Mathematical
Institute

Neural network based policy iteration for values and controls of stochastic games

Yufei Zhang

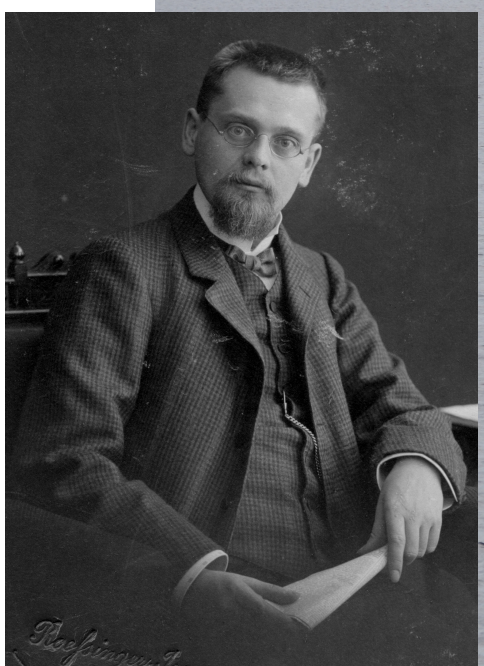
based on work with Kazufumi Ito and Christoph Reisinger

ESSFM, 4 September 2019

Oxford
Mathematics

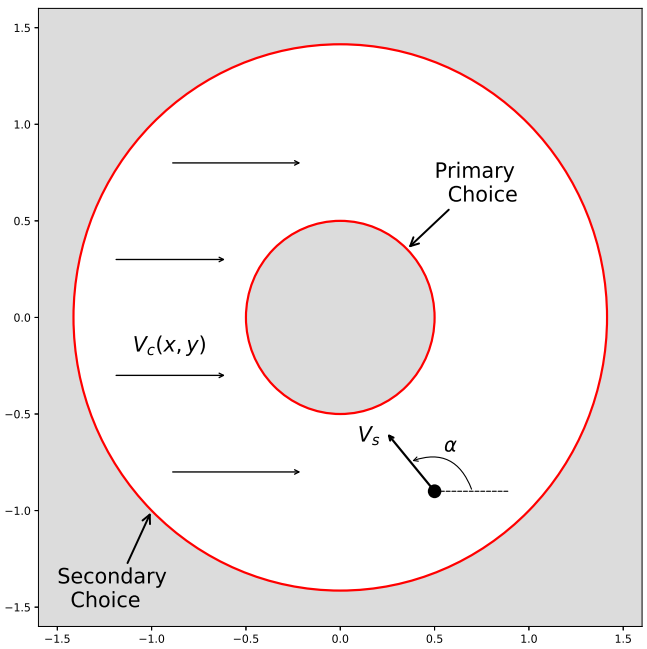
Zermelo's Navigation Problem

UNIVERSITY OF OXFORD
Mathematical Institute



Zermelo, Ernst (1931). Über das Navigationsproblem bei ruhender oder veränderlicher Windverteilung. Zeitschrift für Angewandte Mathematik und Mechanik. 11 (2).

A time-optimal control problem: find the optimal trajectories of a ship navigating a region Ω of strong winds.



- The domain Ω is an annulus.
- The wind "mainly" blows towards the positive x-axis with velocity v_c.
- The ship moves with a velocity v_s, and the captain can control its direction α.
- The captain prefers to exit from the inner circle.

