2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Context-Generative Default Policy for Bounded Rational Agent

Durgakant Pushp, Junhong Xu, Zheng Chen and Lantao Liu

Abstract— **Bounded rational agents often make decisions by evaluating a finite selection of choices, typically derived from a reference point termed the 'default policy,' based on previous experience. However, the inherent rigidity of the static default policy presents significant challenges for agents when operating in unknown environment, that are not included in agent's prior knowledge. In this work, we introduce a context-generative default policy that leverages the region observed by the robot to predict unobserved part of the environment, thereby enabling the robot to adaptively adjust its default policy based on both the actual observed map and the *imagined* unobserved map. Furthermore, the adaptive nature of the bounded rationality framework enables the robot to manage unreliable or incorrect imaginations by selectively sampling a few trajectories in the vicinity of the default policy. Our approach utilizes a diffusion model for map prediction and a sampling-based planning with B-spline trajectory optimization to generate the default policy. Extensive evaluations reveal that the context-generative policy outperforms the baseline methods in identifying and avoiding unseen obstacles. Additionally, real-world experiments conducted with the Crazyflie drones demonstrate the adaptability of our proposed method, even when acting in environments outside the domain of the training distribution.**

I. INTRODUCTION

Robotic autonomy in the real-world demands efficient decision-making despite constraints on sensing and computational capabilities. For instance, when navigating through a cluttered room, a robot cannot see through obstacles or behind walls, leading to a partial and potentially inaccurate understanding of its surroundings. Hence, the decision-making process operates within the constraints of this limited information, often leading to suboptimal choices such as taking inefficient paths that may lead to dead-ends due to occlusion. These limitations highlight a sharp disparity with biological agents like humans, who effortlessly excel in tasks that robots find challenging. Despite these constraints, humans possess the remarkable ability to quickly select among a multitude of options by effectively narrowing down the search space. Leveraging their prior knowledge, humans constrain the range of choices available to them, opting for a satisfactory and sufficient (termed as *satisficing* [1]) solution rather than an exhaustive pursuit of the optimal one. This decision-making paradigm is recognized as a bounded-rational decision-making process. Several models have emerged to tackle the challenges of bounded rationality (BR) among these, information-theoretic bounded rationality [2]–[4] stands out as a comprehensive framework explicitly accounting for a robot's computational limitations by imposing constraints on the amount of information it can process when transitioning from a default policy. The default policy serves as a collection of favorable choices from which an intelligent agent derives a satisficing solution for a given task and hence the efficacy of this default policy is highly reliant on prior experiences and accumulated knowledge. Recent advancements in this field [5], have assumed a default policy in the form of a normal distribution with no prior task-specific knowledge. Subsequent research [6] has improved upon this approach by introducing a task-conditioned default policy that incorporates some prior knowledge. However, this prior work still assumes complete knowledge of the environment and relies on pre-trained goal-conditioned policies for navigating to any goal location within that environment. In scenarios involving partially known environments, we need more informative default policies that consider not only the goal but also the partial observations.

Rather than relying on a static default policy, in this study, we argue that the belief regarding favorable choices should dynamically evolve as new information becomes available, motivating the development of a context-generative default policy. We present a novel approach that leverages observed environmental information to predict the unobserved space to facilitate adaptive adjustments in the agent's prior beliefs, aligning with the evolving knowledge of the world. This prior belief referred to as 'context-generative default policy' is guided by design principles outlined in [6]. We apply this approach to address the navigation challenge in unknown environments, achieving enhanced efficiency, as shown by the results of extensive simulation and real-world experiments.

