A EXPERIMENTAL DETAILS

We used Trieste (Berkeley et al., 2022), Tensorflow (Abadi et al., 2015), and GPFLow (Matthews et al., 2017) to build our work and perform comparisons using MushroomRL (D'Eramo et al., 2021), MultiagentMuJoCo (de Witt et al., 2020), OpenAI Gym (Brockman et al., 2016), and Multi-agent Particle environment (Lowe et al., 2017). When comparing with related work, we used neural network policies of equivalent size. All of our tested policies are < 500 parameters, however the XL models are constructed using 3 layers of 400 neurons each.

To estimate the Hessian, we used the Hessian-Vector product approximation. We relaxed the discrete portions of our metamodel policy into differentiable continuous approximation for this phase using the Sinkhorn-Knopp algorithm for the Role Assignment phase. For role interaction network connectivity, we used a sigmoid to create differentiable "soft" edges between each role. We pragmatically kept all detected edges in the Hessian while maintaining computational feasibility. We observed that our approach could support up to 1500 edges in the dependency graph prior to experiencing computational intractability. We used the Matern-$\frac{5}{2}$ as the base kernel in all our models.

A.1 ABLATION AND INVESTIGATION

In the ablation, we perform experiments on MultiagentMuJoCo with environments Multiagent Ant with 6 segments, Multiagent Swimmer with 6 segments, Predator Prey with 3 predators, and Heterogeneous Predator Prey with 3 predators. In the Predator Prey environment, multiple predators must work together to capture faster and more agile prey. In Heterogeneous Predator Prey, each Predator has differing capabilities of speed and acceleration. This modification is challenging as a policy must not only coordinate between the Predators, but roles based specialization must be considered given the heterogeneous nature of each predator's capabilities.

To generate Fig. 7, we examined policy for Multiagent Ant with 6 agents for the role based policy specialization. The policy modulation plots were generated by examining the PredPrey and Het. PredPrey environments respectively.

A.2 COMPARISON WITH MARL

For the MARL setting, we compare against MADDPG (Lowe et al., 2017), FACMAC (Peng et al., 2021), COMIX (Peng et al., 2021), RODE (Wang et al., 2021b) and CDS (Li et al., 2021) using QPLEX (Wang et al., 2021a) as a base algorithm. We also compare against Comm-MARL approaches SOG (Shao et al., 2022), and G2ANet (Liu et al., 2020). RODE and QPLEX are limited to discrete environments, thus we are unable to provide comparisons on continuous action space tasks such as Multiagent Ant or Multiagent Swimmer. All MARL environments were trained for $2,000,000$ timesteps. The neural network policies were 3-layers each with 15 neurons per layer, and were greater than or equal to the size of the compared Metamodel policy. For Actor-Critic approaches, we did not reduce the size or expressivity of the critic. All used hyperparameters and Algorithmic configurations were as advised by the authors of the work.

In the MARL setting we use Multiagent Ant, Multiagent Swimmer, Predator-Prey, Heterogeneous Predator-Prey. Multiagent Ant, and Multiagent Swimmer are MuJoCo locomotion tasks where each agent controls a segment of an Ant or Swimmer. Predator-Prey (PredPrey N) environment is a cooperative environment where N of agents work together to chase and capture prey agents. In Heterogeneous Predator Prey, each Predator has differing capabilities of speed and acceleration. This modification is challenging as a policy must not only coordinate between the Predators, but roles based specialization must be considered given the heterogeneous nature of each predator's capabilities. We also validated related work on the drone delivery task under which a drone swarm of N agents (Drone Delivery-N) must complete deliveries of varying distances while avoiding collisions and conserving fuel. The code of which is available in supplementary materials and will be open sourced.

We used batching (Picheny et al., 2022) in our comparisons with MARL to allow for a large number of iterations of BO. We used a batch size of 15 in our comparison experiments. In this setting, all MuJoCo environments use the default epoch (total number of interactions with the environment for computing reward) length of 1000, for Predator-Prey environments, epoch length was 25, for Drone Delivery environment, epoch length was 150.

A.3 RL AND MARL UNDER MALFORMED REWARD

For single agent RL we compared against SAC (Haarnoja et al., 2018), PPO (Schulman et al., 2017), TD3 (Fujimoto et al., 2018), and DDPG (Lillicrap et al., 2015) as well as an algorithm using intrinsic motivation (Zheng et al., 2018). In single agent setting, we trained related work for $200,000$ timesteps. In the MARL setting, we trained for $2,000,000$ timesteps. In both single-agent setting and multi-agent setting all policy networks for both HA-GP-UCB and related work was 3 layers of 10 neurons each. The tested environments were standard OpenAI Gym benchmarks of Ant, Hopper, Swimmer, and Walker2D.

In the MARL setting we compared against COVDN (Peng et al., 2021), COMIX, FACMAC, and MADDPG. Comparisons were not possible against other approaches as these do not support continuous action environments and are restricted to discrete action spaces.

For all environments and algorithms, we used the recommended hyperparameter settings as defined by the authors.

A.4 COMPARISON WITH HDBO ALGORITHMS

For this comparison, we compared with several related works in HDBO. We compared with TurBO (Eriksson et al., 2019b), Alebo (Letham et al., 2020), TreeBO (Han et al., 2021), LineBO (Kirschner et al., 2019), and a recent variant of BO for policy search, GIBO (Müller et al., 2021).

For computational efficiency, the epoch length for MuJoCo environments was reduced to 500.

A.5 DRONE DELIVERY TASK

The experimental details follow that of comparisons with MARL.

A.6 COMPUTE

All experiments were performed on commodity CPU and GPUs. Each experimental setting took no more than 2 days to complete on a single GPU.

Table 2: Policy model sizes. Unfilled entries mean this environment was not considered during validation.

	Ant-v3	Hopper-v3	Swimmer-v3	Walker2d-v3	Ant-v3 (MARL)	Hopper-v3 (MARL)	Swimmer-v3 (MARL)	Walker2d-v3 (MARL)
RL (Single Agent)	478	263	222	356				
MARL (CTDE)					310	267	267	353
HA-GP-UCB (Single Agent)	478	263	222	356				
HA-GP-UCB (CTDE)					310	267	267	353

Table 3: Policy model sizes. Unfilled entries mean this environment was not considered during validation.

	Multiagent-Swimmer 4	Multiagent-Swimmer 8	Multiagent-Swimmer 12	Multiagent-Ant 8	Multiagent-Ant 12	Multiagent-Ant 16	PredPrey 6	PredPrey 9	PredPrey 15	Het. PredPrey 6	Het. PredPrey 9	Het. PredPrey 15
MARL (CTDE)	267	267	267	396	396	396	478	478	478	478	478	478
HA-GP-UCB (CTDE)	267	267	267	396	396	396	478	478	478	478	478	478
HA-GP-UCB (MM)	216	244	244	406	434	434	373	373	393	373	393	393

A.7 POLICY SIZES

We list the policy sizes of our models in Table 2 and 3.

Of note is in each environment, the compared against policy of RL or MARL is greater than or equal to in size vs. the policy optimized by HA-GP-UCB.

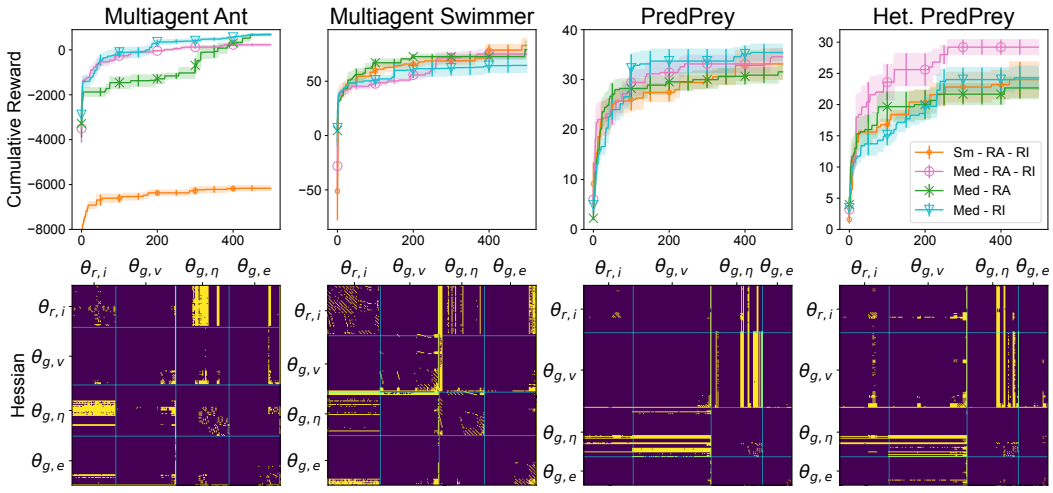

Figure 6: Ablation study. Training curves of HA-GP-UCB and its ablated variants on different multi-agent environments.