Given a bounded open set $\Omega \subset \mathbb{R}^n$, strategies $\{\alpha_t\}_{t\geq 0}$ and $\{\beta_t\}_{t\geq 0}$, we consider $X^{x,\alpha,\beta}$ governed by:

$$dX_t = b(X_t, \alpha_t, \beta_t)dt + \sigma(X_t)\, dW_t, \quad t \in [0,\infty); \quad X_0 = x \in \Omega,$$

and the following value function:

$$u(x) = \inf_{\alpha\in\mathcal{A}} \sup_{\beta\in\mathcal{B}} \mathbb{E}\left[\int_0^{\tau_{x,\alpha,\beta}} f(X_t^{x,\alpha,\beta}, \alpha_t, \beta_t)\, dt + g(X_{\tau_{x,\alpha,\beta}}^{x,\alpha,\beta})\right],$$

where $\tau_{x,\alpha,\beta}$ is the first exit time of $X^{x,\alpha,\beta}$ from Ω, $\mathcal{A}$ and $\mathcal{B}$ contain all admissible controls valued in $\mathbb{A}$ and $\mathbb{B}$, respectively.

The value function u is a solution of the Hamilton–Jacobi–Bellman–Isaacs (HJBI) equation:

$$\begin{aligned} -a^{ij}(x)\partial_{ij}u + H(x, u, \nabla u) = 0, \quad & \text{in } \Omega \subset \mathbb{R}^n, \\ u = g, \quad & \text{on } \partial\Omega, \end{aligned}$$

where $a = \sigma\sigma^T/2$, and the Hamiltonian is given by:

$$H(x, u, \nabla u) = \max_{\alpha\in\mathbb{A}} \min_{\beta\in\mathbb{B}} \left(b^i(x, \alpha, \beta)\partial_i u(x) + c(x, \alpha, \beta)u(x) - f(x, \alpha, \beta) \right).$$

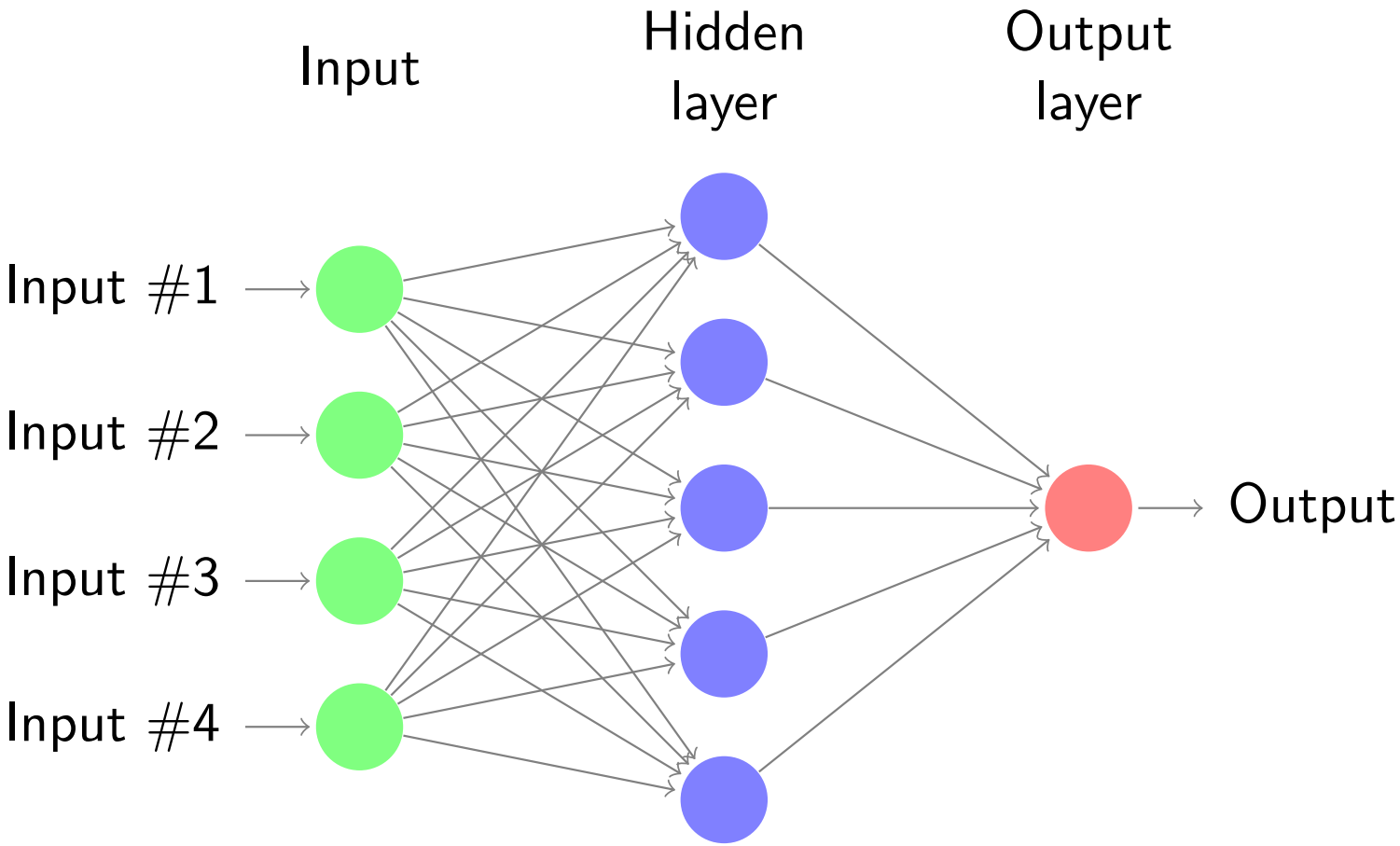
Let u be a solution to the HJBI equation, the optimal (feedback) controls are given by: for all $x \in \Omega$,