II. RELATED WORK

The rationality assumption in the decision making process demands the evaluation of numerous decision options [7], [8]. One approach to mitigate this problem, as discussed in prior work [9]–[12], is through meta-reasoning, wherein a robot examines the costs associated with selecting a choice by finding a balance between expected reward and the associated computational cost to evaluate the choice. However, this approach does not fundamentally reduce the number of choices, as they still optimize a transformed cost function under the assumption of rationality. The concept of bounded rationality [10], [11], [13]–[15] addresses this challenge by reducing the number of options evaluated based on prior task knowledge. Among various modeling approaches, the information-theoretic approach is particularly well-suited to robotics applications [2], [3]. This approach defines the decision problem as the minimization of divergence between the agent's default policy and the optimal policy [16], [17], contending that a more informed default policy can lead to improved performance while evaluating a smaller set of

The authors are with the Luddy School of Informatics, Computing, and Engineering at Indiana University, Bloomington, IN 47408, USA.
E-mail: {dpushp, xu14, zc11, lantao}@iu.edu.
This work was supported by NSF #2006886 and #2047169.

choices. However, prior works have primarily emphasized either uninformative default policies [5] or goal-conditioned informed policies tailored to specific environments [6], which may prove insufficient when confronted with novel environments. We argue that the default policy should account for both the task and environmental factors to generate more informative candidate choices. Our approach achieves this by integrating predictions about environmental characteristics into the information-theoretic framework.

Traditional methods in robotics [18]–[20] usually overlook unobserved areas during the planning. While these methods can rapidly identify feasible trajectories, they often encounter difficulties in escaping local minima. This challenge is tackled by progressively constructing a map during navigation [21]–[23] but results in unnecessary exploration of the environment. Recent works involve using map prediction techniques to enhance the navigation efficiency as outlined in [24]–[26]. These methods introduce various approaches to map prediction and employ existing planners for motion control by assigning a lower cost to the predicted map. However, their performance is limited by the reliability of map prediction methods. Our proposed method adopts a similar approach to finding a default policy. However, unlike existing methods, it does not directly execute actions from the default policy. Instead, it evaluates a finite set of trajectories sampled from the default policy to effectively mitigate the unreliability of prediction methods.

III. Method

We consider the problem of generating trajectories for a robot in an unknown environment represented by $e \in \mathcal{E}$, where e can be any environment representation, i.e., occupancy or semantics; $\mathcal{E}$ is the set of all possible environments. Let o_t be the area observed by the robot at given time t in the given environment e. We define the context $c_t := \bigcup_{t=1}^{T} o_t$ where $c_t \subseteq e$. The goal of the robot is to find the trajectory τ that maximizes an expected reward given by

$$J(\tau, s_t, g, c_t) = \sum_{k=t}^{t+H} R(s_t, a_t, g, e), \tag{1}$$

where $\tau = \{s_{t+1}, a_{t+1}, s_{t+2}, a_{t+2}, ..., s_{t+H-1}, a_{t+H-1}\}$, s is the state, a is the action, g is the goal, H is the planning horizon and c is the context (e.g., a partially observed occupancy map). We consider a deterministic dynamics $s_{t+1} = f(s_t, a_t)$. We interchangeably use the terms τ and a_t since we can iteratively apply action sequences to generate the trajectory τ.

Note that, diverging from the planning objective established in the prior work [6], the utility function J now incorporates the contextual variable c_t derived from the robot's operating environment e. Therefore, in this section, we first discuss the adjustments made in the problem formulation to integrate contextual variable and then introduce the concept of a context-generative default policy. Following that, we delve into the Bounded-Rational Policy Search Algorithm, which leverages the context-generative default policy for informed decision-making.