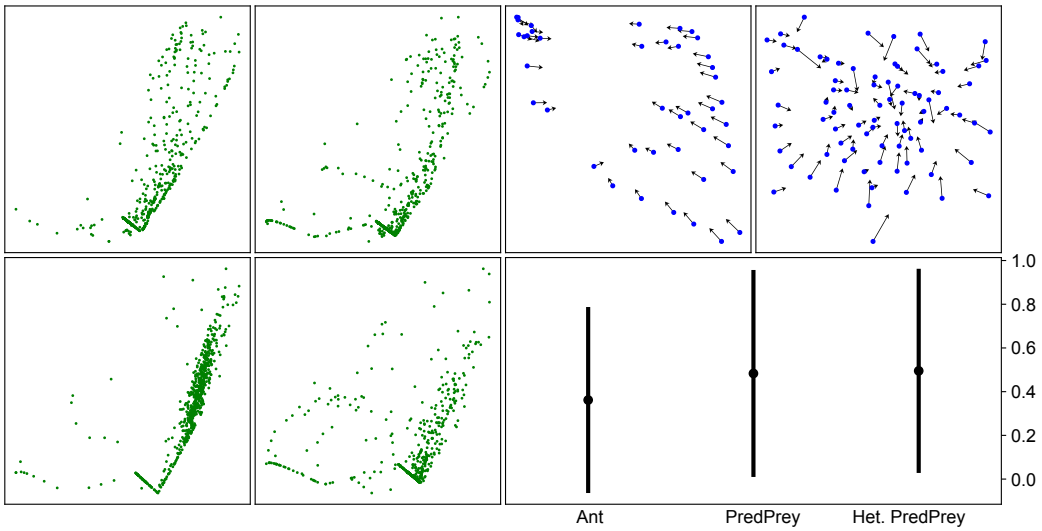

Figure 7: Left: Action distributions of different roles showing diversity in the Multiagent Ant environment with 6 agents. Right above: Policy modulation with role interaction in PredPrey and Het. PredPrey environment with 3 agents. Arrows represent change after message passing. Right below: Mean connectivity and standard deviation in role interaction in Multiagent Ant with 6 agents, PredPrey with 3 agents, and Het. PredPrey with 3 agents.

B Additional Experiments

B.1 Ablation

We present an expanded version of Fig. 3 in Fig. 6 including the ablation for Multiagent Swimmer. Multiagent Swimmer shows similar behavior as the simpler task Multiagent Ant, with stronger block-diagonal Hessian structure.

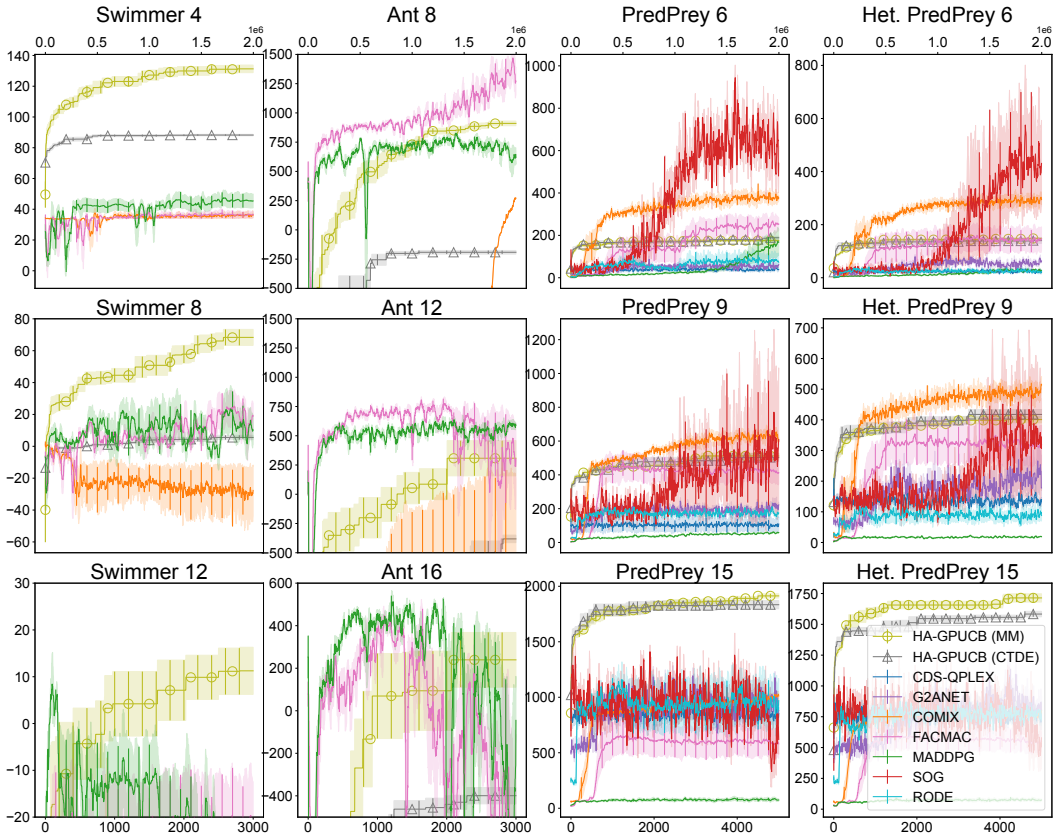

Figure 8: Comparison with MARL approaches with varying number of agents.

B.2 METAMODEL INVESTIGATION

We examined policy for Multiagent Ant with 6 agents for the role based policy specialization. The policy modulation plots were generated by examining the PredPrey and Het. PredPrey environments respectively.

In Fig. 7 we investigate the learned metamodel policies. Our investigation shows that *role* is used to specialize agent policies while maintaining a common theme. *Role interaction* modulates the policy through graphical model inferences. Finally, role interactions are sparse, however noticeably higher for complex coordination tasks such as PredPrey.

B.3 COMPARISON WITH MARL

We present an expanded version of Fig. 5 in Fig. 8 including the results for Multiagent-Ant and Multiagent-Swimmer. We observe that in this relatively uncomplicated task not well-suited for our approach with dense reward, our metamodel approach shows comparable performance to MARL approaches and far outperforms HA-GP-UCB (CTDE). This shows the overall value of our metamodel approach.

B.4 RL AND MARL UNDER MALFORMED REWARD

We present additional experiments under malformed reward for both RL and MARL. We formally define the Sparse reward scenario. Let $v(\theta) \triangleq \sum_{\Gamma=1}^{\hat{\Gamma}} r_\Gamma$ where the value of the policy is determined through $\hat{\Gamma}$ interactions with some unknown environment and each interaction is associated with the reward, r_Γ. Typically, RL algorithms observe the reward, r_Γ after every interaction with the environment. We consider a sparse reward scenario where reward feedback is given every S steps:

$\tilde{r_\Gamma^S} \triangleq \sum_{\Gamma-S}^{\Gamma} r_\Gamma$ if $\Gamma \equiv 0 \mod S$ and 0 o.w. In addition to the sparse reward setting described earlier, we also consider the setting of delayed reward. The delayed reward scenario is defined: $\tilde{r_\Gamma^D} \triangleq r_{\Gamma-D}$ if $\Gamma > D$ and 0 o.w. Thus in the delayed reward scenario, feedback on an action taken is *delayed*. This scenario is important as it arises in long term planning tasks where the value of an action is not immediately clear, but rather is ascertained after significant delays. We present the complete table comparing related works in RL with HA-GP-UCB in Table 4. As can be seen, similar to the Sparse reward scenarios, significant degradation can be observed across all tested RL algorithms with HA-GP-UCB outperforming RL algorithms with moderate to severe amount of sparsity or delay. This degradation cannot be overcome by increasing the size of the policy, as we verify with the "XL" models which are orders of magnitude larger with 3 layers of 400 neurons.

We repeat these experimental scenarios in the MARL setting with similar results in Table 5 where MARL approaches are compared against HA-GP-UCB in the CTDE setting. Thus our validation shows that in both RL and MARL strong performance requires dense, informative feedback which may not be present outside of simulator settings. In these settings, our approach of optimizing small compact policies using HA-GP-UCB outperforms related work in both RL and MARL.

Table 4: RL under sparse reward. Sparse n refers to sparse reward. Delay n refers to delayed reward. Averaged over 5 runs. Parenthesis indicate standard error. Standard error for MARL algorithms excluded due to space constraints.