$$(\alpha(x), \beta(x)) \in \arg\max_{\alpha\in\mathbb{A}} \min_{\beta\in\mathbb{B}} \left(b^i(x,\alpha,\beta)\partial_i u(x) + c(x,\alpha,\beta)u(x) - f(x,\alpha,\beta)\right).$$

Artificial neural networks

Example one hidden layer

UNIVERSITY OF OXFORD
Mathematical Institute



UNIVERSITY OF OXFORD
Mathematical Institute

Definition

Let $\varrho : \mathbb{R} \to \mathbb{R}$ be a given function. For each $l = 1, \dots, L$, let

$$T_l : \mathbb{R}^{N_{l-1}} \to \mathbb{R}^{N_l}, \quad T_l(x) = W_l x + b_l$$

for some $W_l \in \mathbb{R}^{N_l \times N_{l-1}}$ and $b_l \in \mathbb{R}^{N_l}$. A function $F : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$,

$$F(x) = \big(T_L \circ (\varrho \circ T_{L-1}) \circ \cdots (\varrho \circ T_1)\big)(x), \quad x \in \mathbb{R}^{N_0},$$

is called (the realisation of) a feedforward neural network.

- L is the depth of F.
- $N_1, \dots, N_{L-1}$ are the dimensions of the hidden layers.
- The number of unknown $\{W_l, b_l\}_{l=1}^{N}$ is the complexity $\mathscr{C}$ of F.

UNIVERSITY OF OXFORD
Mathematical Institute

- Choose a family of neural networks $\mathcal{F}_M$ parametrized by $\theta := \{W_l, b_l\}_{l=1}^L \in \mathbb{R}^M$, i.e., $\mathcal{F}_M = \{u^\theta \mid \theta \in \mathbb{R}^M\}$.
- Formulate the semilinear Dirichlet problem

$$-a^{ij}(x)\partial_{ij}u + H(x, u, \nabla u) = 0, \quad \text{in } \Omega,$$
$$u = g, \quad \text{on } \partial\Omega,$$

into a nonlinear optimization problem over $\mathcal{F}_M$:

$$\underbrace{\min_{\theta\in\mathbb{R}^M}}_{\min_{u^\theta\in\mathcal{F}_M}} \| - a^{ij}\partial_{ij}u^\theta + H(\cdot, u^\theta, \nabla u^\theta)\|^2_{L^2(\Omega)} + \|u^\theta - g\|^2_Y.$$

Note that $Y = L^2(\partial\Omega)$ in most published works.