A. Problem Modeling

In principle, determining the optimal trajectory typically involves exhaustively evaluating every potential trajectory $\mathcal{T}$ within the state space when the environment e is entirely known. However, such an exhaustive approach is often infeasible due to computational constraints inherent in robotic systems. Information-theoretic bounded rationality (ITBR) addresses this issue more realistically by explicitly accounting for agents' computational constraints in balancing computational resources against the number of evaluated trajectories [3]. This is achieved through a constraint on maximum expected utility by quantifying the amount of information the agent can process to transition from a default policy Q to the optimal policy and formulated using KL-divergence as

$$\tau_t^* = \arg\max_{\tau \sim Q} \{J(\tau, s_t, g, c_t)\}, \text{ s. t. } D_{KL}(\pi_t || Q) \leq K, \tag{2}$$

where τ is the trajectory sampled from the default policy, π_t is the current policy of the robot, D_{KL} is the KL divergence between the two stochastic policies, and K is a constant denoting the amount of information an agent can deviate from the default policy. The default policy characterizes the robot's expected behavior prior to starting the navigation task. In bounded rational settings, a robot aims to find a *satisficing* trajectory within the proximity of the default policy. Using Lagrange multiplier, Eq. (2) can further be reduced as:

$$\tau_t^* = \arg\max_{\tau \sim Q} \{J(\tau, s_t, g, c_t) \ - \Gamma(\beta, Q)\}, \tag{3}$$

where $\Gamma(\beta, Q) = \frac{1}{\beta} D_{KL}(\pi_t || Q)$, and $\beta > 0$ indicates the *rationality level*. Eq. (3) offers an enhanced capacity for modulating trade-off dynamics through the rationality parameter β and the default policy Q. In practice, the rationality level β is linked to the computational capacity of the robot's hardware. Enhanced computational power allows for a more exhaustive evaluation of trajectories prior to decision-making, resulting in more rational decision outcomes. Since hardware flexibility is typically limited in robots, the degree of rationality predominantly relies on the informativeness of choices derived from the default policy. In this work, we introduce a novel perspective by associating the degree of rationality with the completeness of the robot's perception. Specifically, a greater range of perception enables evaluating trajectories more rationally, thereby facilitating escaping from potential local optima. Therefore, we propose a dynamic default policy that relies on the current state s_t, goal g, and context c_t, defined as: $Q_t(a_t \,|\, s_t, g, c_t)$ where Q_t denotes the default policy and a_t represents an action at time t. In this work, we treat the β as a constant and aim to find a more informative default policy to improve robot's behavior.

Formally, the robot navigation under bounded-rational decision process is defined as a tuple $(\beta, \mathcal{T}_{Q_t}, Q_t, J)$, where $\mathcal{T}_{Q_t} = \{\tau \sim Q_t \,|\, \text{Distance}(\tau, Q_t) \leq \epsilon\}$ represents the finite set of trajectories that are within a neighborhood (defined by ϵ) of the default policy Q_t. The goal of the robot is to find a *satisificing* trajectory $\tau^* \sim \mathcal{T}_{Q_t} \subset \mathcal{T}$ that maximizes the

following equation, derived from Eq. (3) following the prior work [5], [6]:

$$\tau_t^* = \underset{\tau \sim \mathcal{T}_{Q_t}}{\arg\max} \{-D_{KL}(\pi_t || \phi_t)\}, \tag{4}$$

where $\phi_t(a_t|s_t, g, c_t) \propto Q_t(a_t \mid s_t, g, c_t) e^{\beta J(\tau, s_t, g, c_t)}$.

B. Context-Generative Default Policy

In the ITBR framework, the quality of solutions is significantly influenced by the default policy choice. The default policy guides which part of the state space to search for the bounded-optimal policy. Previous work leverages goal-conditioned reinforcement learning to compute an informative default policy. When the robot is deployed in a similar environment with a *known* map, with the goal-conditioned informed policy, the agent finds a better policy using a small number of trajectory evaluations than the uniform (uninformed) one. However, when the environment map is not known a-prior, this informed policy can no longer provide correct guidance for the policy search. Real-world environments typically exhibit consistent patterns and structures. Such regularities suggest that the unobserved segments of the environment can be extrapolated from the observed regions. Conditioning the informative default policy on the unobserved regions allows the agent to utilize these environmental regularities to better guide the bounded-rational policy search to avoid potentially unobserved yet impassable regions.