	Ant-v3					Hopper-v3					Swimmer-v3					Walker2d-v3				
	DDPG	PPO	SAC	TD3	Intrinsic	DDPG	PPO	SAC	TD3	Intrinsic	DDPG	PPO	SAC	TD3	Intrinsic	DDPG	PPO	SAC	TD3	Intrinsic
Baseline	−90.77(318.36)	1105.69(72.49)	2045.24(477.20)	2606.17(96.63)	2144.00(56.19)	604.20(153.01)	1760.65(84.29)	2775.66(390.61)	1895.76(495.13)	1734.00(152.46)	44.45(7.03)	121.38(5.25)	58.73(1.86)	48.78(0.48)	1950.00(136.79)	2203.80(177.40)	892.81(208.68)	4297.03(163.71)	1664.46(490.97)	2210.00(221.94)
Sparse 2	−32.88(548.48)	1007.80(93.20)	2563.97(440.27)	1407.40(239.70)	1964.00(97.47)	877.93(141.24)	1567.14(113.01)	3380.60(76.51)	1570.84(627.28)	2074.00(219.94)	35.59(5.65)	99.50(11.44)	46.75(0.52)	47.23(1.98)	1758.80(254.49)	1470.62(173.60)	1471.33(440.18)	1673.46(704.37)	2297.43(584.68)	1952.00(175.53)
Sparse 5	−2687.97(302.46)	961.31(100.83)	711.56(87.40)	762.61(38.90)	1916.00(96.98)	814.59(110.96)	1616.79(330.77)	3239.20(93.46)	2290.67(560.51)	1972.00(85.15)	26.66(5.52)	68.69(14.60)	43.84(0.99)	40.12(0.76)	1856.00(237.63)	961.30(231.98)	697.93(274.48)	1697.25(682.00)	2932.27(361.78)	1924.00(213.72)
Sparse 20	−2809.89(292.76)	624.07(41.20)	694.30(16.55)	379.12(145.78)	1838.00(189.72)	783.95(175.83)	1629.28(112.73)	2535.17(368.16)	1436.33(535.61)	1537.20(336.19)	19.12(12.07)	54.63(9.12)	37.78(2.06)	37.03(2.63)	2108.00(130.24)	663.04(114.66)	365.39(75.80)	1010.63(490.13)	276.56(69.79)	1810.00(253.57)
Sparse 50	−3067.37(116.57)	−67.43(33.33)	663.28(21.26)	253.66(149.78)	1091.20(102.39)	816.25(145.64)	1010.73(165.81)	1238.03(499.72)	551.43(107.79)	642.00(155.66)	23.73(4.84)	51.52(7.55)	38.78(3.69)	30.01(7.84)	812.00(124.17)	572.12(115.80)	428.29(67.93)	349.47(29.70)	298.28(52.01)	834.75(89.95)
Sparse 100	−3323.43(43.74)	−4021.56(1893.85)	679.30(23.68)	−115.43(387.63)	450.40(51.04)	988.36(15.38)	324.51(53.61)	260.52(31.38)	342.48(36.19)	406.80(51.77)	9.64(9.63)	21.09(4.97)	27.98(6.35)	30.10(1.95)	376.60(80.28)	523.89(114.07)	205.93(18.11)	200.16(26.90)	147.22(24.45)	480.60(75.14)
Sparse 200	−3098.37(73.21)	−8167.98(2801.42)	−107.14(614.25)	−147.86(339.75)	258.60(46.74)	765.05(146.14)	222.76(30.59)	300.36(23.10)	281.68(24.31)	350.80(58.42)	−9.97(6.70)	21.69(1.97)	33.35(3.40)	30.48(2.96)	342.80(10.65)	182.84(56.03)	193.43(21.18)	187.16(33.73)	148.06(23.58)	353.20(39.59)
Sparse XL 100	−2887.23(182.60)	−5101.30(1979.27)	448.67(118.92)	−730.72(720.12)	429.28(141.31)	608.79(116.39)	481.72(68.89)	338.95(10.91)	188.06(67.16)	570.60(59.04)	9.56(8.54)	31.47(1.56)	36.70(2.86)	50.99(11.73)	473.60(67.95)	867.98(65.29)	222.13(23.32)	250.27(12.96)	95.01(43.31)	421.80(31.56)
Sparse XL 200	−2681.07(338.79)	−11084.05(2144.89)	378.36(134.59)	−1331.39(593.91)	373.40(28.56)	1025.16(27.15)	361.34(80.88)	331.69(17.64)	217.04(78.00)	323.00(27.99)	9.87(3.57)	26.53(2.31)	36.61(12.84)	35.21(1.31)	414.40(20.72)	843.38(70.49)	206.66(31.57)	228.23(32.14)	270.81(40.87)	339.20(39.73)
Lag 1	−645.56(227.58)	1028.21(103.66)	1981.58(547.20)	2570.06(222.18)	2378.00(111.98)	872.14(200.36)	1907.03(177.63)	2957.36(394.10)	2817.97(434.89)	1870.00(147.92)	22.56(8.02)	111.21(11.94)	70.53(17.88)	47.56(0.70)	1980.00(106.70)	2197.41(266.42)	1180.54(362.97)	3696.70(712.13)	3806.52(125.83)	1788.00(165.73)
Lag 5	−289.09(349.92)	848.20(43.75)	866.79(5.83)	860.53(14.72)	1761.00(365.85)	803.61(165.08)	1812.64(219.72)	3280.25(113.50)	3377.79(48.14)	1844.00(124.02)	16.78(6.09)	89.70(14.44)	55.33(4.99)	47.22(4.34)	2018.00(245.36)	1328.87(246.34)	939.46(325.01)	3786.54(650.86)	3192.36(411.20)	1856.00(136.84)
Lag 20	−2578.96(232.02)	239.10(69.51)	824.88(16.22)	54.63(619.95)	1718.00(131.74)	692.46(145.45)	1944.12(263.78)	3385.03(99.86)	2172.20(584.96)	1526.00(141.38)	21.16(9.25)	57.67(4.33)	30.10(6.27)	42.81(3.63)	1910.00(165.43)	424.51(86.68)	740.85(108.44)	2117.99(652.73)	2098.28(608.93)	1726.00(66.79)
Lag 50	−3093.43(184.86)	−176.74(92.67)	796.54(16.98)	791.36(13.30)	1155.60(240.31)	907.45(108.70)	815.34(169.38)	596.78(111.18)	632.85(47.74)	1086.80(109.28)	22.80(7.69)	35.28(1.43)	29.66(2.61)	26.24(2.97)	1092.20(154.49)	816.82(282.02)	368.41(82.30)	388.86(89.26)	463.18(69.68)	942.80(34.55)
Lag 100	−2784.59(92.56)	−5569.67(2110.05)	691.43(22.64)	640.74(150.22)	257.60(39.28)	1028.14(4.61)	216.30(36.22)	264.75(39.10)	280.71(28.86)	292.60(37.57)	11.72(7.89)	23.17(4.52)	37.77(2.77)	29.70(3.84)	377.20(54.35)	623.61(158.22)	197.16(17.30)	209.38(33.69)	164.23(46.02)	382.40(47.09)
Lag 200	−2655.00(279.69)	−9419.68(2630.52)	605.73(38.85)	163.37(233.15)	392.00(35.41)	786.96(167.20)	271.60(46.30)	278.98(25.34)	261.85(28.64)	299.60(28.97)	4.79(7.07)	19.18(2.56)	30.63(5.42)	22.95(15.78)	340.40(42.30)	488.44(128.77)	182.44(4.52)	147.78(20.99)	128.41(9.08)	304.40(29.79)
Lag XL 100	−2722.44(340.81)	−3140.81(1713.55)	503.88(61.88)	−404.24(557.86)	406.80(41.65)	280.53(151.98)	577.16(15.41)	306.99(9.95)	196.16(71.21)	431.40(34.74)	23.54(3.69)	25.59(3.22)	26.03(5.69)	24.72(9.90)	442.60(26.17)	656.25(107.48)	227.56(33.60)	275.18(34.80)	153.97(46.48)	478.00(33.40)
Lag XL 200	−2507.38(142.96)	−8436.44(461.63)	300.83(115.24)	−959.92(648.64)	377.60(26.88)	439.52(175.28)	390.18(80.50)	331.84(15.65)	191.93(69.82)	342.60(33.82)	9.84(5.67)	23.67(3.79)	17.53(1.37)	9.07(7.23)	317.60(54.21)	844.83(40.32)	219.77(23.85)	237.50(19.80)	168.55(49.49)	336.40(44.66)
HA-GP-UCB	1147.21(36.88)					1009.3(17.95)					175.73(15.54)					1008.90(11.95)				

Table 5: MARL under sparse reward. Sparse n refers to sparse reward. Delay n refers to delayed reward. Averaged over 5 runs. Parenthesis indicate standard error. Standard error for MARL algorithms excluded due to space constraints.

	Ant-v3				Hopper-v3				Swimmer-v3				Walker2d-v3			
	COVDN	COMIX	MADDPG	FACMAC	COVDN	COMIX	MADDPG	FACMAC	COVDN	COMIX	MADDPG	FACMAC	COVDN	COMIX	MADDPG	FACMAC
Baseline	970.50(5.96)	959.20(2.83)	982.58(54.19)	909.37(40.08)	38.92(0.06)	38.88(0.06)	305.58(107.48)	638.80(245.03)	−0.05(7.75)	13.43(0.56)	11.21(0.52)	14.06(0.11)	249.21(9.65)	275.85(10.04)	280.30(22.16)	543.71(180.78)
Sparse 5	877.09(20.95)	906.38(7.48)	912.61(12.48)	882.94(46.35)	138.07(72.33)	39.84(0.21)	320.74(123.44)	38.76(0.13)	10.13(2.16)	7.10(4.77)	14.19(0.81)	13.05(0.36)	341.89(51.92)	260.58(29.98)	236.37(28.88)	267.70(118.68)
Sparse 50	242.69(252.04)	−212.74(75.21)	772.29(41.13)	799.71(11.76)	909.88(71.73)	38.88(0.16)	362.66(264.39)	172.37(109.04)	8.49(2.52)	10.50(2.19)	12.06(0.34)	12.10(1.01)	190.92(8.87)	121.51(67.31)	227.60(31.76)	287.36(67.07)
Sparse 100	403.99(208.18)	−756.61(201.61)	490.19(62.22)	564.73(75.80)	443.95(170.88)	27.74(9.52)	47.47(7.08)	38.78(0.06)	−3.44(3.28)	1.12(3.02)	10.11(3.29)	14.14(1.19)	106.24(66.51)	194.91(17.35)	159.24(14.53)	444.41(215.77)
Sparse 200	65.04(89.82)	−197.26(64.03)	−632.05(967.04)	584.23(28.34)	687.82(249.11)	38.78(0.02)	50.74(9.86)	37.10(1.36)	−0.15(3.39)	−1.65(1.20)	14.43(0.73)	13.84(0.13)	165.50(67.12)	277.89(4.89)	123.91(10.61)	229.11(129.53)
Sparse XL 100	766.66(208.18)	209.58(201.61)	632.15(62.22)	552.28(75.80)	813.86(170.88)	236.74(9.52)	50.62(7.08)	72.11(0.06)	9.52(3.28)	14.45(3.02)	13.69(3.29)	14.16(1.19)	135.95(66.51)	28.22(17.35)	127.66(14.53)	184.52(215.77)
Sparse XL 200	442.92(89.82)	322.54(64.03)	479.03(967.04)	553.20(28.34)	996.85(249.11)	39.47(0.02)	26.41(9.86)	71.31(1.36)	3.58(3.39)	4.67(1.20)	13.28(0.73)	11.76(0.13)	167.62(67.12)	207.81(4.89)	157.37(10.61)	88.47(129.53)
Lag 1	656.69(5.96)	426.57(2.83)	901.14(54.19)	835.77(40.08)	55.16(0.06)	38.90(0.06)	113.16(107.48)	38.94(245.03)	1.99(7.75)	9.73(0.56)	13.60(0.52)	14.34(0.11)	413.26(9.65)	318.73(10.04)	312.03(22.16)	357.73(180.78)
Lag 5	255.56(20.95)	302.46(7.48)	873.81(12.48)	910.90(46.35)	629.94(72.33)	38.85(0.21)	376.38(123.44)	852.61(0.13)	10.07(2.16)	4.16(4.77)	14.51(0.81)	13.20(0.36)	170.91(51.92)	289.44(29.98)	302.41(28.88)	640.21(118.68)
Lag 50	56.94(252.04)	−272.78(75.21)	870.20(41.13)	885.37(11.76)	383.76(71.73)	15.95(0.16)	612.84(264.39)	358.68(109.04)	0.79(2.52)	6.26(2.19)	14.55(0.34)	11.16(1.01)	97.29(8.87)	145.98(67.31)	183.41(31.76)	289.66(67.07)
Lag 100	112.82(208.18)	10.21(201.61)	844.80(62.22)	843.90(75.80)	997.51(170.88)	358.31(9.52)	275.65(7.08)	466.20(0.06)	10.41(3.28)	3.49(3.02)	13.67(3.29)	11.60(1.19)	114.41(66.51)	83.99(17.35)	149.93(14.53)	102.84(215.77)
Lag 200	366.62(89.82)	−125.65(64.03)	674.74(967.04)	723.30(28.34)	434.69(249.11)	38.95(0.02)	26.44(9.86)	38.74(1.36)	5.27(3.39)	7.93(1.20)	13.28(0.73)	13.86(0.13)	51.59(67.12)	147.23(4.89)	163.30(10.61)	277.79(129.53)
Lag XL 100	869.35(208.18)	292.96(201.61)	721.18(62.22)	814.25(75.80)	1000.04(170.88)	38.85(9.52)	532.12(7.08)	26.30(0.06)	9.23(3.28)	0.48(3.02)	10.12(3.29)	10.19(1.19)	105.65(66.51)	119.80(17.35)	191.74(14.53)	280.79(215.77)
Lag XL 200	657.07(89.82)	260.13(64.03)	611.73(967.04)	801.94(28.34)	692.92(249.11)	38.85(0.02)	38.78(9.86)	26.27(1.36)	5.62(3.39)	6.48(1.20)	13.57(0.73)	5.56(0.13)	56.72(67.12)	227.96(4.89)	122.66(10.61)	222.77(129.53)
HA-GP-UCB (CTDE)	988.44(3.39)				213.50(22.26)				31.95(1.3)				397.68(14.42)			