See, e.g., J. Sirignano and K. Spiliopoulos, DGM: A deep learning algorithm for solving partial differential equations, J. Comput. Phys., 375 (2018), pp. 1339–1364.

- Discretize the optimization problem by using samples $\Omega_d \subset \Omega$ and $\partial\Omega_d \subset \partial\Omega$: solve $\min_{\theta\in\mathbb{R}^M} J(\theta)$, where

$$J(\theta) = \frac{1}{|\Omega_d|} \sum_{x_i\in\Omega_d} |-a^{ij}(x_i)\partial_{ij}u^\theta(x_i) + H(x_i, u^\theta(x_i), \nabla u^\theta(x_i))|^2 + \frac{1}{|\partial\Omega_d|} \sum_{x_i\in\partial\Omega_d} |u^\theta(x_i) - g(x_i)|^2.$$

- Discretize the optimization problem by using samples $\Omega_d \subset \Omega$ and $\partial\Omega_d \subset \partial\Omega$: solve $\min_{\theta\in\mathbb{R}^M} J(\theta)$, where

$$J(\theta) = \frac{1}{|\Omega_d|} \sum_{x_i\in\Omega_d} |-a^{ij}(x_i)\partial_{ij}u^\theta(x_i) + H(x_i, u^\theta(x_i), \nabla u^\theta(x_i))|^2 + \frac{1}{|\partial\Omega_d|} \sum_{x_i\in\partial\Omega_d} |u^\theta(x_i) - g(x_i)|^2.$$

- Find $\theta \in \arg\min_{\theta\in\mathbb{R}^M} J(\theta)$ by minimizing $J(\theta)$ with stochastic gradient descent (SGD) method:

$$\theta^{k+1} := \theta^k - \eta\nabla_\theta J(\theta^k), \quad \forall k \geq 1.$$

Pros:

- It is a mesh-free method.
- Neural networks can approximate certain high-dimensional functions with reasonable complexity.

Cons:

- It is a non-convex non-smooth optimization problem due to the non-convexity of the Hamiltonian H.
- SGD requires evaluations of H and its gradient, which may not be possible if $\mathbb{A}$ and $\mathbb{B}$ are high-dimensional.
- The choice of $Y = L^2(\partial\Omega)$ does not ensure the convergence of ∇u, hence no convergence of optimal controls.

Neural network based policy iteration algorithm

Mathematical Institute

Iteratively linearize the HJBI equation:

$$-a^{ij}\partial_{ij}u + H(\cdot, u, \nabla u) = 0, \quad \text{in } \Omega,$$

and then solve the linear equations by neural networks.

Recall the optimal controls associated with u are given by:

$$(\alpha(\cdot), \beta(\cdot)) \in \arg\max_{\alpha\in\mathbb{A}} \min_{\beta\in\mathbb{B}} \left(b^i(\cdot, \alpha, \beta)\partial_i u + c(\cdot, \alpha, \beta)u - f(\cdot, \alpha, \beta)\right).$$

UNIVERSITY OF OXFORD
Mathematical Institute

Algorithm (Policy iteration for HJBI equations)

Initialization: choose an initial guess u^0 *and set* $k = 0$.

1. *Given the function* u^k, *update the control laws:* $\forall x \in \Omega$,

$$(\alpha^k(x), \beta^k(x)) \in \arg\max_{\alpha\in\mathbb{A}} \min_{\beta\in\mathbb{B}} \left(b^i(x,\alpha,\beta)\partial_i u^k(x) + c(x,\alpha,\beta)u^k(x) - f(x,\alpha,\beta)\right)$$

2. *Solve the linear problem for* u^{k+1}:

$$-a^{ij}\partial_{ij}u + b^i_k\partial_i u + c_k u - f_k = 0, \quad \text{in } \Omega; \quad u = g, \quad \text{on } \partial\Omega, \quad (1)$$

where $\phi_k(\cdot) := \phi(\cdot, \alpha^k(\cdot), \beta^k(\cdot))$ *for* $\phi = b^i, c, f$.