Assume that we are given a model $\psi(\tilde{e}|c_t; \theta)$ that facilitates the prediction of a complete map conditioned on the context c_t. We use $\tilde{e}$ to denote the map sampled from the model ψ parameterized by θ. To address the optimization problem outlined in (4), it is imperative to define Q_t. Hence, we adopt the design principles proposed in [6] that states the designed default policy should be informative i.e., the sampled action sequences covers the high-utility regions, and it should be adaptive to changes. To ensure the informativeness, we use the model ψ to anticipate the unseen environment and then use a sampling based-planner such as *RRT** on predicted map $\tilde{e}$ along with the *B-spline* trajectory optimization to get the mean of Q_t. Finally, the path obtained from this planning process is truncated to the planning horizon H. We define the default policy as:

$$Q_t(\tau \mid \bar{p}, \Sigma) = \frac{1}{N} \exp\left(-\frac{1}{2}(\tau - \bar{p})^T \Sigma^{-1} (\tau - \bar{p})\right) \tag{5}$$

where $Q_t(\tau \mid \bar{p}, \Sigma) = Q_t(\tau \mid \gamma(s_t, g, \psi(\tilde{e}|c; \theta)), \Sigma)$ with mean trajectory $\bar{p}$ obtained from the path planning heuristics γ along which the policy is centered, the covariance matrix Σ, $N = \sqrt{(2\pi)^{H-1}|\Sigma|}$ is the normalization constant and τ is the trajectory obtained by rolling out the sampled actions from the distribution. Here, Σ remains a tunable parameter, initially set to a low value at s_t and gradually increased until H is reached as shown in Fig. 1. The utility of the trajectories sampled from Q_t depends on the accuracy of model ψ. If the predicted map significantly deviates from reality, it can diminish the informativeness of Q_t. Consequently, addressing inaccuracies in the map prediction becomes necessary to avoid such scenarios.

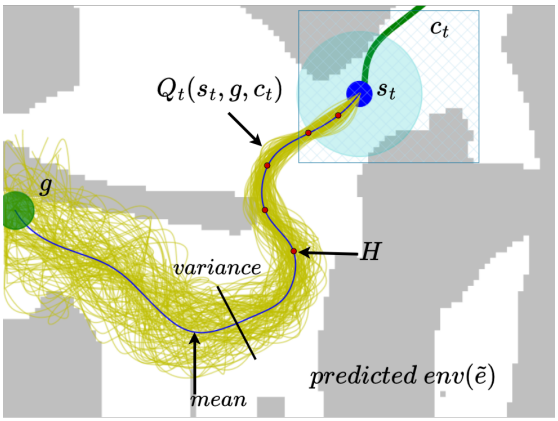


Fig. 1. Snapshot of the navigation task at time t, illustrating the default policy distribution. The predicted environment $\tilde{e}$ is derived from the context c_t, indicated by the cross-hatched square. The yellow trajectories, sampled from Q_t, extend the horizon to the goal to enhance comprehension of the default policy but they can be truncated to any desired value during code implementation, as depicted by the red dots.

C. Handling Map Prediction Errors

Initially, in scenarios where robots have not yet gathered sufficient environmental data, we observe that $\tilde{e}$ often tends to be inaccurate. This inaccuracy arises due to insufficient ground truth map input for the map prediction model. This situation is analogous to when biological agents navigate through unfamiliar environment; their expectations regarding unobserved area are typically inaccurate. However, intelligent agents possess the capability to make decisions even in the presence of highly uncertain or potentially erroneous data. This capacity is referred to in the literature as 'bounded information,' and real-world intelligent agents adeptly handle these limitations, making decisions efficiently despite constraints on time or the availability of accurate information. Addressing these complex issues is challenging, and there is a well-documented argument suggesting that humans employ simple heuristics to navigate such scenarios, as discussed in [27], [28]. Hence, we take the following measures:

- We assume that the goal always lies within navigable space and that a feasible path to the goal exists. When γ fails to find a viable path, we adapt it to seek the nearest feasible path to the goal by directly linking the last accessible node to the goal and assigning minimal rewards to unreliable predictions, while still considering them as free to navigate.
- We further evaluate a finite number of trajectories proportional to the rationality number β, sampled from the default policy to find more informative path.