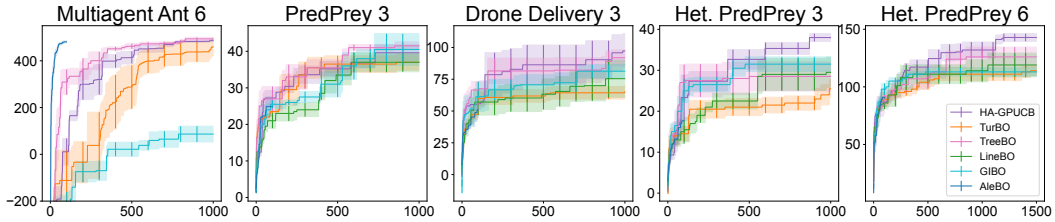

Figure 9: Comparison with BO algorithms. HA-GP-UCB outperforms on complex multi-agent coordination tasks.

B.5 Comparison with HDBO Algorithms

We compare with several related work in High-dimensional BO including TurBO (Eriksson et al., 2019b), AleBO (Letham et al., 2020), LineBO (Kirschner et al., 2019), TreeBO (Han et al., 2021), and GIBO (Müller et al., 2021). This is presented in Fig. 9. We experienced out-of-memory issues with AleBO after approximately 100 iterations, hence the AleBO plots are truncated. We compare against these algorithms at optimizing our metamodel policy for solving various multi-agent policy search tasks. We validated on Multiagent Ant with 6 agents, PredPrey with 3 agents, Het. PredPrey with 3 agents, Drone Delivery with 3 agents, and also Het. PredPrey with 6 agents. We observe that these competing works offer competitive performance for simpler tasks such as Multiagent Ant and PredPrey with 3 agents. However for more complex tasks that require role based interaction and coordination, our approach outperforms related work. This is evidenced in Het. PredPrey 3, Het. PredPrey 6 as well as the Drone Delivery task with 3 agents.

Thus our validation shows that for simpler task, competing related works are able to optimize for simple policies of low underlying dimensionality. However, for more complex tasks which require sophisticated interaction using both Role and Role Interaction, related work is less capable of optimizing for strong policies *due to the complexity of the high-dimensional* BO *task*. In contrast, our work offers the capability of finding stronger policies for these complex tasks and scenarios.

C On the Applicability of Our Assumptions to RBF and Matern Kernel

We show that our assumption is satisfied by the RBF Kernel when $\Theta = [0,1]^D$, and is quasi-satisfied by the Matern$-\frac{5}{2}$ kernel. We also show that in the setting where $\Theta = [0,r]^D$ for some bounded r, our assumptions are quasi-satisfied as although these kernels may take on small negative values, these values decay exponentially with respect to the distance. These Lemmas show that our assumptions are reasonable.

Lemma 1. *Let $k(\theta,\theta') \triangleq \exp(\frac{-d^2}{2})$ be the RBF kernel with $d \triangleq ||\theta - \theta'||$, then*

$$k^{\partial i \partial j}(\theta,\theta') = k(\theta,\theta')\left(1-(\theta^i-\theta'^i)^2\right)\left(1-(\theta^j-\theta'^j)^2\right).$$

Proof. As shown in (Rasmussen & Williams, 2006) Section 9.4, the derivative of a Gaussian Process is also a Gaussian Process. Let $GP(0, k(\theta,\theta'))$ be the GP from which f is sampled. This implies:

$$\frac{\partial f}{\partial \theta^a} \sim GP\left(0, \frac{\partial^2 k(\theta,\theta')}{\partial \theta^a \partial \theta'^a}\right).$$

Applying this rule once more for the Hessian, we have:

$$\frac{\partial^2 f}{\partial \theta^b \theta^a} \sim GP\left(0, \frac{\partial^4 k(\theta,\theta')}{\partial \theta^b \partial \theta'^b \partial \theta^a \partial \theta'^a}\right).$$

Given the above identities, we compute the partial derivatives for the RBF kernel:

$$\frac{\partial^2 k(\theta,\theta')}{\partial \theta^a \partial \theta'^a} = \exp\left(-\frac{||\theta-\theta'||^2}{2}\right)\left(1-(\theta^a-\theta'^a)^2\right).$$

Deriving once more we have:

$$\frac{\partial^4 k(\theta,\theta')}{\partial \theta^b \partial \theta'^b \partial \theta^a \partial \theta'^a} = \exp\left(-\frac{||\theta-\theta'||^2}{2}\right)\left(1-(\theta^a-\theta'^a)^2\right)\left(1-(\theta^b-\theta'^b)^2\right).$$

This completes the proof noting that $k(\theta,\theta') \triangleq \exp(\frac{-d^2}{2})$ with $d \triangleq ||\theta-\theta'||$. □

Corollary 1. *Let $k(\theta,\theta') \triangleq \exp(\frac{-d^2}{2})$, and $\theta,\theta' \in [0,1]^D$, then $k^{\partial i \partial j}(\theta,\theta') \geq 0$.*

Proof. The above is straightforward to see as $\exp(\cdot) \geq 0$ and with $\theta,\theta' \in [0,1]^D$ we have $\left(1-(\theta^a-\theta'^a)^2\right) \geq 0\left(1-(\theta^b-\theta'^b)^2\right) \geq 0$. □

Corollary 2. *Let $k(\theta,\theta') \triangleq \exp(\frac{-d^2}{2})$, and $\theta,\theta' \in [0,r]^d$, then $k^{\partial i \partial j}(\theta,\theta') \geq c\exp(-d^2)$ for some constant c dependent on r.*

Proof. The above is straightforward given the above Lemma. We note that although the RBF kernel may take on negative values in the domain $\Theta = [0,r]^d$, this values experience strong tail decay showing the quasi-satisfaction of our assumptions. □

The above Lemma and Corollary shows that our assumptions are satisfied by the RBF Kernel when $\Theta = [0,1]^D$, and quasi satisfied when $\Theta = [0,r]^D$ after choosing a suitable p_h and σ_h^2. We show how these assumptions are quasi-satisfied by the Matern-$\frac{5}{2}$ kernel.

Lemma 2. *Let* $k(\theta, \theta') \triangleq (1 + \sqrt{5}d + \frac{5}{3}d^2)\exp(-\sqrt{5}d)$ *be the Matern-*$\frac{5}{2}$ *kernel with* $d \triangleq ||\theta - \theta'||$*, then with* $d_i \triangleq \theta^i - \theta'^i$ *we have*

$$k^{\partial i \partial j}(\theta, \theta') = \exp(-\sqrt{5}d)\left(\frac{5\sqrt{5}}{3} - \frac{25}{3d}d_i^2 - \frac{25}{3d}d_j^2 + \frac{25\sqrt{5}}{3d^2}d_i^2 d_j^2 + \frac{25}{3d^3}d_i^3 d_j^3\right).$$

Proof. Following the proof of Lemma 1, we state the partial derivatives of the Matern-$\frac{5}{2}$ kernel:

$$\frac{\partial^2 k(\theta, \theta')}{\partial\theta^a \partial\theta'^a} = \exp\left(-\sqrt{5}||\theta - \theta'||\right)\left(\frac{5}{3} + \frac{5\sqrt{5}}{3}||\theta - \theta'|| - \frac{25}{3}(\theta^a - \theta'^a)^2\right).$$

Differentiating one more we have

$$\begin{aligned}\frac{\partial^4 k(\theta, \theta')}{\partial\theta^b \partial\theta'^b \partial\theta^a \partial\theta'^a} &= \exp\left(-\sqrt{5}||\theta - \theta'||\right) \\ &\left(\frac{5\sqrt{5}}{3} - \frac{25}{3d}(\theta^a - \theta'^a)^2 - \frac{25}{3d}(\theta^b - \theta'^b)^2 + \frac{25\sqrt{5}}{3d^2}(\theta^a - \theta'^a)^2(\theta^b - \theta'^b)^2\right. \\ &\left. + \frac{25}{3d^3}(\theta^a - \theta'^a)^3(\theta^b - \theta'^b)^3\right).\end{aligned}$$

This completes the proof noting that $d_i \triangleq \theta^i - \theta'^i$ and $d \triangleq ||\theta - \theta'||$. □

Corollary 3. *Let* $k(\theta, \theta') \triangleq (1 + \sqrt{5}d + \frac{5}{3}d^2)\exp(-\sqrt{5}d)$ *and* $\theta, \theta' \in [0, 1]^D$*. Then* $k^{\partial i \partial j}(\theta, \theta') \geq \exp(-\sqrt{5}d)\left(\frac{5\sqrt{5}}{3} - \frac{25}{3d} - \frac{25}{3d} - \frac{25}{3d^3}\right)$.