Algorithm (Inexact policy iteration for HJBI equations)

Initialization: choose a family of trial functions $\mathcal{F} \subset H^2(\Omega)$, *an initial guess* $u^0 \in \mathcal{F}$, *a sequence* $\{\eta_k\}_{k\in\mathbb{N}}$ *of non-negative scalars.*

1. *For each* $k \geq 0$, *given the function* $u^k \in \mathcal{F}$, *update the control laws* (α^k, β^k) *as before.*
2. *Find* $u^{k+1} \in \mathcal{F}$ *such that*

$$\|(-a^{ij}\partial_{ij} + b_k^i\partial_i + c_k)u^{k+1} - f_k\|^2_{L^2(\Omega)} + \|u^{k+1} - g\|^2_{H^{3/2}(\partial\Omega)} \leq \eta_{k+1}\|u^{k+1} - u^k\|^2_{H^2(\Omega)},$$

where b_k, c_k *and* f_k *depend on* (α^k, β^k) *as in* (1).

Theorem

Suppose $\mathcal{F}$ is dense in $H^2(\Omega)$, $\lim_{k\to\infty} \eta_k = 0$, and the coefficients are sufficiently regular.

Theorem

Suppose $\mathcal{F}$ is dense in $H^2(\Omega)$, $\lim_{k\to\infty} \eta_k = 0$, and the coefficients are sufficiently regular. Let $u^* \in H^2(\Omega)$ be the solution to the HJBI equation.

Then for any initial guess $u^0 \in \mathcal{F}$, we have:

1. *$\{u^k\}_{k\in\mathbb{N}}$ converge superlinearly to u^* in $H^2(\Omega)$, i.e.,* $\lim_{k\to\infty} \|u^{k+1} - u^*\|_{H^2(\Omega)} / \|u^k - u^*\|_{H^2(\Omega)} = 0.$

Theorem

Suppose $\mathcal{F}$ is dense in $H^2(\Omega)$, $\lim_{k\to\infty}\eta_k = 0$, and the coefficients are sufficiently regular. Let $u^* \in H^2(\Omega)$ be the solution to the HJBI equation.

Then for any initial guess $u^0 \in \mathcal{F}$, we have:

1. *$\{u^k\}_{k\in\mathbb{N}}$ converge superlinearly to u^* in $H^2(\Omega)$, i.e., $\lim_{k\to\infty} \|u^{k+1} - u^*\|_{H^2(\Omega)} / \|u^k - u^*\|_{H^2(\Omega)} = 0$.*
2. *$\lim_{k\to\infty}(u^k, \partial_i u^k, \partial_{ij} u^k)(x) = (u^*, \partial_i u^*, \partial_{ij} u^*)(x)$ for a.e. $x \in \Omega$, and for all $i, j = 1, \ldots, n$.*

