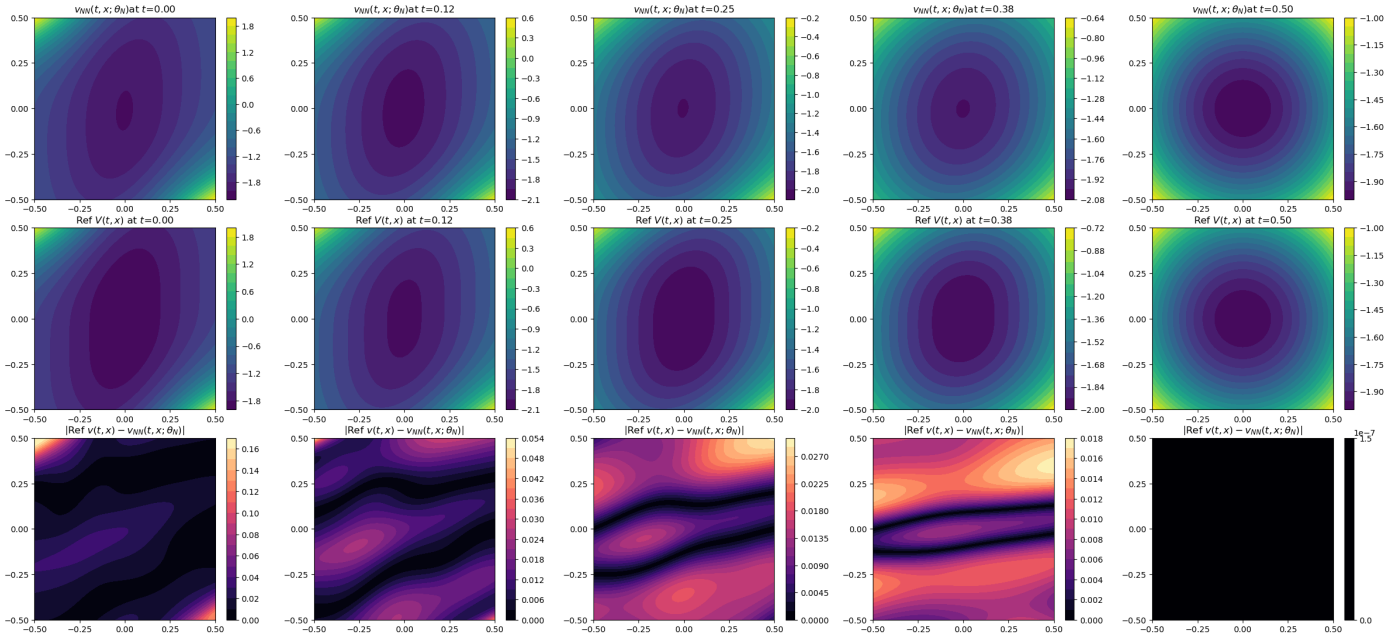
MSE: 1.65e-06
Relative L^2-error: 1.54e-03

MSE: 3.97e-16
Relative L^2-error: 2.38e-08

Experiment: 2N+1 Publisher-Subscriber

- 5D results

	Mean	std
PINN PI	1.46e-02	2.30e-03
Direct PINN	1.83e-02	1.53e-03



MSE: 6.41e-04
Relative L^2-error: 1.65e-02

MSE: 1.58e-04
Relative L^2-error: 7.91e-03

MSE: 1.59e-04
Relative L^2-error: 7.69e-03

MSE: 9.30e-05
Relative L^2-error: 5.78e-03

MSE: 1.59e-15
Relative L^2-error: 2.38e-08

Conclusion

- Stochastic differential games lead to a viscous HJI PDE with a minimax Hamiltonian
- We use Policy Iteration (PI)
  (1) policy evaluation = solve a linear PDE with frozen policies
  (2) policy improvement = pointwise minimax update $p = \nabla$

- PINN-based PI makes this practical: we approximate each evaluation PDE by a PINN, use hard terminal constraint, and warm-start across PI steps
- Theory-wise, we can control the error by the PINN residual (plus PI contraction term)
- In experiments (moving obstacle, publisher-subscriber), PINN-PI improves accuracy compared to a direct PINN baseline

Thanks