Proof. The above is an immediate consequence of Lemma 2 and noting that $||d_i|| \leq 1$. □

Corollary 4. *Let* $k(\theta, \theta') \triangleq (1 + \sqrt{5}d + \frac{5}{3}d^2)\exp(-\sqrt{5}d)$ *and* $\theta, \theta' \in [0, r]^d$*. Then* $k^{\partial i \partial j}(\theta, \theta') \geq c\exp(-d)$ *for some* c *dependent on* r.

Proof. The above is an immediate consequence of Lemma 2 and noting that $||d_i|| \leq r$. □

Although the above corollary shows that the Matern-$\frac{5}{2}$ kernel may take on negative values, we note that these values experience strong tail decay due to the presence of the $\exp\left(-\sqrt{5}d\right)$ term. Thus, the negative values are likely to be extremely small, thus quasi-satisfying our assumptions. In our experiments, we observed no shortcoming in using the Matern-$\frac{5}{2}$ kernel in HA-GP-UCB.

D PROOF OF PROPOSITION 1

We restate Proposition 1 for clarity.

Proposition 1. *Let $\mathcal{G}_d = (V_d, E_d)$ represent an additive dependency structure with respect to $v(\theta)$, then the following holds true: $\forall a, b \; \frac{\partial^2 v}{\partial \theta^a \partial \theta^b} \neq 0 \implies (\Theta^a, \Theta^b) \in E_d$ which is a consequence of v formed through addition of independent sub-functions $v^{(i)}$, at least one of which must contain θ^a, θ^b as parameters for $\frac{\partial^2 v}{\partial \theta^a \partial \theta^b} \neq 0$ which implies their connectivity within E_d.*

Proof. The above follows from the linearity of addition, which naturally implies a lack of curvature. In the multivariate case, this corresponds to zero or non-zero entries in the Hessian.

To be precise, we prove the contrapositive:

$$(\Theta^a, \Theta^b) \notin E_d \implies \frac{\partial^2 v}{\partial \theta^a \partial \theta^b} = 0.$$

Let a, b be arbitrary dimensions with $(\Theta^a, \Theta^b) \notin E_d$. As a consequence of the definition of the dependency graph, $\nexists \Theta^{(i)}$ s.t. $\{\Theta^a, \Theta^b\} \subseteq \Theta^{(i)}$. That is, no subfunction $v^{(i)}$ takes both θ^a and θ^b as arguments.

By the linearity of the partial derivative, we see that:

$$\frac{\partial^2}{\partial \theta^a \partial \theta^b} v(\theta) = \frac{\partial^2}{\partial \theta^a \partial \theta^b} \sum_{i=1}^{M} v^{(i)}(\theta^{(i)}) = \sum_{i=1}^{M} \frac{\partial^2}{\partial \theta^a \partial \theta^b} v^{(i)}(\theta^{(i)}) = 0$$

where the last equality follows from no subfunction $v^{(i)}$ taking both θ^a and θ^b as arguments. □

E PROOF OF THEOREM 1

Our proof of Theorem 1 relies in being able to determine whether an edge does or does not exist in the dependency graph. To be able to do this, we examine the Hessian. As we have shown in Proposition 1, examining the Hessian answers this question. The challenge of Theorem 1 is detecting this dependency under noisy observations of the Hessian, as well as in domains where the variance of the second partial derivative is often zero, i.e., $k^{\partial i \partial j}(\theta, \theta') = 0$ with high probability. To overcome this challenge, we sample the Hessian multiple times to both find portions of the domain where $k^{\partial i \partial j}(\theta, \theta') \geq \sigma_h^2$, and also reduce the effect of the noise on learning the dependency structure. To proceed with the analysis, we first prove a helper lemma showing that if we can construct two Normal variables of sufficiently different variances, then it's possible to accurately determine which Normal variable has low, and high variance by taking a singular sample from each. This helper lemma will be used later to help determine edges in the dependency graph. As we shall soon show, If an edge exists, we are able to construct a Normal variable with high variance. Correspondingly, if an edge does not exist, we are able to construct a Normal variable with low variance.

Lemma 3. *Let $X_l \sim \mathcal{N}(0, \sigma_l^2)$ and $X_h \sim \mathcal{N}(0, \sigma_h^2)$ be two random univariate gaussian variables. For any $\delta \in (0, 1)$, $\exists\, c_h$ s.t. $|X_l| \leq c_h \leq |X_h|$ with probability $1 - \delta$ when $\frac{\sigma_h^2}{\sigma_l^2} > \frac{8}{\delta^2} \log \frac{2}{\delta}$ and precisely when $\frac{\sigma_h \delta}{2} > c_h > \sigma_l \sqrt{2 \log \frac{2}{\delta}}$.*

Proof. First we note that $|X_l|$ and $|X_h|$ are Half-Normal random variables, with cumulative distribution function of $F_l(x) = \operatorname{erf} \frac{x}{\sigma_l \sqrt{2}}$ and $F_h(x) = \operatorname{erf} \frac{x}{\sigma_h \sqrt{2}}$ respectively. Thus to show that $|X_l| \leq \sigma_l \sqrt{2 \log \frac{2}{\delta}}$ and $|X_h| \geq \frac{\sigma_h \delta}{2}$ with high probability, we utilize well known bounds on the erf and erfc function. The proofs of the below can be found in several places, e.g., Chu (1955) and Ermolova & Häggman (2004) respectively.

$$\operatorname{erf} x \leq \sqrt{1 - \exp -2x^2}\,; \;\; \operatorname{erfc} x \leq \exp -x^2.$$

Given the above, we show that $p(c_h \leq |X_l|) \leq \frac{\delta}{2}$ and $p(c_h \geq |X_h|) \leq \frac{\delta}{2}$ and utilizing the union bound completes the proof.

$$c_h > \sigma_l \sqrt{2 \log \frac{2}{\delta}} \implies c_h^2 > 2\sigma_l^2 \log \frac{2}{\delta} \implies \frac{c_h^2}{2\sigma_l^2} > -\log \frac{\delta}{2} \implies -\frac{c_h^2}{2\sigma_l^2} < \log \frac{\delta}{2}$$

$$\implies \exp -\frac{c_h^2}{2\sigma_l^2} \leq \frac{\delta}{2} \implies \operatorname{erfc} \frac{c_h}{\sqrt{2}\sigma_l} < \frac{\delta}{2} \implies 1 - \operatorname{erf} \frac{c_h}{\sqrt{2}\sigma_l} \geq 1 - \frac{\delta}{2} \implies F_l(c_h) \geq 1 - \frac{\delta}{2}$$

$$\implies p\left(c_h \leq |X_l|\right) < \frac{\delta}{2}.$$

Following a similar line of reasoning we have:

$$c_h < \frac{\sigma_h \delta}{2} \implies \frac{c_h^2}{\sigma_h^2} < \frac{\delta^2}{4} \implies \frac{-c_h^2}{\sigma_h^2} > -\frac{\delta^2}{4} \implies \frac{-c_h^2}{\sigma_h^2} > \log 1 - \frac{\epsilon^2}{4} \implies \exp -\frac{c_h^2}{\sigma_h^2} > 1 - \frac{\delta^2}{4}$$

$$\implies 1 - \exp -\frac{c_h^2}{\sigma_h^2} < \frac{\delta^2}{4} \implies \sqrt{1 - \exp -\frac{c_h^2}{\sigma_h^2}} < \frac{\delta}{2} \implies \operatorname{erf} \frac{c_h}{\sigma_h \sqrt{2}} < \frac{\delta}{2} \implies F_h(c_h) < \frac{\delta}{2}$$

$$\implies p(c_h \geq |X_h|) < \frac{\delta}{2}.$$

Finally, to complete the proof, we show that the interval $(\sigma_l \sqrt{2 \log \frac{2}{\delta}}, \frac{\sigma_h \delta}{2})$ is not the empty set when $\frac{\sigma_h^2}{\sigma_l^2} > \frac{8}{\delta^2} \log \frac{2}{\delta}$.

$$\frac{\sigma_h^2}{\sigma_l^2} > \frac{8}{\delta^2}\log\frac{2}{\delta} \implies \frac{\sigma_h}{\sigma_l} > \frac{2\sqrt{2}}{\delta}\sqrt{\log\frac{2}{\delta}} \implies \frac{\sigma_h\delta}{2} > \sigma_l\sqrt{2\log\frac{2}{\delta}}.$$

□

We are now ready to prove Theorem 1.