Theorem

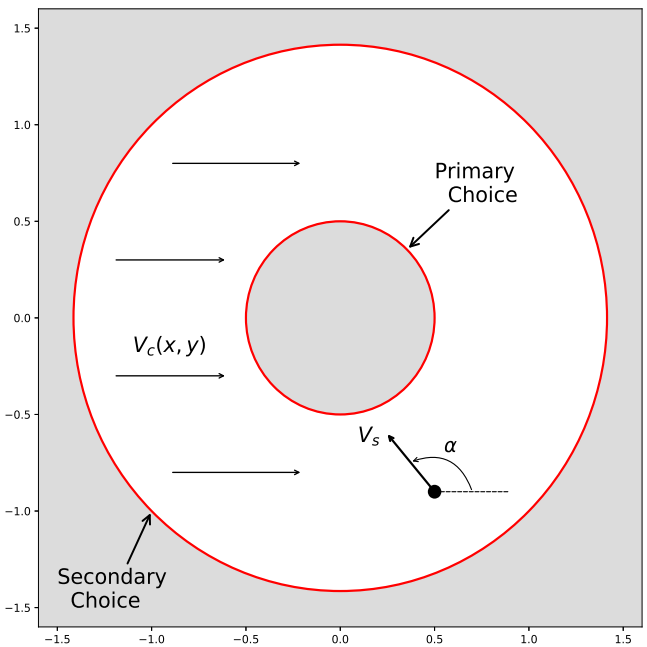
Suppose $\mathcal{F}$ is dense in $H^2(\Omega)$, $\lim_{k\to\infty}\eta_k = 0$, and the coefficients are sufficiently regular. Let $u^* \in H^2(\Omega)$ be the solution to the HJBI equation.
Then for any initial guess $u^0 \in \mathcal{F}$, we have:

1. *$\{u^k\}_{k\in\mathbb{N}}$ converge superlinearly to u^* in $H^2(\Omega)$, i.e., $\lim_{k\to\infty} \|u^{k+1} - u^*\|_{H^2(\Omega)} / \|u^k - u^*\|_{H^2(\Omega)} = 0$.*
2. *$\lim_{k\to\infty}(u^k, \partial_i u^k, \partial_{ij} u^k)(x) = (u^*, \partial_i u^*, \partial_{ij} u^*)(x)$ for a.e. $x \in \Omega$, and for all $i, j = 1, \ldots, n$.*
3. *The control laws $\{(\alpha^k, \beta^k)\}_{k\in\mathbb{N}}$ converge to the optimal feedback control (α^*, β^*).*

Numerical experiments: Zermelo's Navigation Problem

The position of $\vec{X}_t^{x,\alpha} = (X_t^{x,\alpha}, Y_t^{x,\alpha})$ follows:

$$\begin{pmatrix} dX_t^{x,\alpha} \\ dY_t^{x,\alpha} \end{pmatrix} = \begin{pmatrix} v_c(X_t^{x,\alpha}, Y_t^{x,\alpha}) + v_s \cos(\alpha_t) \\ v_s \sin(\alpha_t) \end{pmatrix} dt + \begin{pmatrix} \sigma_x & 0 \\ 0 & \sigma_y \end{pmatrix} dW_t$$

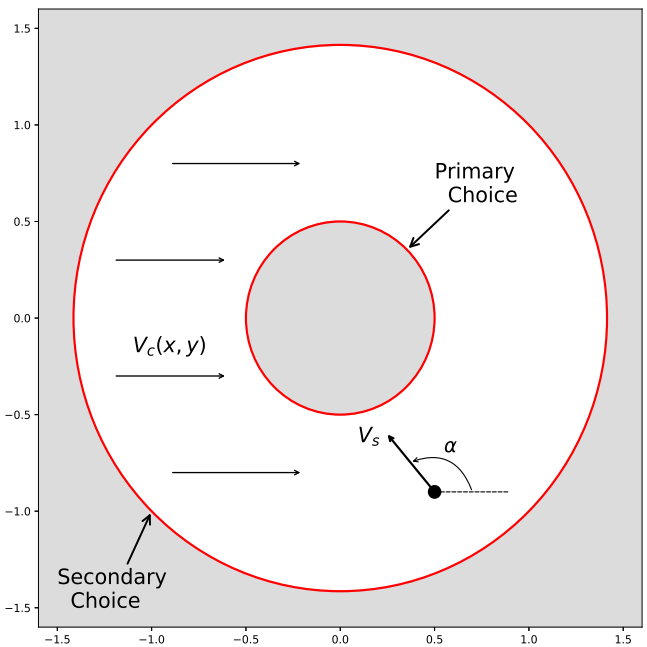


- The velocity of the wind is $v_c = 1 - 0.2 \sin\left(\pi \frac{x^2+y^2-r^2}{R^2-r^2}\right)$.
- $\sigma_x = 0.5$ and $\sigma_y = 0.2$.
- $r = 0.5$ and $R = \sqrt{2}$.

