Supplementary material

Recoverability-shielded reinforcement learning for far-range spacecraft rendezvous

This supplement contains the complete parameter and benchmark tables, exact-navigation reference comparisons, extended fault and stress-test results, training-stage comparisons and remaining-cost surface construction. Equation and figure references without an appendix letter refer to the main article.

A. Parameters and complete navigation comparisons

Table A.1 collects the task, controller and learning settings. Tables A.2 and A.3 retain all evaluated methods in the navigation sweep; Fig. A.1 shows their cost and success curves.

Table A.1: Numerical settings for the rendezvous task, navigation model, command checks and policy training. Fault ranges labelled "training" define the randomized training envelope; OOD and OOD-2 denote the extended test envelopes. Unmodelled faults change individual thruster efficiencies, their time evolution or their availability. Exact-navigation runs use zero state error and zero reported navigation uncertainty.

Group	Parameter	Value
Task	initial range, angle to $-y$, speed	9.5 km to 10.5 km, $\leq 60°$, ≤ 0.5 m/s
	thrust limit, control interval, time limit	0.02 m/s^2, 1 s, 12 h
	corridor radius and half angle; speed zone	200 m, 30°; 100 m, 1 m/s
	keep-out radius; handover	10 m; 100 m at ≤ 1 m/s
	time weight w of the performance index	0.33 m/s per hour
Faults	efficiency, misalignment, bias (training)	$[0.4, 1]$, $\leq 5°$, $\leq 2 \times 10^{-4}$ m/s^2
	efficiency, misalignment, bias (OOD)	$[0.25, 1]$, $\leq 8°$, $\leq 4 \times 10^{-4}$ m/s^2
	onset; share of faulty training missions	≤ 3 h; 80 %
	efficiency, misalignment, bias (OOD-2)	$[0.2, 1]$, $\leq 10°$, $\leq 5 \times 10^{-4}$ m/s^2
	unmodelled: thruster efficiency; degradation; outages	[0.4, 1] each; 0.5 h to 4 h; 10 % of time
Navigation	position error s.d. per axis	0.5 m + $f\rho$
	level f: test; training	0, 0.5, 1, 2, 3, 5 %; $\mathcal{U}[0, 4\,\%]$
	third RL stage: exact navigation; near-constraint starts	20 %; 15 % of missions
	error period, damping; accelerometer noise	2 h, 0.7; 2×10^{-5} m/s^2
Action	decision interval; maximum increment; dead band	300 s; 2 m/s; 0.02 m/s
Policy	encoder, GRU and head units; inputs	128, 128, 128; 34
Teacher	arrival grid; entry point; drift speed	0.75 h to 8 h; 300 m; 0.1 m/s to 0.8 m/s
	rear-plane margin; hysteresis; DOB time constant	30 m; 0.01 m/s; 60 s
	transfer-leg dead band	$\max(0.02\,\text{m/s}, 3\sqrt{3}\,\sigma_v)$
Imitation	rounds × missions; epochs (first, later, final)	9×512; 100, 30, 60
Shield	blend levels	$\{0, 0.25, 0.5, 0.75, 1\}$
	backup acceleration; assumed braking	0.015 m/s^2; 0.006 m/s^2
	prediction margin ε; position margin; reserve	2×10^{-4} m/s^2; 0.5 m; 0.5 m
	speed margin; disturbance time constant	0.02 m/s; 30 s
	navigation margin (imitation: $3\sigma_p$)	$4\sigma_p(\hat{\rho})$
RL	updates; parallel missions; decisions per update	304 + 304 + 305; 512; 16
	discount; GAE λ; clip; target KL	0.999; 0.95; 0.2; 0.02
	learning rates (actor, critic); imitation weight	10^{-4}, 10^{-3}; $0.1 \cdot 0.98^j$
	initial standard deviation (gate, increment)	0.3, 0.02
RL, Imp-RL-exact	updates; parallel missions; actor learning rate	200; 128; 3×10^{-5}
	increment standard deviation; imitation weight	0.004; $0.1 \cdot 0.99^j$

Table A.2: Performance index J (m/s, mean over successful missions) and success rate (%, in parentheses) over the navigation-error level, nominal missions, native Tudat; 200 missions per level and seed. Imp-RL and Imp-RL-exact pool three seeds; Imp-BC uses one. All methods use the recoverability shield with the margin of four standard deviations.