Theorem 1. *Suppose*[10] *there exists* σ_h^2, p_h *s.t.* $\forall i,j\ \mathbb{P}_{\theta\sim\mathcal{U}(\Theta)}\left[k^{\partial i\partial j}(\theta,\theta)\geq\sigma_h^2\right]\geq p_h$ *and* $\forall i,j,\theta,\theta'$ $k^{\partial i\partial j}(\theta,\theta')\geq 0$. *Then for any* $\delta_1,\delta_2\in(0,1)$ *after* $t\geq T_0$ *steps of* HA-GP-UCB *we have:* $\bigcap_{i,j}P(\widetilde{E}_d^{i,j}=E_d^{i,j})\geq 1-\delta_1-\delta_2$ *when* $T_0=C_1>\frac{8D^2}{\delta_1^2}\log\frac{2D^2}{\delta_1}\frac{\sigma_n^2}{\sigma_h^2}+\frac{D^2}{p_h\delta_2}$, $c_h\triangleq T_0\sigma_n\sqrt{2\log\frac{2D^2}{\delta_1}}$.

Proof. We prove the above for a single pair of variables, i.e., $k^{\partial i\partial j}$ and utilize the union bound to complete the proof. The first challenge to overcome is to sufficiently sample enough points in the domain such that we are able to find enough points $\theta\in\Theta$ where $k^{\partial i\partial j}(\theta,\theta)\geq\sigma_h^2$. To achieve this we sample T_0 different θ in the domain. After sampling T_0 points if there exists an edge between Θ^a, and Θ^b, then with probability $1-\frac{\delta_2}{D^2}$ we have sampled $T_0-\frac{D^2}{p_h\delta_2}$ points where $k^{\partial i\partial j}(\theta,\theta)\geq\sigma_h^2$. To show the above we use bounds on the cumulative distribution of the Binomial theorem. A well known bound is given T_0 trials, with p_h probability of success, the probability of having fewer than s successes is upper bounded as follows:

$$\frac{p_h}{T_0-s}.$$

Given the above, we use δ_2 and derive:

$$\frac{p_h}{T_0-(T_0-\frac{D^2}{p_h\delta_2})}\leq\frac{\delta_2}{D^2}.$$

Given the above, with at least $(T_0-\frac{D^2}{p_h\delta_2})$ points where $k^{\partial i\partial j}(\cdot,\cdot)\geq\sigma_h^2$, as well as our assumption $k^{\partial i\partial j}(\theta,\theta)\geq 0$, we apply Bienaymé's identity which we restate for convenience:

$$\operatorname{Var}\left[\sum_{\ell=1}^{C_1}h_{t,\ell}\right]=\sum_{\ell=1}^{C_1}\sum_{\ell'=1}^{C_1}\operatorname{Cov}\left(h_{t,\ell},h_{t,\ell'}\right).$$

Noting each of the $(T_0-\frac{D^2}{p_h\delta_2})$ successes is sampled $C_1=T_0$ times with $\operatorname{Cov}\left(h_{t,\ell},h_{t,\ell'}\right)\geq\sigma_h^2$ for each of the successes and $\operatorname{Cov}\left(h_{t,\ell},h_{t,\ell'}\right)\geq 0$ for all samples by our assumption. Applying Bienaymé's identity and the sum of (correlated) Normal variables is also a normal variable, we have $\operatorname{Var}\left[\sum_{t=1}^{C_1}\sum_{\ell=1}^{C_1}h_{t,\ell}\right]\geq(T_0-\frac{D^2}{p\delta_2})T_0^2\sigma_h^2$. Compare this quantity with the variance if no edge exists between Θ^a, and Θ^b, where the variance results from i.i.d. noise: $\operatorname{Var}\left[\sum_{t=1}^{T_0}\sum_{\ell=1}^{T_0}h_{t,\ell}\right]=T_0^2\sigma_n^2$. Comparing these two quantities, with an appropriately picked c_h determines the edge between Θ^a and Θ^b using Lemma 3. By Lemma 3, letting $c_h\triangleq T_0\sigma_n\sqrt{2\log\frac{2D^2}{\delta_1}}$ ensures that $p(h^{i,j}<c_h)<\frac{\delta_1}{D^2}$ if edge $E_d^{i,j}$ exists, and $p(h^{i,j}>c_h)<\frac{\delta_1}{D^2}$ if edge $E_d^{i,j}$ does not exist. Applying the union bound over D^2 pairs of variables completes the proof with $\bigcap_{i,j}P(\widetilde{E}_d^{i,j}=E_d^{i,j})\geq 1-\delta_1-\delta_2$.

□

[10]RBF kernel satisfies these assumptions when $\Theta=[0,1]^D$.

F PROOF OF THEOREM 2

Our proof of Theorem 2 is presented under the same setting and assumptions as the work of Srinivas et al. (2010).

To prove Theorem 2, we rely on several helper lemmas. The high-level sketch of the proof is to use the properties of Erdős-Rényi graph to bound both the *size of the maximal clique* as well as *the number of maximal cliques* with high probability. Once these two quantities are bounded, we are able to analyze the mutual information of the kernel constructed by *summing the kernels corresponding to the maximal cliques* of the sampled Erdős-Rényi graph as indicated in Assumption 1. Finally, once this mutual information is bounded, we use similar analysis as Srinivas et al. (2010) to complete the regret bound.

We begin by bounding the size of the maximal cliques.

Lemma 4. *Let $\mathcal{G}_d = (V_d, E_d)$ be sampled from a Erdős-Rényi model with probability p_g: $\mathcal{G}_d \sim G(D, p_g)$, then $\forall \delta \in (0,1)$ the largest clique of $\mathcal{G}_d$ is bounded above by*

$$|\textit{Max-Clique}(\mathcal{G}_d)| \leq 2\log_{\frac{1}{p_g}}|V_d| + 2\sqrt{\log_{\frac{1}{p_g}}\frac{|V_d|}{\delta} + 1}$$

with probability at least $1-\delta$.

Proof. The above relies on well known upper bounds on the maximal clique size on a graph sampled from an Erdős-Rényi model. As shown in (Bollobás & Erdös, 1976) and (Matula, 1976) the expected number of Cliques of size k, $\mathbb{E}[C_k]$ is given by:

$$\mathbb{E}[C_k] = \binom{|V_d|}{k}\frac{1}{p_g}^{-\binom{k}{2}} \leq |V_d|^k \frac{1}{p_g}^{-\frac{k(k-1)}{2}} = \frac{1}{p_g}^{\frac{k}{2}\left(2\log_{\frac{1}{p_g}}|V_d| - k + 1\right)}.$$

In the sequel, we omit the base of the log: $\frac{1}{p_g}$ for clarity. To bound the size of the maximal clique, we find a suitable k such that $\mathbb{E}[C_k] \leq \frac{\delta}{n}$ and utilize the union bound over $[C_i]_{i=k,\ldots,n}$ where we have $|[C_i]_{i=k,\ldots,n}| \leq n$. Finally, we utilize Markov's inequality to complete the proof.

$$\text{Let } k = 2\log|V_d| + 2\sqrt{\log\frac{|V_d|}{\delta} + 1}.$$

We utilize the above bound on $\mathbb{E}[C_k]$.

$$\begin{aligned}
\implies \frac{k}{2}\left(2\log_{\frac{1}{p_g}}|V_d| - k + 1\right) &= \\
&\left(\log|V_d| + \sqrt{\log\frac{n}{\delta}}\right)\left(2\log|V_d| - 2\log|V_d| - 2\sqrt{\log\frac{n}{\delta}+1} + 1\right) \\
&\leq -\log|V_d| - \log\frac{n}{\delta} + 1 \leq \log\frac{\delta}{n} \\
\implies \mathbb{E}[C_k] &\leq \frac{1}{p_g}^{\log\frac{\delta}{n}} = \frac{\delta}{n}.
\end{aligned}$$

The proof is complete by noting that by Markov inequality, $p(C_k \geq 1) \leq \mathbb{E}[C_k]$ and taking the union bound over at most n members of $[C_i]_{i=k,\ldots,n}$. □

Next, we bound the total number of maximal cliques:

Lemma 5. *Let $\mathcal{G}_d = (V_d, E_d)$ be sampled from a Erdős-Rényi model with probability p: $\mathcal{G}_d \sim G(D, p_g)$, then $\forall \delta \in (0,1)$ the number of total maximal cliques in $\mathcal{G}_d$ is bounded above by*

$$\frac{1}{\delta}\sqrt{|V_d|^{\log_{\frac{1}{p_g}}|V_d|+5}}$$

with probability at least $1-\delta$.