This approach acknowledges the evident inaccuracies in the map prediction and offers a practical means of managing inaccurate information within bounded rational robotic agents.

D. Context-Generative Bounded-Rational Policy Search

The proposed method combines the above-mentioned context-generative default policy with ITBR framework. We compute the utility function by replacing e with $\tilde{e}$ in Eq. (1) assuming the predicted map $\tilde{e}$ is equivalent to the ground

truth map e given as:

$$J(\tau, s_t, g, c_t) = \sum_{k=t}^{t+H} R(s_t, a_t, g, \psi(\tilde{e}|c_t; \theta)), \tag{6}$$

Note that the Eq. (4) obtained its maximum value when $D_{KL}(\pi_t||\phi_t) = 0$ as KL-divergent being non-negative i,e., $\pi_t = \phi_t$ leading to the optimal action as shown below:

$$\pi^*(a_t \mid s_t, g, c_t) = \frac{1}{\lambda} Q_t(a_t) e^{\beta J(\tau, s_t, g, c_t)} \tag{7}$$

where $\lambda = \int Q_t(a_t) e^{\beta \sum_{k=t}^{t+H} R(s_t, a_t, g, \psi(\tilde{e}|c_t; \theta))} \, da_t$ is the normalization constant. Finally, the expected optimal action sequences is obtained by importance sampling given by:

$$\mathbb{E}_{a_t^* \sim \pi_t^*}[a_t \mid s_t, g, c_t] \approx \frac{\sum_{k=1}^{L} w(a_{t,k}) a_{t,k}}{\sum_{k=1}^{L} w(a_{t,k})}. \tag{8}$$

where $w(a_{t,k}) = e^{\beta \sum_{k=t}^{t+H} R(s_{t,k}, a_{t,k}, g, \psi(\tilde{e}|c_t; \theta))}$ and L is the number of samples to be evaluated. In summery, the robot navigation involves sampling L actions from a finite number of choices $\mathcal{T}_{Q_t}$ obtained from Q_t. Subsequently, each sequence is assigned a weight w that reflects its significance according to the agents' utilities and level of rationality and the reward function is computed by rolling out future states along the action sequence until the planning horizon ends. Finally, the expected optimal action is obtained by Eq. (8).

IV. Simulation Experiments

A. Simulation Setup and Performance Metrics

We conducted two experiments to validate our approach. The first evaluates the robot's navigation abilities, focusing on its capacity to avoid challenging obstacles in a 2D environment. The second experiment analyzed the influence of initial context (landmark knowledge) on the performance.

Map Prediction Model: We use the *nuScenes* dataset for training our map-prediction module, as documented in [29]. We convert the semantic images into occupancy maps by categorizing the classes into two groups: navigable and non-navigable. The resulting dataset is divided into two subsets: a training set comprising 28,008 images and a separate testing set consisting of 500 images. All the simulation testing environments are drawn from this pool of 500 images. We use the SePaint model [30] to predict the environment.

Performance Metric: We employ a set of robust metrics that includes *path length*, *navigation efficiency* (N_{eff}), and *map prediction accuracy* (M_{Acc}). The navigation efficiency (N_{eff}) is a critical measure, quantifying the change in distance to the goal relative to the explored area at a specific time step t within the given environment. Higher N_{eff} values signify more efficient navigation, indicating that the agent covers a shorter spatial extent while making substantial progress toward its goal. This not only highlights the agent's ability to minimize detours but also emphasise its energy-saving potential for future tasks. Map prediction accuracy (M_{Acc}) is defined as $M_{Acc} = (n_{tp} + n_{tn})/(n_{tp} + n_{tn} + n_{fp} + n_{fn})$, where n_{tp}, n_{tn}, n_{fp}, and n_{fn} represent true positives, true negatives, false positives, and false negatives, respectively.

