TD3 for Competitive Air Hockey

RL Course WS 2025/26 — Final Project

Julian Jurcevic · Team: alphabet-td3

February 2026

TD3 — Three Key Ideas

1. Clipped Double Q-Learning

Two critics; target uses the *minimum*:

$$y = r + \gamma(1-d) \min_{i=1,2} Q_{\phi_i'}(s', \tilde{a}')$$

2. Target Policy Smoothing

$$\tilde{a}' = \mathrm{clip}\big(\pi_{\theta'}(s') + \epsilon,\ -c,\ c\big)$$

3. Delayed Actor Updates

- Critic: every step; Actor: every 2nd
- Polyak averaging: $\phi' \leftarrow \tau\,\phi + (1-\tau)\,\phi'$

Architecture

- Actor & twin critics: 2×256 (tanh)
- Actor: $a \in [-1,1]^4$
- Critics: concatenated (s,a) input

Three-Stage Curriculum with Self-Play

Stage I
10k episodes
100% Weak
0% Strong
0% Self-Play

Puck control

→

Stage II
25k episodes
45–55% Strong
40–45% Weak
0–10% Self-Play

Mixed opponents

→

Stage III
12k episodes
30–60% Strong
30–70% Weak
0–30% Self-Play

Generalization

Self-Play Pool: Snapshots every $k=150$ eps · Pool $N_{\text{pool}}=25$ · Difficulty-weighted sampling ($\times 1.2$ loss, $\times 0.95$ win)

Curriculum Training Progression

Training curves for a representative single-seed run (see report Figure 1).

Stage I
WR Weak: ~95%
WR Strong: ~30%
Overfits to weak

→

Stage II
WR Weak: ~95%
WR Strong: ~80%
Mixed schedule

→

Stage III
WR Weak: ~95%
WR Strong: ~85%
Competition-ready

- Without curriculum: strong win rate stays at ~30% (pure weak training)
- Staged scheduling resolves this while retaining high weak win rates

Ablation: Noise Comparison (3 seeds, Stage II)

Noise Type	WR Weak (%)	WR Strong (%)	Ret. Weak	Ret. Strong
Gaussian	92.5 ± 4.5	81.0 ± 0.5	8.22 ± 0.80	5.69 ± 0.10
Ornstein–Uhlenbeck	$\mathbf{94.7 \pm 0.6}$	$\mathbf{89.0 \pm 2.7}$	$\mathbf{8.56 \pm 0.13}$	$\mathbf{7.06 \pm 0.46}$
Pink	92.6 ± 4.3	86.1 ± 2.8	8.17 ± 0.57	6.40 ± 0.34
Uniform	91.2 ± 2.8	80.2 ± 8.4	8.10 ± 0.55	5.51 ± 1.51

- OU best: **+8%** vs. Gaussian against strong (temporal correlation $\rightarrow$ smoother trajectories)
- Pink noise second-best (also correlated) · Uniform: highest variance

Ablation: Self-Play & Prioritized Replay (3 seeds)

Variant	WR Weak (%)	WR Strong (%)	Ret. Weak	Ret. Strong
No PER, No SP	**93.1 $\pm$ 3.8**	**78.3 $\pm$ 3.1**	**8.33 $\pm$ 0.66**	**5.00 $\pm$ 0.70**
No PER, Self-Play	90.7 $\pm$ 5.9	72.6 $\pm$ 7.6	7.62 $\pm$ 1.26	4.06 $\pm$ 1.56
PER, No SP	75.8 $\pm$ 9.2	66.1 $\pm$ 4.7	4.22 $\pm$ 2.00	1.99 $\pm$ 1.04
PER, Self-Play	78.3 $\pm$ 2.2	65.3 $\pm$ 5.1	4.71 $\pm$ 0.54	1.78 $\pm$ 1.01

PER hurts performance

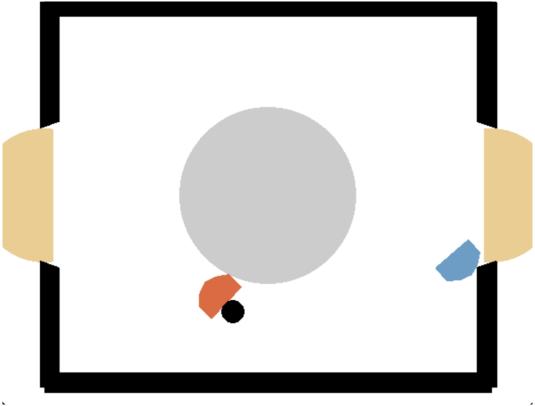
- Non-stationary opponents amplify priority variance
- $\approx$15–20% WR drop $\Rightarrow$ **not used**

Self-Play — trade-off

- Lower benchmark scores but **retained** for tournament robustness
- Pool diversity decreases as policies converge

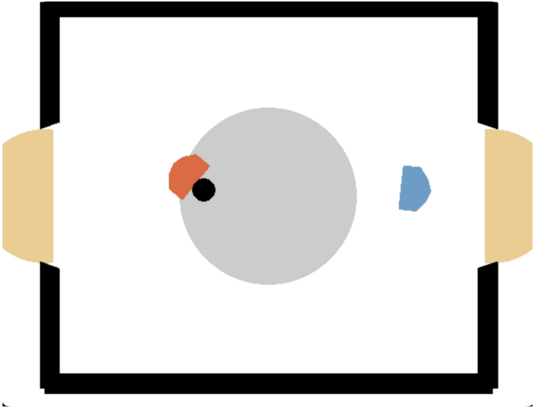
Gameplay Examples

vs. Weak Opponent



File: gameplay_weak.mov

vs. Strong Opponent



File: gameplay_strong.mov

Conclusion & Takeaways

What worked

- **Curriculum**: most impactful; prevents overfitting to single opponent
- **OU noise**: +8% WR (strong) via temporal correlation
- **Self-play**: retained for tournament generalization

What didn't work

- **PER**: ∼15% drop under non-stationarity

Final Agent Config

Algorithm	TD3
Noise	OU (annealed)
Replay	Uniform (300k)
Curriculum	3 stages
Self-Play	Pool of 25
WR Weak	∼95%
WR Strong	∼85%

Limitations

- Single-seed training curves
- Manual curriculum tuning
- Self-play pool convergence