Proof. We prove the above by bounding $\max_k C_k$ with high probability and noting that the number of maximal cliques is bounded by $\sum_k C_k \leq n \max_k C_k$ with high probability. To bound $\max C_k$, we first consider $\max_k \mathbb{E}[C_k]$.

$$\max_k \mathbb{E}[C_k] = \max_k \frac{1}{p_g}^{\frac{k}{2}\left(2\log_{\frac{1}{p_g}}|V_d|-k+1\right)} = \frac{1}{p_g}^{\max_k \frac{k}{2}\left(2\log_{\frac{1}{p_g}}|V_d|-k+1\right)}.$$

Taking the partial derivative of $\frac{k}{2}\left(2\log_{\frac{1}{p_g}}|V_d|-k+1\right)$ with respect to k we determine the maximum:

$$\arg\max_k \frac{k}{2}\left(2\log_{\frac{1}{p_g}}|V_d|-k+1\right) = \log_{\frac{1}{p_g}}|V_d|+1.$$

Thus we are able to bound:

$$\frac{\log_{\frac{1}{p_g}}|V_d|+1}{2}\left(2\log_{\frac{1}{p_g}}|V_d| - \log_{\frac{1}{p_g}}|V_d| - 1 + 1\right) = \frac{\log_{\frac{1}{p_g}}|V_d|+1}{2}\left(\log_{\frac{1}{p_g}}|V_d|\right) = \frac{1}{2}\log^2_{\frac{1}{p_g}}|V_d| + \frac{1}{2}\log_{\frac{1}{p_g}}|V_d|$$

Which yields the bound:

$$\mathbb{E}[C_k] \leq \frac{1}{p_g}^{\frac{1}{2}\log^2_{\frac{1}{p_g}}|V_d|+\frac{1}{2}\log_{\frac{1}{p_g}}|V_d|} = \sqrt{|V_d|^{\log_{\frac{1}{p_g}}|V_d|+1}}.$$

To complete the proof, we utilize Markov's inequality with $p\left(C_k \geq \frac{|V_d|}{\delta}\sqrt{|V_d|^{\log_{\frac{1}{p_g}}|V_d|+1}}\right) \leq \frac{\delta}{|V_d|}$ and utilize the union bound over n choices of k:

$$\sum_k C_k \leq \sum_k \frac{|V_d|}{\delta}\sqrt{|V_d|^{\log_{\frac{1}{p_g}}|V_d|+1}} = \frac{1}{\delta}\sqrt{|V_d|^{\log_{\frac{1}{p_g}}|V_d|+5}}$$

with probability $1-\delta$. □

Now that we have bounded both the number of cliques, as well as the sizes of the maximal cliques with high probability, we now consider the mutual information of the kernel constructed by summing the kernels corresponding to the maximal cliques of the dependency graph.

Lemma 6. *Define $I(\mathbf{y}_A; v) \triangleq \mathrm{H}(\mathbf{y}_A) - \mathrm{H}(\mathbf{y}_A \mid v)$ as the mutual information between $\mathbf{y}_A$ and v with $\mathrm{H}(\mathcal{N}(\mu,\Sigma)) \triangleq \frac{1}{2}\log|2\pi e\Sigma|$ as the entropy function. Define $\gamma_T^k \geq \max_{A\subset\Theta:|A|=T} I(\mathbf{y}_A; v)$ when $v \sim GP(0, k(\theta,\theta'))$. Let $[k_i]_{i=1,\ldots,M}$ be arbitrary kernels defined on the domain Θ with upper bounds on mutual information $[\gamma_T^{k_i}]_{i=1,\ldots,M}$, then the following holds true:*

$$\gamma_T^{\sum_i k_i} \leq M^2 \max[\gamma_T^{k_i}]_{i=1,\ldots,M}.$$

To prove the above, we first state Weyl's inequality for convenience:

Lemma 7. *Let $H, P \in \mathbb{R}^{n\times n}$ be two Hermitian matrices and consider the matrix $M = H + P$. Let $\mu_i, \nu_i, \rho_i, i = 1, \ldots, n$ be the eigenvalues of M, H, and P respectively in decreasing order. Then, for all $i \geq r + s - 1$ we have*

$$\mu_i \leq \nu_r + \rho_s.$$

The above has an immediate Corollary as noted by Rolland et al. (2018):

Corollary 5. *Let $K_i \in \mathbb{R}^{n\times n}$ be Hermitian matrices for $i = 1, \ldots, M$ with $K \triangleq \sum_i^M K_i$. Let $[\lambda_\ell^{K_i}]_{\ell=1,\ldots,n}$ denote the eigenvalues of K_i in decreasing order. Then for all $\ell \in \mathbb{N}_0$ such that $\ell M + 1 \leq n$ we have*

$$\lambda_{\ell M+1}^{K} \leq \sum_{i=1}^{M} \lambda_{\ell+1}^{K_i}.$$

We are now ready to prove Lemma 6 using Weyl's inequality and its corollary as a key tool.

Proof. Given the definition of $I(\mathbf{y}_A; v) \triangleq \frac{1}{2}\log|\mathbf{I} + \sigma^{-2}\mathbf{K}_A^k|$ (Srinivas et al., 2010) we bound the eigenvalues of $M\mathbf{I} + \sigma^{-2}\sum_i^M \mathbf{K}_A^{k_i}$ using the eigenvalues of $[I + \sigma^{-2}\mathbf{K}_A^{k_i}]_{i=1,\ldots,M}$ where $k \triangleq \sum_{i=1}^M k_i$. Using the above Corollary we see that:

$$\lambda_\ell^{M\mathbf{I}+\sigma^{-2}K} \leq \sum_{i=1}^{M} \lambda_{\lceil \frac{\ell}{M} \rceil}^{I+\sigma^{-2}K_i}.$$

Given the above, we see that $M^2 \max[\gamma_T^{k_i}]_{i=1,\ldots,M} \geq \frac{1}{2}\log|\mathbf{I} + \sigma^{-2}\mathbf{K}_A^k|$ as $\sum_i^M M\gamma_T^{k_i} \geq \frac{1}{2}\log|M\mathbf{I} + \sigma^{-2}\sum_i^M \mathbf{K}_A^{k_i}|$.

□

Finally, we require an additional helper lemma to bound the supremum and infimum of a function sampled from a GP. This helper lemma helps bound the regret during the first phase of HA-GP-UCB where we randomly sample the Hessian over the domain.

Lemma 8. *Let $k(\theta, \theta')$ be four times differentiable on the continuous domain $\Theta \triangleq [0, r]^D$ for some bounded r (i.e., compact and convex) with $f \sim GP(0, k(\theta, \theta'))$ then for all $\delta \in (0, 1)$ the following holds true:*

$$\sup_{\theta\in[0,r]^D} f \leq c_b\sqrt{D\log\delta^{-1}} = \mathcal{O}\left(\sqrt{D\log\delta^{-1}}\right).$$

$$\inf_{\theta\in[0,r]^D} f \geq -c_b\sqrt{D\log\delta^{-1}} = \Omega\left(-\sqrt{D\log\delta^{-1}}\right).$$

for some constant c_b dependent on δ and r, with probability $1 - \delta$.

Proof. We refer readers to Srinivas et al. (2010) Lemma 5.8 for the proof of the above. □

We are now ready to prove Theorem 2.

Theorem 2. *Let k be the kernel as in Assumption 1, and Theorem 1. Let $\gamma_T^k(d) : \mathbb{N} \to \mathbb{R}$ be a monotonically increasing upper bound function on the* mutual information *of kernel k taking d arguments. The cumulative regret of* HA-GP-UCB *is bounded with high probability as follows:*

$$R_T = \tilde{\mathcal{O}}\left(D^{4.5}\log^2 D + \sqrt{T\beta_T D^{\mathcal{O}(\log D)}\gamma_T^k(\mathcal{O}(\log D))}\right). \tag{4}$$

We restate the above theorem with more precision:

Theorem 2. *Let k be the kernel as in Assumption 1, and Theorem 1 and for some constants a, b,*

$$P\left[\sup_{\theta\in\Theta}\left|\frac{\partial v}{\partial\theta_i}\right| > L\right] \leq ae^{-(L/b)^2}, i = 1, \ldots, D.$$

Let $\gamma_T^k(d) : \mathbb{N} \to \mathbb{R}$ be a monotonically increasing upper bound function on the mutual information *of kernel k taking d arguments. Let $k(\theta, \theta')$ be four times differentiable on the continuous domain $\Theta \triangleq [0, r]^d$ for some bounded r (i.e., compact and convex). For any $\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6 \in (0, 1)$. Let, $\tilde{t} \triangleq t - T_0C_1$ and let*

$$\beta_t = 2\log(\tilde{t}^2 2\pi^2/3\delta_6^2) + 2D\log(\tilde{t}^2 Dbr\sqrt{\log(4Da/\delta_6)})$$

The cumulative regret of HA-GP-UCB *is bounded:*

$$P\left[R_T \leq 2C_1^2 c_b\sqrt{D\log\delta_5^{-1}} + \sqrt{C_2 T\beta_T\gamma_T} + 2 \quad \forall T \geq 1\right] \geq 1 - \delta_1 - \delta_2 - \delta_3 - \delta_4 - \delta_5 - \delta_6$$

when $C_1 = \frac{8D^2}{\delta_1^2}\log\frac{2D^2}{\delta_1}\frac{\sigma_n^2}{\sigma_h^2} + \frac{D^2}{p_h\delta_2} + 1$, $C_2 = 8/\log(1 + \sigma^{-2})$, and $\gamma_T = \frac{1}{\delta_4^2}D^{\log_{1/p_g} D+5}\gamma_T^k\left(2\log_{1/p_g} D + 2\sqrt{\log_{1/p_g} D/\delta_3 + 1}\right)$ where c_b is some constant dependent on δ_5.