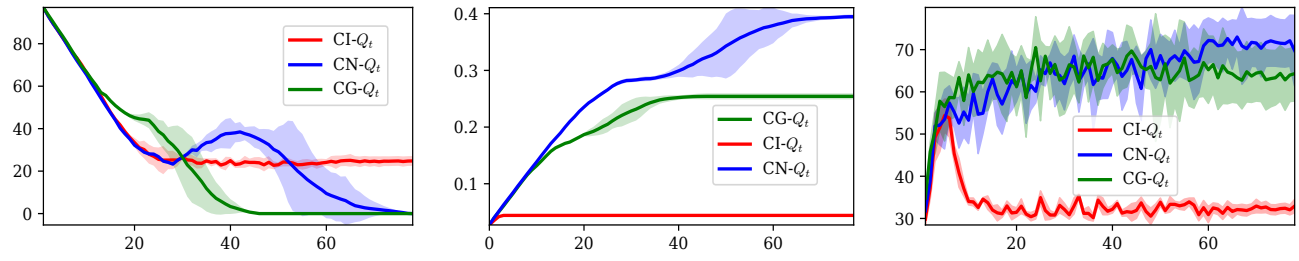
Constant Parameters in ITBR: We employ consistent settings for key parameters. The agents' transition functions are determined by a deterministic single integrator model, where their speed is bounded between $0m/s$ (minimum speed) and $1m/s$ (maximum speed). Our one-step reward function is thoughtfully designed to discourage collisions and substantial deviations from the goal. To ensure consistency in our experiments, we maintain a fixed number of sampled trajectories at 100. Additionally, we keep the rationality level ($\beta = 0.04$) constant throughout our simulations.

Baselines: We use two baselines to compare with the proposed method. The first baseline disregards the context acquired by the agent, treating the unknown areas as free to navigate and generating a default policy similar to [6]. We refer to this baseline as the *"Context-Ignorant Default Policy (CI-Q_t)"*. The second baseline uses the observed map but ignores its potential to predict the unknown map and only rely on the known map for generating the default policy, referred as *"Context-Neutral Default Policy (CN-Q_t)"*. Our proposed method is referred as *"Context-Generative Default Policy (CG-Q_t)"*.