Numerical experiments: Zermelo's Navigation Problem

Let $\alpha_t \in \mathbb{A} = [0, 2\pi]$ be the direction of the boat. We consider

$$u(x) = \inf_{\alpha \in \mathcal{A}} \sup_{\mathbb{Q} \in \mathcal{M}} \mathbb{E}_{\mathbb{Q}} \left[\tau_{x,\alpha} + g(\vec{X}^{x,\alpha}_{\tau_{x,\alpha}}) \right].$$



- The first exit time is $\tau_{x,\alpha} := \inf\{t \geq 0 \mid \vec{X}^{x,\alpha}_t \notin \Omega\}$.
- $g \equiv 0$ on $\partial B_r(0)$ and $g \equiv 1$ on $\partial B_R(0)$.
- $\mathcal{M}$ denotes the uncertainty from the unknown law of the random perturbation.

Zermelo's Navigation Problem

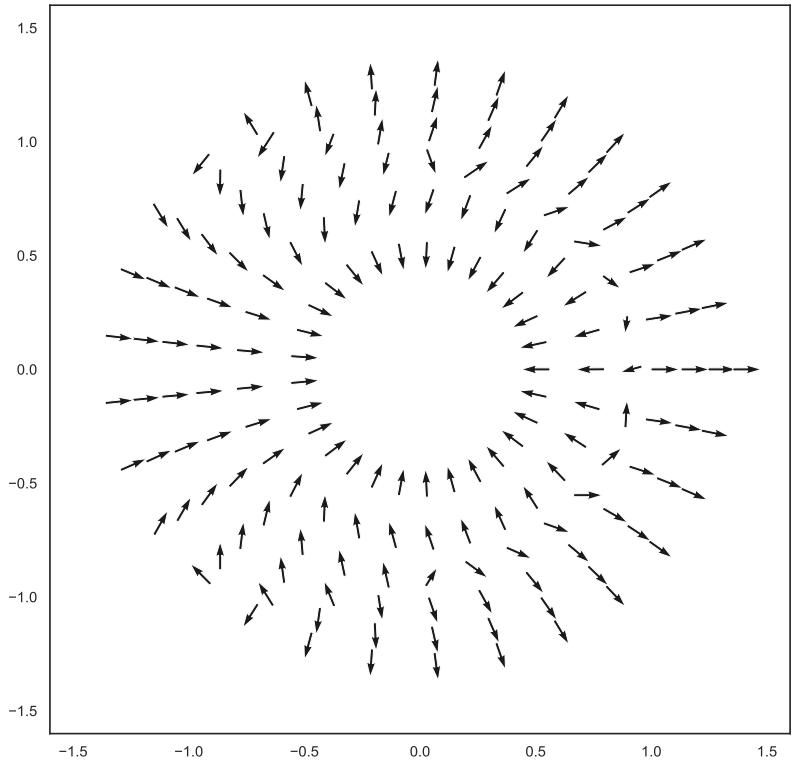


Figure: The feedback control for the scenario where the ship moves slower than the wind ($v_s = 0.5$).