Method	exact	0.5 %	1 %	2 %	3 %	5 %
Imp-RL (proposed)	2.76 (100)	2.82 (100)	2.91 (100)	3.18 (100)	3.56 (99.2)	4.88 (91.7)
Imp-BC (imitation start)	3.16 (100)	3.07 (99.5)	3.59 (99.5)	4.04 (96.5)	4.61 (94.0)	6.36 (76.0)
Imp-RL-exact	2.93 (100)	3.64 (99.3)	4.49 (98.5)	6.19 (95.8)	7.86 (90.0)	11.47 (66.0)
Imp-CE-adapt-3σ (teacher)	2.56 (100)	2.93 (100)	3.73 (99.0)	3.83 (95.5)	4.63 (93.5)	6.50 (76.5)
Imp-CE-adapt-2σ	2.56 (100)	2.98 (100)	3.31 (99.0)	4.09 (98.5)	5.05 (96.5)	6.93 (87.5)
Imp-CE-adapt	2.56 (100)	3.61 (100)	4.53 (100)	6.37 (99.5)	8.24 (99.5)	12.23 (94.0)
Imp-CE	2.51 (100)	3.48 (89.5)	4.41 (87.5)	6.15 (86.0)	8.25 (84.5)	12.03 (64.5)
Imp-CE-axis	2.56 (100)	3.61 (100)	4.53 (100)	6.37 (99.5)	8.24 (99.5)	12.23 (94.0)
Imp-T3	2.59 (100)	3.53 (89.5)	4.71 (82.5)	6.97 (71.5)	9.65 (61.5)	14.92 (36.5)
Imp-T4	2.64 (100)	4.02 (98.0)	5.57 (95.5)	8.77 (81.0)	12.14 (61.0)	18.60 (35.0)
Imp-T5	2.94 (100)	4.66 (93.5)	6.65 (92.5)	10.62 (86.5)	14.78 (73.0)	23.06 (36.5)
Drift-CE	2.60 (100)	3.08 (100)	3.47 (100)	4.46 (98.5)	5.56 (96.5)	7.32 (88.0)
Drift-T2	2.98 (100)	3.37 (100)	3.82 (100)	6.02 (100)	11.41 (99.5)	7.96 (78.0)
Drift-T3	2.62 (100)	3.12 (100)	3.61 (100)	5.88 (99.5)	9.69 (97.0)	8.43 (73.5)
Drift-T4	2.72 (100)	3.32 (100)	3.95 (100)	5.34 (98.5)	8.03 (97.0)	9.51 (71.5)
Drift-T5	2.94 (100)	3.68 (100)	4.36 (100)	5.90 (99.5)	8.25 (96.5)	11.23 (73.0)
Drift-T6	3.22 (100)	4.10 (100)	4.94 (99.5)	6.76 (98.5)	8.82 (93.5)	12.70 (70.0)
Two-phase DOB-LQR	3.14 (100)	3.71 (100)	4.31 (100)	5.41 (92.5)	6.58 (75.5)	9.18 (35.5)
Glideslope	3.77 (100)	4.32 (100)	4.88 (100)	6.05 (100)	7.32 (99.5)	10.12 (96.0)
APF	4.34 (100)	4.82 (100)	5.77 (100)	8.02 (100)	10.42 (100)	15.44 (100)
Waypoint DOB-LQR	5.16 (100)	5.17 (100)	5.17 (100)	5.27 (100)	5.43 (99.5)	6.03 (90.0)

Table A.3: Performance index J (m/s) and success rate (%, in parentheses) over the navigation-error level, combined-fault missions, native Tudat; layout as in Table A.2.

Method	exact	0.5 %	1 %	2 %	3 %	5 %
Imp-RL (proposed)	3.60 (99.8)	3.66 (99.8)	3.75 (99.7)	4.13 (99.8)	4.66 (99.0)	6.22 (89.8)
Imp-BC (imitation start)	4.58 (99.5)	4.42 (100)	4.56 (99.0)	5.32 (94.5)	6.15 (88.0)	7.62 (64.0)
Imp-RL-exact	3.96 (97.8)	4.49 (95.3)	5.24 (95.0)	6.98 (93.8)	8.69 (90.7)	12.16 (70.5)
Imp-CE-adapt-3σ (teacher)	3.57 (100)	3.82 (100)	4.07 (99.5)	4.97 (96.0)	5.56 (87.5)	7.57 (65.5)
Imp-CE-adapt-2σ	3.57 (100)	3.87 (99.5)	4.27 (100)	5.06 (97.5)	5.95 (97.0)	8.06 (87.0)
Imp-CE-adapt	3.57 (100)	4.47 (100)	5.64 (100)	7.29 (99.0)	9.13 (97.0)	13.15 (91.5)
Imp-CE	3.64 (96.0)	4.44 (93.0)	5.39 (92.0)	7.08 (90.0)	8.82 (86.0)	12.76 (67.5)
Imp-CE-axis	3.77 (98.5)	4.68 (98.5)	5.55 (98.5)	7.26 (97.0)	9.12 (96.0)	12.97 (88.0)
Imp-T3	3.89 (95.5)	4.87 (92.0)	5.82 (89.5)	8.20 (83.5)	10.55 (77.0)	15.30 (47.0)
Imp-T4	4.49 (95.0)	5.50 (92.0)	6.78 (91.0)	9.70 (88.0)	13.05 (80.0)	19.25 (51.0)
Imp-T5	5.42 (95.0)	6.41 (89.5)	7.98 (88.5)	11.66 (84.5)	15.51 (79.0)	23.40 (56.5)
Drift-CE	3.95 (100)	4.31 (100)	4.72 (99.5)	6.62 (99.0)	9.22 (99.0)	9.16 (80.5)
Drift-T2	3.89 (100)	4.31 (100)	4.92 (100)	9.24 (99.0)	13.05 (97.0)	9.76 (77.0)
Drift-T3	3.98 (100)	4.50 (100)	5.08 (100)	7.71 (100)	10.96 (96.5)	10.90 (78.5)
Drift-T4	4.55 (100)	5.22 (100)	5.86 (99.5)	7.68 (99.5)	10.43 (94.5)	12.50 (76.0)
Drift-T5	5.27 (100)	6.06 (100)	6.93 (100)	8.89 (100)	11.91 (95.0)	15.31 (74.5)
Drift-T6	6.02 (100)	7.04 (100)	8.13 (100)	10.61 (100)	12.87 (92.0)	17.12 (74.0)
Two-phase DOB-LQR	5.06 (100)	5.67 (99.5)	6.30 (99.0)	7.68 (87.5)	8.87 (59.0)	11.74 (27.0)
Glideslope	6.12 (100)	6.74 (100)	7.43 (100)	8.91 (100)	10.53 (100)	14.06 (96.0)
APF	5.84 (100)	6.37 (100)	7.40 (100)	9.90 (100)	12.61 (100)	18.28 (100)
Waypoint DOB-LQR	6.58 (100)	6.60 (100)	6.65 (100)	6.76 (100)	6.94 (100)	7.57 (95.5)

Figure A.1 compares Imp-RL with one classical method per panel, ordered by the latter's nominal J at 2 % navigation error. The cost curves use a shared logarithmic scale and normal 95 % intervals from pooled mission records. Zero error denotes exact navigation.