Proof. The proof is a consequence of the helper lemmas and theorems we have proved. First we consider Phase 1 of HA-GP-UCB where $t \leq T_0$. By Theorem 1, at most $T_0C_1 = C_1^2$ queries will be made during Phase 1, and Lemma 8 indicates the maximum regret for any query. Consulting the respective Theorem and Lemma, we are able to bound the cumulative regret during Phase 1 by:

$$2C_1^2 c_b\sqrt{D\log\delta_5^{-1}} = \mathcal{O}(D^{4.5}\log^2 D).$$

Considering Phase 2, we utilize Lemma 4, Lemma 5, Lemma 6 to bound the mutual information of the sampled kernel with high probability. The number of cliques is given by:

$$\frac{1}{\delta_4}\sqrt{D^{\log_{1/p_g} D+5}} = D^{\mathcal{O}(\log D)}.$$

The size of the largest clique is given by:

$$2\log_{1/p_g} D + 2\sqrt{\log_{1/p_g} D/\delta_3 + 1} = \mathcal{O}(\log D).$$

Following Lemma 6, we may bound the mutual information by:

$$\mathcal{O}(D^{\mathcal{O}(\log D)}\gamma_T^k(\mathcal{O}(\log D))).$$

The proof is complete by leveraging the connection between mutual information and cumulative regret as shown by Srinivas et al. (2010) where $\tilde{\mathcal{O}}$ is the same as $\mathcal{O}$ with the $\log$ factors suppressed. □

G ON THE SURROGATE HESSIAN, $\mathbf{H}_\pi$

In Section 4.5 we remarked that although we cannot observe $\mathbf{H}_v$, we can observe a surrogate hessian, $\mathbf{H}_\pi$ which is related to $\mathbf{H}_v$ by the chain rule. We justify our choice here with showing how $\mathbf{H}_\pi$ is an important sub-component of $\mathbf{H}_v$ (Skorski, 2019). Although the reasoning we give is in one dimension, an analogous argument can be made in arbitrary dimensions using the chain rule for vector-valued functions yielding the Hessian tensor (Magalhães, 2020). We have $v : \Theta \to \mathbb{R}$ is a function of the policy π and can be expressed as a composition of functions:

$$v : \Theta \to \mathbb{R} = \hat{v}\left(\pi\left(\theta\right)\right).$$

In the above we use $\pi\left(\theta\right)$ as shorthand for $\pi\left(\mathbf{s}^\alpha, \mathbf{a}^\alpha; \theta\right)$ with $\hat{v}$ representing some unknown function. Using the definition of the Hessian we have:

$$\mathbf{H}_v \triangleq \left[\frac{\partial^2 v}{\partial\theta^a \partial\theta^b}\right]_{a,b=1,\ldots,D} = \left[\frac{\partial^2}{\partial\theta^a \partial\theta^b}\ \hat{v}\left(\pi\left(\theta\right)\right)\right]_{a,b=1,\ldots,D}$$

Where the above identity follows from the definition of v in Eq. equation G. We can now apply chain rule to express:

$$\frac{\partial^2}{\partial\theta^a \partial\theta^b}\ \hat{v}\left(\pi\left(\theta\right)\right) = \underbrace{\left[\mathbf{H}_{\hat{v}}(\pi(\theta))\frac{\partial\pi}{\partial\theta^a}(\theta)\right] \cdot \frac{\partial\pi}{\partial\theta^b}(\theta)}_{r(\theta)} + \underbrace{\frac{\partial^2\pi}{\partial\theta^a\partial\theta^b}\left(\theta\right)}_{\mathbf{H}_\pi(\theta)} \cdot \underbrace{\nabla\hat{v}(\pi(\theta))}_{g(\theta)}$$

As we see in the above as a consequence of the chain rule, $\frac{\partial^2\pi}{\partial\theta^a\partial\theta^b}$ forms an important sub-component $\frac{\partial^2 v}{\partial\theta^a\partial\theta^b}$. Given the above, we can simplify the above in the following manner:

$$\mathbf{H}_v = r + \mathbf{H}_\pi \circ g$$

where r, g, and $\mathbf{H}_\pi$ arise from the corresponding highlighted terms in Eq. equation G with r representing some unknown remainder term and $\circ$ representing the Hadamard product. Given the above, it is straightforward to see how $\mathbf{H}_\pi$ serves as a surrogate hessian for $\mathbf{H}_v$. Indeed if $r \neq -\mathbf{H}_\pi \circ g$ and g has no zero entries then $\mathbf{H}_\pi \neq 0 \implies \mathbf{H}_v \neq 0$. In our use case, we are most concerned with non-zero entries in the Hessian, $\mathbf{H}_v$, and the surrogate Hessian, $\mathbf{H}_\pi$ is well served for determining $\mathbf{H}_v \neq 0$ due to the above.

Since $\pi\left(\theta\right)$ is shorthand for $\pi\left(\mathbf{s}^\alpha, \mathbf{a}^\alpha; \theta\right)$, to approximate $\mathbf{H}_\pi$ we average $\mathbf{H}_{\pi(\mathbf{s}^\alpha, \mathbf{a}^\alpha; \theta)}$ over state action pairs, $(\mathbf{s}^\alpha, \mathbf{a}^\alpha)$ formed through interaction of the policy with the unknown task environment.

A possible avenue of overcoming this limitation is considering Hessian estimation through zero'th order queries. Several works along this direction have recently appeared using Finite Differences (Cheng et al., 2021), as well as Gaussian Processes (Müller et al., 2021). We consider removing this dependency on the surrogate Hessian for future work.

H DRONE DELIVERY TASK

Drones fly from delivery point to delivery point where completing a delivery gives a large amount of reward, but running out of fuel and collisions give a small amount of negative reward. After completing a delivery, the delivery point is randomly removed within the environment. A collision gives a small amount of negative reward and momentarily stops the drone. Completing a delivery refills the drone fuel and allows it to continue to make more deliveries. The amount of reward given increases quadratically with the distance of the delivery to highly reward long distance deliveries which require long term planning. To compound this requirement for long term planning, fuel consumption also dramatically increases at high velocities to encourage long-term fuel efficiency planning. In this complex scenario requiring long term planning, RL approaches can easily fall into local minima of completing short distance, low reward deliveries and fail to sufficiently explore (under sparse reward) policies which complete long distance deliveries with careful planning.

Implementation code of this task can be found in supplementary materials.

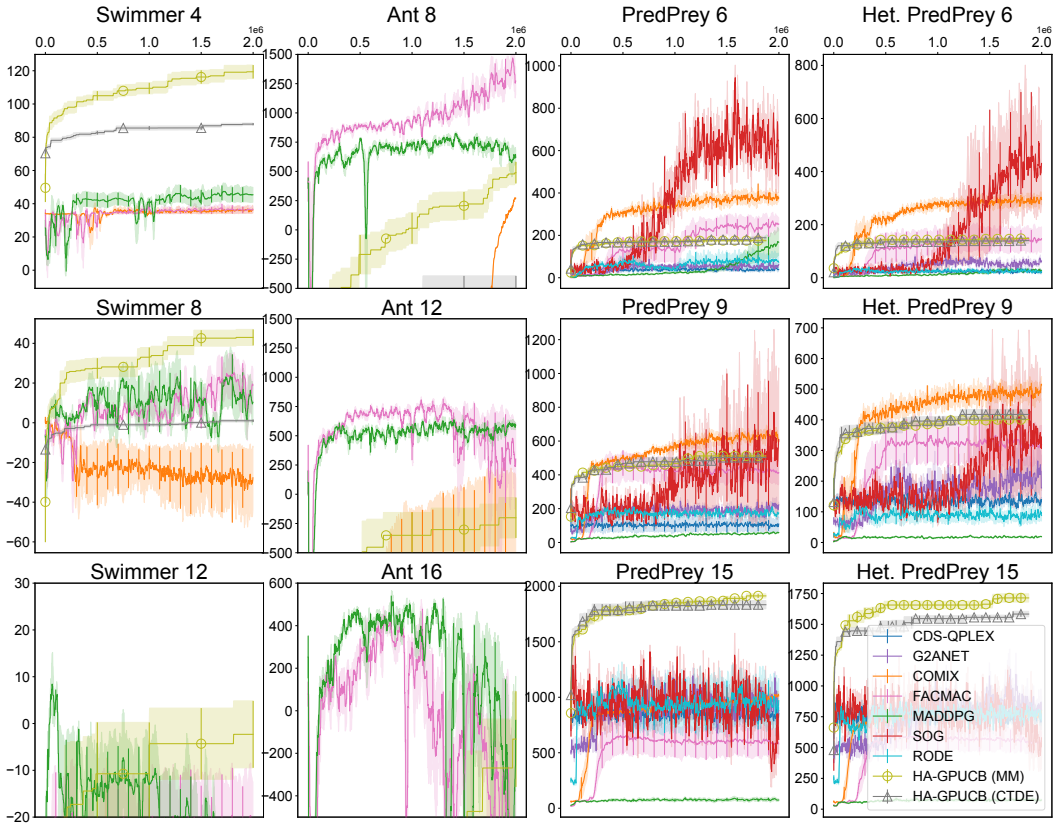

Figure 10: Comparison with MARL approaches with varying number of agents.

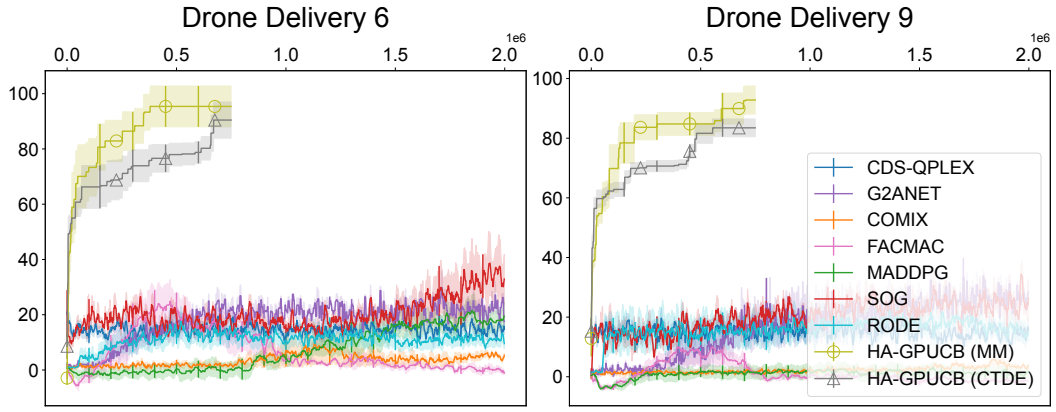

Figure 11: Comparison with MARL approaches on the drone delivery task.

I REPLOT WITH TIMESTEPS

We replot the relevant figures in Fig. 10 and Fig. 11 while maintaining total environment interactions as the singular independent variable. We note that there is no significant change to our conclusions as a consequence of this replotting. We also highlight that although total environment interactions is considered the important independent variable in RL and MARL, in BO typically the total evaluated policies is considered the more important independent variable as each evaluation is assumed to be costly.