Zermelo's Navigation Problem

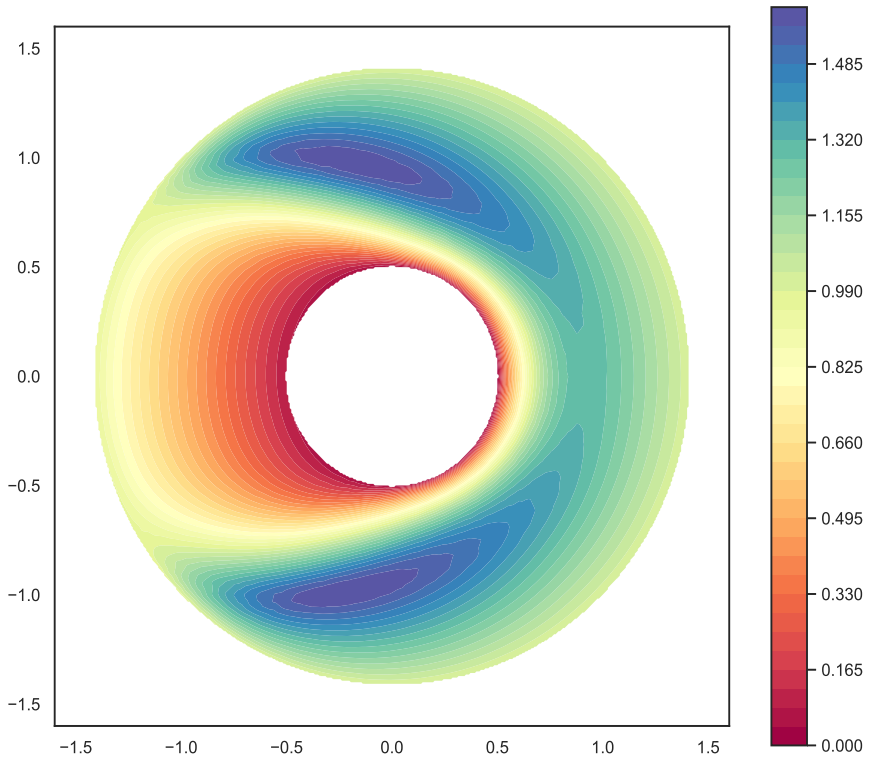


Figure: The value function for the scenario where the ship moves slower than the wind ($v_s = 0.5$).

Numerical experiments: Zermelo's Navigation Problem

UNIVERSITY OF OXFORD
Mathematical Institute

$$\text{Residual} := \| - a^{ij}\partial_{ij}u^k + H(\cdot, u^k, \nabla u^k)\|^2_{2,\Omega,\text{val}} + \|u^k - g\|^2_{3/2,\partial\Omega,\text{val}}.$$

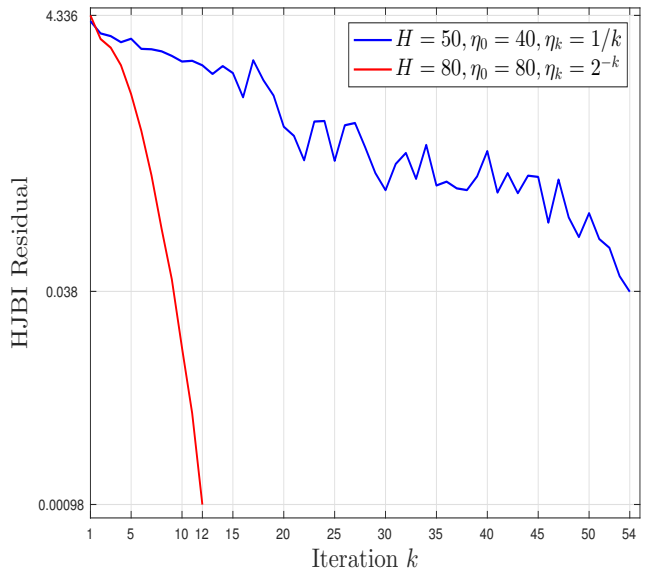


Figure: Residuals with respect to the number of policy iterations.

Conclusions

Iterative fitting for stochastic games with controlled drifts, using

- Sobolev spaces (for feedback controls);
- Newton methods (a.k.a. policy iteration) in function space;
- stochastic gradient descent (weakest link).

Extensions to finite-horizon games and optimal stopping problems.

Reference:

1. K. Ito, C. Reisinger, Y. Zhang. *A neural network based policy iteration algorithm with global H^2-superlinear convergence for stochastic games on domains*, arXiv:1906.02304.