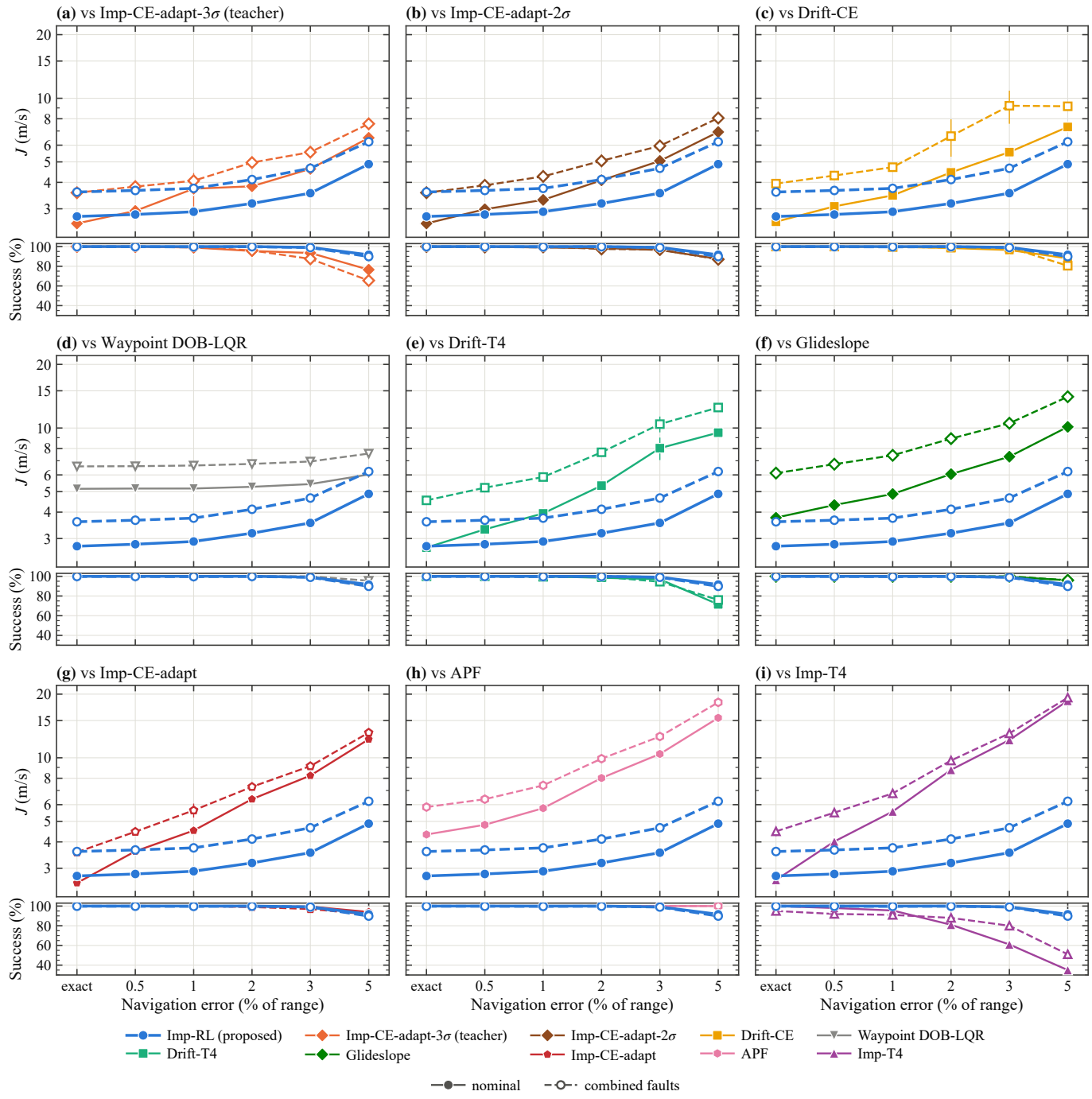


Figure A.1: Pairwise navigation-error comparisons: successful-mission mean J above and success rate below. Solid lines and filled markers denote nominal missions; dashed lines and open markers denote combined faults.

A.1. Fault scenarios with navigation errors

The single-fault and OOD scenarios test whether the navigation-error result persists across fault mechanisms (Table A.4); glideslope guidance and the artificial potential function are included in these comparisons. With navigation errors the learned guidance has the lowest performance index of all methods flown in these scenarios, in every scenario and at every level. Paired with its teacher, it lowers the performance index by 8.3, 17.8 and 19.0 % at 1, 2 and 3 % with efficiency faults, by 7.5, 17.1 and 26.4 % with misalignment, by 15.0, 18.9 and 22.1 % with bias and by 3.6 to 7.0 % in the OOD scenario. Relative to Imp-CE-adapt-2σ the reductions are 2.3 to 29.3 %, relative to Imp-CE-adapt 18.6 to 56.8 %, relative to Drift-CE 17.5 to 38.6 % and relative to the feedback guidance laws 21.0 to 66.2 %. The pooled-pair bootstrap 95 % intervals lie below zero in the single-fault comparisons. In OOD, the intervals include zero against the teacher at 1 and 3 % and against Imp-CE-adapt-2σ at 1 %. In each of these twelve cases it also uses

less propellant than every classical design, and it arrives 0.19 to 0.41 h earlier than its teacher in the single-fault scenarios and within 0.07 h of it in the OOD scenario.

The learned guidance completes 99.7 to 100 % of the single-fault missions at every level, whereas its teacher completes 90.5 to 94.5 % of them at 3 %. In the OOD scenario, whose faults exceed the training ranges, it completes 92.8, 93.5 and 90.7 % of the missions at 1, 2 and 3 %, about as many as its teacher (96.0, 93.0 and 87.5 %) and fewer than Drift-CE (96.0, 97.5 and 94.0 %) and the feedback guidance laws, which complete 99.5 to 100 % of them at 1.3 to 2.2 times its performance index.

With exact navigation the certainty-equivalent planners are again the most economical designs: paired with Imp-CE-adapt-3σ, the performance index of the learned guidance is 9.8 % higher with efficiency faults, 7.5 % higher with misalignment and 1.7 % higher with bias, and in the OOD scenario the difference of +3.6 % has a 95 % interval from −1.0 to +8.5 %.

Table A.4: Performance index J (m/s, mean over successful missions) and success rate (%, in parentheses) in the single-fault scenarios and the OOD scenario with exact navigation and at navigation errors of 1 to 3 % of the range, native Tudat.

Scenario	Navigation error	**Imp-RL**	Imp-CE-adapt-3σ (teacher)	Imp-CE-adapt-2σ	Imp-CE-adapt	Drift-CE	Waypoint DOB-LQR
Efficiency	exact	3.22 (100)	2.93 (100)	2.93 (100)	2.93 (100)	3.07 (100)	5.94 (100)
	1 %	3.37 (99.8)	3.67 (99.0)	3.85 (100)	5.06 (99.5)	4.23 (100)	5.98 (100)
	2 %	3.76 (100)	4.57 (97.0)	4.58 (98.5)	7.15 (99.0)	5.46 (98.5)	6.10 (100)
	3 %	4.26 (99.7)	5.21 (90.5)	5.61 (98.0)	9.01 (98.0)	6.90 (96.5)	6.29 (100)
Misalignment	exact	2.77 (100)	2.58 (100)	2.58 (100)	2.58 (100)	2.63 (100)	5.18 (100)
	1 %	2.91 (100)	3.14 (97.5)	3.36 (99.5)	4.57 (100)	3.53 (99.5)	5.23 (100)
	2 %	3.18 (100)	3.81 (94.0)	4.09 (98.5)	6.36 (99.5)	4.33 (98.0)	5.35 (100)
	3 %	3.56 (99.7)	4.81 (92.0)	5.00 (97.5)	8.20 (97.0)	5.49 (96.5)	5.60 (100)
Bias	exact	3.04 (100)	2.99 (100)	2.99 (100)	2.99 (100)	3.10 (100)	5.60 (100)
	1 %	3.16 (100)	3.72 (99.5)	3.71 (100)	4.93 (100)	3.89 (100)	5.63 (100)
	2 %	3.42 (100)	4.20 (97.0)	4.41 (98.0)	6.61 (99.5)	5.07 (99.5)	5.72 (100)
	3 %	3.82 (100)	4.91 (94.5)	5.39 (95.5)	8.43 (98.0)	6.15 (97.5)	5.93 (99.5)
OOD	exact	5.42 (95.7)	5.18 (98.5)	5.18 (98.5)	5.18 (98.5)	6.19 (97.5)	8.72 (100)
	1 %	5.57 (92.8)	5.85 (96.0)	5.77 (96.5)	6.85 (96.5)	7.46 (96.0)	8.74 (100)
	2 %	6.10 (93.5)	6.53 (93.0)	7.02 (96.5)	9.08 (94.0)	8.09 (97.5)	8.90 (100)
	3 %	7.04 (90.7)	7.37 (87.5)	7.36 (89.0)	10.98 (87.0)	9.60 (94.0)	9.11 (100)

B. Exact-navigation reference

The main article presents the trajectory and performance distributions in Section 5.4. This appendix gives the complete numerical comparisons. With exact navigation, this appendix analyses the guidance trained with exact navigation, Imp-RL-exact, whose teacher is Imp-CE; in these conditions the guidance trained with navigation errors has a 4.6 % lower performance index nominally and a 10.0 % lower one under combined faults (Section 5.1). Table B.1 compares the methods on the performance index (10) in the six fault scenarios, and Table B.2 pairs every method with the fixed-time drift design Drift-T4 on identical missions. Table B.1 also lists the guidance trained with navigation errors, Imp-RL, which in every scenario arrives 0.34 h to 0.57 h earlier than Imp-RL-exact on the same or less propellant and completes 95.7 % of the OOD missions against 83.0 %. Imp-RL-exact reaches the handover in 100 % of the nominal, efficiency and misalignment missions, 99.3 % of the bias missions and 97.8 % under combined faults, with mean flight times of 3.3 h to 3.5 h.

Against the guidance laws that fly the corridor with feedback, Imp-RL-exact is both more economical and faster in every scenario. Nominally it uses 1.78 m/s against 2.38 m/s for glideslope guidance, 3.00 m/s for the artificial potential function and 3.79 m/s for waypoint DOB-LQR (25–53 % less), and it arrives 0.55 h to 0.7 h earlier; across the scenarios their performance index is 28–76 % higher than that of Imp-RL-exact. In the OOD scenario these slower feedback laws complete all missions, against 83.0 % for Imp-RL-exact.

Fixed-time CW targeting spans a trade-off between propellant and flight time through its transfer time. Against the four-hour transfers Drift-T4 and Imp-T4, the default of the classical designs, Imp-RL-exact arrives 0.7 h to 1.0 h earlier in every scenario. Under bias and combined faults it lowers the performance index relative to Drift-T4 by 6.2 % and 11.7 % (paired), and it completes more missions than fixed-time impulsive targeting, whose open-loop burns time out in 10.5 % of the bias missions (99.3 % against 89.5 % success). Nominally the four-hour transfers fly slower on less propellant, and the performance index of Imp-RL-exact is 7.9 % higher than that of Drift-T4. The three-hour transfers Drift-T3 and Imp-T3 arrive up to 0.34 h earlier than Imp-RL-exact and on less propellant in most scenarios; Imp-RL-exact uses less propellant than Drift-T3 under combined faults and completes more missions than Imp-T3 under bias, combined and OOD faults.

The certainty-equivalent planner reaches the handover in the same time as Imp-RL-exact; Imp-RL-exact uses 12–30 % more propellant and comes within 8–17 % of its performance index. Its completion rates track those of its teacher (99.3 % against 98.0 %

under bias, 97.8 % against 96.0 % under combined faults, 83.0 % against 83.5 % in the OOD scenario), and under combined faults it comes within 0.5 % of the certainty-equivalent drift design Drift-CE. The two extensions of the planner raise its reliability under faults: with the corridor re-entry rule and a per-axis efficiency estimate, Imp-CE-adapt completes all in-distribution missions and 98.5 % of the OOD missions, and Imp-RL-exact comes within 12.5–16.9 % of its performance index in distribution.

Table B.1: Performance index $J = \Delta v + w\,T$ (m/s, mean over successful missions) and success rate per scenario with exact navigation in native Tudat; 200 missions per scenario and seed, learned methods pool three seeds. Imp-RL is the proposed guidance, trained with navigation errors, and Imp-RL-exact the guidance trained with exact navigation. All methods use the recoverability shield.

Method	Nominal	Efficiency	Misalign.	Bias	Combined	OOD
Imp-RL (proposed)	2.76 (100)	3.22 (100)	2.77 (100)	3.04 (100)	3.60 (99.8)	5.42 (95.7)
Imp-RL-exact	2.93 (100)	3.38 (100)	2.96 (100)	3.49 (99.3)	3.96 (97.8)	5.85 (83.0)
Imp-BC-exact (imitation only)	2.99 (100)	3.32 (100)	3.15 (100)	3.73 (98.5)	4.05 (98.0)	5.85 (80.0)
Imp-CE (teacher of Imp-RL-exact)	2.51 (100)	2.99 (100)	2.61 (100)	2.97 (98.0)	3.64 (96.0)	5.31 (83.5)
Imp-CE-axis	2.56 (100)	3.03 (100)	2.58 (100)	2.99 (100)	3.77 (98.5)	5.89 (90.5)
Imp-CE-adapt	2.56 (100)	2.93 (100)	2.58 (100)	2.99 (100)	3.57 (100)	5.18 (98.5)
Imp-T3	2.59 (100)	3.04 (100)	2.80 (100)	3.04 (86.5)	3.89 (95.5)	5.43 (70.0)
Imp-T4	2.64 (100)	3.04 (100)	2.72 (100)	3.56 (89.5)	4.49 (95.0)	6.37 (65.5)
Imp-T5	2.94 (100)	3.24 (100)	3.16 (99.5)	4.26 (89.0)	5.42 (95.0)	7.98 (63.0)
Drift-CE	2.60 (100)	3.07 (100)	2.63 (100)	3.10 (100)	3.95 (100)	6.19 (97.5)
Drift-T2	2.98 (100)	3.57 (100)	3.01 (100)	3.25 (100)	3.89 (100)	4.84 (99.5)
Drift-T2.5	2.70 (100)	3.24 (100)	2.74 (100)	3.13 (100)	3.81 (100)	5.16 (99.5)
Drift-T3	2.62 (100)	3.13 (100)	2.66 (100)	3.23 (100)	3.98 (100)	5.73 (99.5)
Drift-T3.5	2.64 (100)	3.09 (100)	2.67 (100)	3.45 (100)	4.23 (100)	6.45 (100)
Drift-T4	2.72 (100)	3.14 (100)	2.74 (100)	3.73 (100)	4.55 (100)	7.24 (100)
Drift-T5	2.94 (100)	3.33 (100)	2.97 (100)	4.33 (100)	5.27 (100)	8.85 (100)
Drift-T6	3.22 (100)	3.59 (100)	3.23 (100)	4.96 (100)	6.02 (100)	10.50 (100)
Glideslope	3.77 (100)	4.80 (100)	3.79 (100)	4.75 (100)	6.12 (100)	8.98 (100)
APF	4.34 (100)	5.36 (100)	4.30 (100)	4.65 (100)	5.84 (100)	8.13 (100)
Waypoint DOB-LQR	5.16 (100)	5.94 (100)	5.18 (100)	5.60 (100)	6.58 (100)	8.72 (100)
Two-phase DOB-LQR	3.14 (100)	3.60 (100)	3.14 (100)	4.17 (100)	5.06 (100)	8.40 (100)

Table B.2: Paired difference of the performance index to Drift-T4 (%, with pooled-pair bootstrap 95 % interval), over missions both methods complete, in native Tudat.

Method	Nominal	Efficiency	Misalign.	Bias	Combined	OOD
Imp-RL-exact	+7.9 [+6.7, +9.3]	+7.6 [+6.4, +8.9]	+8.0 [+7.0, +9.2]	−6.2 [−7.6, −4.7]	−11.7 [−13.8, −9.2]	−10.2 [−15.8, −3.4]
Imp-BC-exact (imitation only)	+9.9 [+7.5, +12.8]	+5.7 [+3.9, +7.5]	+14.7 [+10.1, +20.5]	+0.4 [−4.2, +6.1]	−10.2 [−13.3, −7.0]	−10.3 [−18.3, −0.7]
Imp-CE (teacher of Imp-RL-exact)	−7.6 [−8.3, −6.8]	−4.8 [−6.2, −3.4]	−4.7 [−6.9, −1.8]	−19.8 [−22.4, −17.0]	−18.4 [−21.3, −15.3]	−18.2 [−26.0, −8.0]
Imp-CE-axis	−5.9 [−6.9, −5.0]	−3.5 [−5.0, −2.0]	−6.1 [−7.1, −5.2]	−19.7 [−21.4, −18.0]	−16.4 [−18.9, −13.8]	−13.5 [−20.3, −5.3]
Imp-CE-adapt	−5.9 [−6.9, −5.0]	−6.6 [−7.6, −5.7]	−6.0 [−6.9, −5.1]	−19.7 [−21.4, −18.0]	−21.4 [−23.6, −19.4]	−27.1 [−31.8, −22.0]
Imp-T3	−4.8 [−7.1, −2.1]	−3.3 [−5.2, −1.5]	+2.2 [−2.5, +7.2]	−16.8 [−18.5, −14.9]	−13.0 [−15.3, −10.5]	−9.3 [−17.9, +3.7]
Imp-T4	−2.8 [−4.0, −0.9]	−3.2 [−3.9, −2.5]	−0.9 [−3.2, +2.0]	−3.0 [−3.3, −2.7]	+0.5 [−1.6, +3.7]	+10.2 [+3.5, +19.8]
Drift-CE	−4.2 [−4.7, −3.6]	−2.2 [−2.9, −1.5]	−4.0 [−4.5, −3.4]	−17.0 [−18.4, −15.5]	−13.2 [−14.9, −11.6]	−13.3 [−16.4, −10.1]
Drift-T3	−3.6 [−4.2, −3.0]	−0.3 [−1.1, +0.6]	−3.2 [−3.8, −2.6]	−13.4 [−14.3, −12.4]	−12.4 [−13.7, −11.1]	−20.6 [−22.8, −18.4]
Glideslope	+38.6 [+37.6, +39.5]	+52.8 [+50.1, +55.5]	+38.1 [+37.1, +39.1]	+27.5 [+26.6, +28.4]	+34.6 [+32.8, +36.4]	+24.1 [+22.3, +26.0]
APF	+59.5 [+57.6, +61.4]	+70.7 [+67.5, +73.9]	+56.7 [+55.0, +58.6]	+24.7 [+22.3, +27.0]	+28.5 [+25.9, +31.2]	+12.4 [+9.0, +15.9]
Waypoint DOB-LQR	+89.8 [+88.0, +91.4]	+89.2 [+86.2, +92.2]	+88.6 [+86.9, +90.3]	+50.2 [+47.6, +52.7]	+44.8 [+41.6, +47.7]	+20.5 [+16.6, +24.5]

The fault features connect the observed execution error to the next burn. Fig. B.1 follows one combined-fault test mission of Imp-RL-exact through the onset of its fault, which sets in after 1.42 h with axis efficiencies of 0.46, 0.55 and 0.94, a misalignment of 2.4° and a bias of 1.27×10^{-4} m/s^2. At the first decision after the onset, the mean coasting innovation observed by the policy tracks the hidden bias, and the policy fires a 0.38 m/s correction. During the following burns its thrust-error feature settles near the

efficiency loss of the firing axes, and the policy keeps trimming the arc with small burns. It reaches the handover after 2.71 h with 2.55 m/s, 0.37 h earlier than its teacher on 4 % more propellant; the CW replay reproduces the Tudat results of both methods on this mission. Panel (c) compares the coasting-innovation components, scaled by 2×10^{-4} m/s^2, with the hidden biases; panel (d) compares the scalar thrust-error feature with the hidden per-axis efficiency losses $\eta_i - 1$. These hidden quantities appear as dotted reference lines.

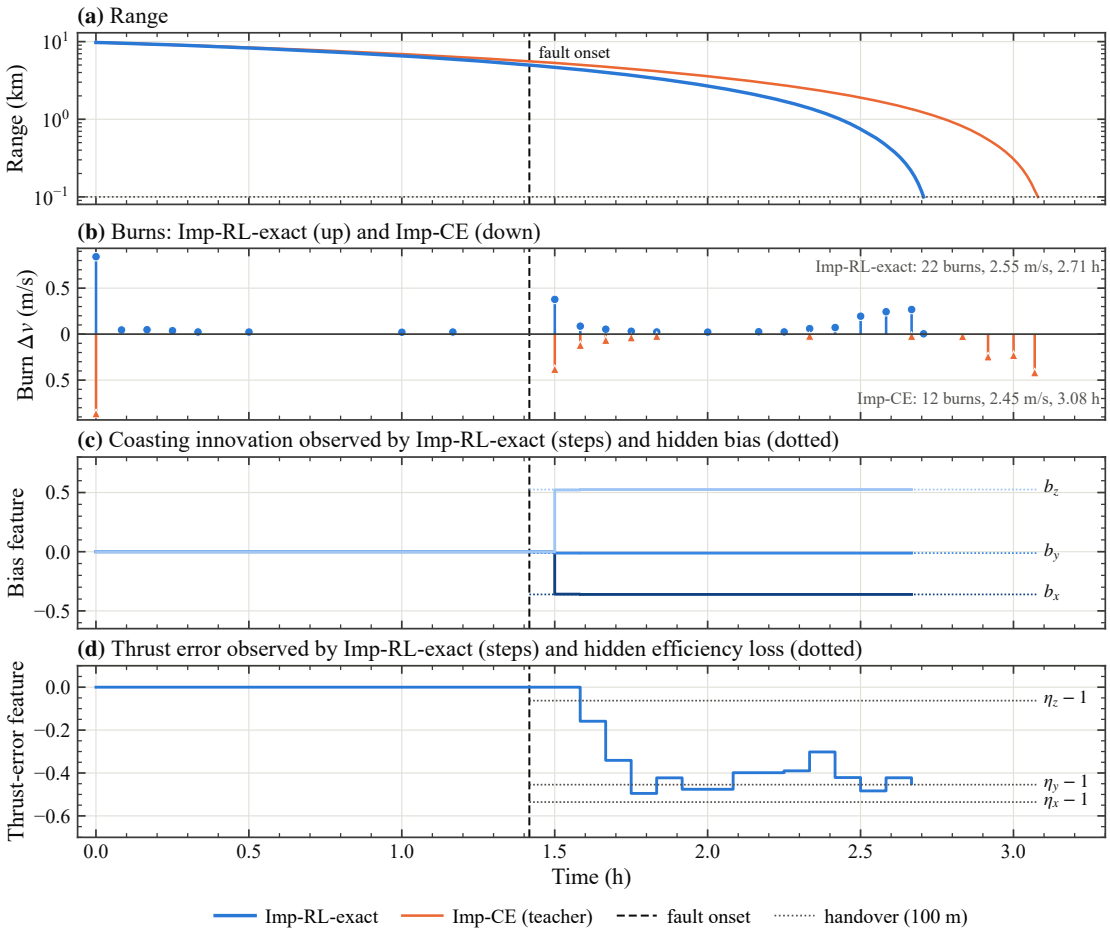


Figure B.1: Combined-fault response in a CW replay: (a) range, (b) burn increments, (c) coasting innovation and (d) thrust-error features. The dashed line marks fault onset.

B.1. Paired imitation comparison

Table B.3: Paired difference of Imp-RL-exact to its imitated policy before reinforcement learning with exact navigation, in native Tudat (mean and pooled-pair bootstrap 95 % interval).

Scenario	Δ propellant (%)	Δ flight time (h)	ΔJ (%)
Nominal	+2.1 [+0.3, +3.8]	−0.27 [−0.31, −0.23]	−1.8 [−3.1, −0.5]
Efficiency	+6.0 [+4.6, +7.4]	−0.20 [−0.23, −0.17]	+1.8 [+0.9, +2.8]
Misalign.	−3.4 [−6.9, −0.2]	−0.36 [−0.42, −0.30]	−5.8 [−8.5, −3.5]
Bias	−6.3 [−10.3, −3.1]	−0.28 [−0.33, −0.23]	−6.7 [−9.8, −4.2]
Combined	+0.3 [−2.5, +3.4]	−0.21 [−0.25, −0.16]	−1.5 [−3.8, +1.0]
OOD	+1.5 [−3.7, +7.0]	−0.12 [−0.18, −0.07]	+0.5 [−4.0, +4.9]

C. Extended fault and stress-test results

Two further test families extend the evaluation to more severe fault magnitudes and changing actuator behaviour; both are flown with exact navigation. (i) *More severe faults*: all three faults of (3) with $\eta_i \geq 0.2$, angles up to 10° and $\|\boldsymbol{b}\| \leq 5 \times 10^{-4}$ m/s^2 (OOD-2).

(ii) *Unmodelled fault types*, each with the misalignment and bias of the training envelope: each of the six thrusters (± per axis) has its own efficiency in [0.4, 1], so that the axis efficiency depends on the firing direction (Thruster); the efficiencies degrade linearly from 1 at the onset to their final value over 0.5 h to 4 h (Degradation); or all thrust is lost intermittently, 10 % of the time in outages of 60 s mean duration (Outage). The scenario Unmodelled combines the three with the OOD magnitudes. All scenarios reuse the test missions and faults. Imp-RL-W imitates Imp-CE-adapt and trains with exact navigation on the wider OOD fault ranges.

Table C.1: Exact-navigation tests beyond the training envelope in native Tudat. Entries give mean J (m/s) over successful missions and success rate (%, in parentheses); 200 missions per scenario and seed, with three seeds for each learned method.

Method	OOD	OOD-2	Thruster	Degradation	Outage	Unmodelled
Imp-RL (proposed)	5.42 (95.7)	6.54 (88.5)	3.67 (99.7)	3.36 (100)	4.01 (95.8)	4.62 (89.8)
Imp-RL-exact	5.85 (83.0)	7.49 (63.7)	4.11 (97.0)	3.87 (99.0)	4.14 (91.3)	4.61 (73.7)
Imp-RL-W (wide faults)	6.35 (96.5)	8.22 (93.5)	4.43 (100)	3.90 (100)	4.95 (97.5)	6.41 (89.0)
Imp-CE	5.31 (83.5)	5.44 (68.0)	3.81 (98.5)	3.37 (98.0)	3.73 (87.0)	3.87 (74.5)
Imp-CE-axis	5.89 (90.5)	6.27 (76.5)	4.05 (100)	3.43 (99.5)	3.99 (91.0)	4.38 (79.0)
Imp-CE-adapt	5.18 (98.5)	6.46 (96.5)	3.77 (100)	3.36 (100)	4.06 (95.0)	5.37 (92.5)
Imp-T4	6.37 (65.5)	7.79 (55.0)	4.80 (94.0)	4.28 (95.0)	4.61 (90.0)	6.23 (62.5)
Drift-CE	6.19 (97.5)	8.35 (94.5)	4.06 (100)	3.55 (100)	4.31 (97.5)	5.05 (96.0)
Drift-T3	5.73 (99.5)	7.18 (99.5)	4.11 (100)	3.72 (100)	4.32 (99.5)	5.42 (97.5)
Drift-T4	7.24 (100)	9.17 (100)	4.70 (100)	4.40 (100)	4.92 (100)	7.26 (99.5)
Glideslope	8.98 (100)	10.92 (100)	6.24 (100)	5.74 (100)	6.67 (100)	8.71 (99.5)
APF	8.13 (100)	9.59 (100)	5.87 (100)	5.37 (100)	6.51 (100)	7.70 (100)

The audit records zero violations through termination in these additional scenarios and in the stress tests, with exact navigation or 2 % error. For Imp-RL, the 3000 runs beyond the training envelope and 3000 stress runs bring the audited test total to 22 800.

In the severe and unmodelled fault tests (Fig. 11a, b, Table C.1), the proposed guidance completes 99.7 %, 100 % and 95.8 % of the Thruster, Degradation and Outage missions and 89.8 % of the Unmodelled missions. Paired with Imp-CE-adapt, its performance index is 2.5 % lower in the Thruster scenario, within 1 % of it in the Degradation and Outage scenarios, whose 95 % intervals include zero, and 12.5 % lower in the Unmodelled scenario. Under the more severe faults of OOD-2 it completes 88.5 % of the missions, against 96.5 % for Imp-CE-adapt and 63.7 % for Imp-RL-exact, whose teacher has no efficiency estimate; relative to Imp-RL-exact, it lowers the performance index by 9.3 to 24.7 % in four of the five scenarios and by 4.1 % under Outage. Without retraining, Imp-RL-exact completes 97.0 %, 99.0 % and 91.3 % of the Thruster, Degradation and Outage missions, with a performance index within 3–16 % of Imp-CE-adapt on the missions both complete. Trained on the wider envelope, Imp-RL-W completes all Thruster and Degradation missions, 97.5 % of the Outage and 89.0 % of the Unmodelled missions, and it raises the success rate from 83.0 % to 96.5 % in the OOD scenario and from 63.7 % to 93.5 % in OOD-2. It also completes the stress tests more often, 100 % instead of 52 % of the missions that start behind the rear plane and 100 % instead of 78 % of those outside the cone, while its in-distribution performance index is 0.7 % higher nominally and 6.8 % higher under combined faults. In the Thruster and Degradation scenarios the proposed guidance and Imp-CE-adapt are the most economical designs. Outages disturb the efficiency estimate of Imp-CE-adapt: under Outage the non-adaptive Imp-CE has a 6.8 % lower performance index on the missions both complete but completes only 87.0 % of the missions, and in the Unmodelled scenario Drift-CE completes more missions (96.0 % against 92.5 %) at an 8.2 % lower performance index.

The stress tests in Fig. 11c, d start next to a constraint; the third training stage draws 15 % of its missions from the same distributions with other seeds. With exact navigation, the proposed guidance completes 100 % of the corridor tests, 99.3 % of those with faults and 100 % of the rear-plane tests, against 78.0, 71.7 and 52.3 % for Imp-RL-exact; Imp-CE-adapt, Drift-CE, glideslope guidance, the artificial potential function and waypoint DOB-LQR complete all of them. At 2 %, the shield finds no admissible increment for the first navigation solution in 2 and 3 % of the corridor missions, 17 % of the rear-plane missions and 11 and 15 % of the speed-zone missions, which therefore end at the first step for every method. The proposed guidance completes 97.3 % of the corridor missions, 95.3 % of those with faults and 82.0 % of the rear-plane missions, as many as Imp-CE-adapt-3σ, Drift-CE, glideslope guidance and waypoint DOB-LQR within two percentage points (94 to 98 % and 80 to 83 %); the artificial potential function completes 69 to 73 %, Imp-CE and Imp-T4 25 to 56 % and Imp-RL-exact 0 to 25.7 %. Beyond the missions that end at the first step, 10 of its 900 corridor and rear-plane missions at 2 % fail, 7 by timeout and 3 when the shield finds no admissible increment later in the approach. In the speed-zone tests the methods complete 99 to 100 % of the missions with exact navigation and 85 % and 88 to 89 % at 2 %.

D. Training-stage comparisons

In Fig. 5, shaded intervals denote critic warm-up, and the success and violation histories in panels (b, c) are smoothed over nine updates.

Development-set cost falls through the three training stages (Fig. 5a). Across the three seeds, mean cost decreases from 4.33 m/s after imitation to 3.52–3.74 m/s at the selected stage-2 checkpoints, compared with 4.15 m/s for the teacher; stage 2 contributes 0.17–0.22 m/s of this reduction. On the development missions drawn from the stage-3 mixture, the range of seed-wise mean costs decreases from 3.79–4.05 m/s to 3.36–3.45 m/s, compared with 3.64 m/s for the teacher. The exact-navigation development cost difference changes from +17.4 to +7.3 % nominally and from +7.4 to −1.5 % under combined faults. At 2 %, development stress-test completion rises from 69.3, 73.7 and 46.7 % to 98.0, 96.5 and 72.8 % for the corridor, faulty-corridor and rear-plane starts. Across 0.5–5 %, paired cost differences from the teacher change by at most 2.6 percentage points; combined-fault completion at 5 % changes from 92.7 to 89.7 %.

The exploration-setting comparison shows larger cost gains with the increased increment noise and actor learning rate. On development missions at 1–5 % navigation error, increment noise 0.004 keeps the policy within a few percent of its teacher. With noise 0.02 and actor learning rate 10^{-4}, cost is 6.5–14.7 % below the teacher after stage 1, 8.5–21.8 % below after stage 2 and 8.8–24.4 % below after stage 3.

E. Remaining-cost surface construction

Figure 4 uses 300 nominal CW missions at 2 % navigation error. For each method, the surfaces retain successful missions whose total cost is at most three times that method's successful-mission median. Samples are taken every 30 s from 120 s. At each sample, J_{go} is the subsequently spent propellant plus the weighted remaining flight time. Gaussian smoothing produces a position-conditioned mean over the retained trajectory samples.

The plotted domain requires local support from at least ten distinct missions and retains the largest connected component, with enclosed holes filled. Each surface thus summarizes the positions and histories visited by its own method. Panels follow the ordering of nominal Tudat costs at 2 % navigation error. Dark traces project example trajectories from their first plotted positions (squares) to handover (stars).