B. Anticipating Obstacles: Beyond Dead Ends

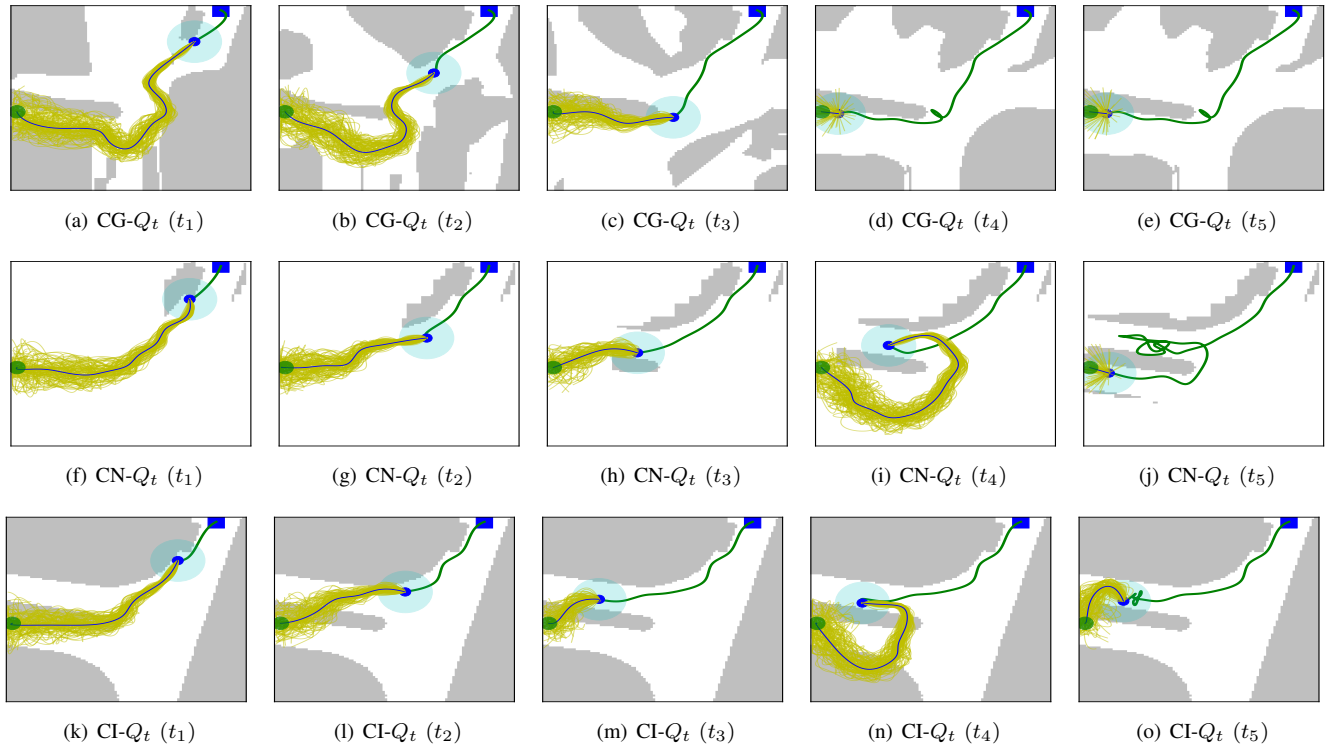
We empirically demonstrate the remarkable anticipatory capabilities of our proposed method in navigating complex environments, specifically in scenarios involving challenging obstacles. To evaluate this, we conducted controlled experiments where we randomly sampled a map from the testing environment, deliberately selecting scenarios with complex obstacles, such as U-shaped barriers placed between the starting point and the goal.

As depicted in Fig. 2(a), our proposed method and the CN-Q_t approach both successfully completed the navigation task. In contrast, the CI-Q_t approach encountered difficulties, evident from the consistent and prolonged tail in the curve, indicative of being trapped in local minima. Furthermore, while the CN-Q_t approach ultimately completed the task, it briefly deviated from the optimal path, leading to an increase in distance to the goal - a clear indication of navigating into a dead end. Fig. 3 shows a detailed comparative illustration of this experiment, with simultaneous snapshots for reference. To evaluate the influence of acquired information during navigation on map prediction methods, we provide the explored area by all the methods at a given time (shown in Fig. 2(b)) as contextual information to the map prediction process and recorded their prediction accuracy. Our findings indicate the substantial impact of incorporating increased context on the map prediction accuracy as shown in Fig. 2(c). This experiment was repeated across 10 randomly selected maps, consistently yielding the same results. These outcomes establish our proposed method as a promising asset for addressing long-range navigation tasks, where ample context is acquired during the initial exploration, leading to improved prediction accuracy and offering a potential solution to real-world challenges in robotics.



(a) Distance to Goal vs Time (b) Explored Area vs Time (c) M_{Acc} vs Time

Fig. 2. Illustration of the performance evaluation with respect to navigation task. The X-axis indicates the time while the Y-axis indicates the distance to goal in (a), the explored area in (b), and map prediction accuracy in (c).



(a) CG-Q_t (t_1) (b) CG-Q_t (t_2) (c) CG-Q_t (t_3) (d) CG-Q_t (t_4) (e) CG-Q_t (t_5)

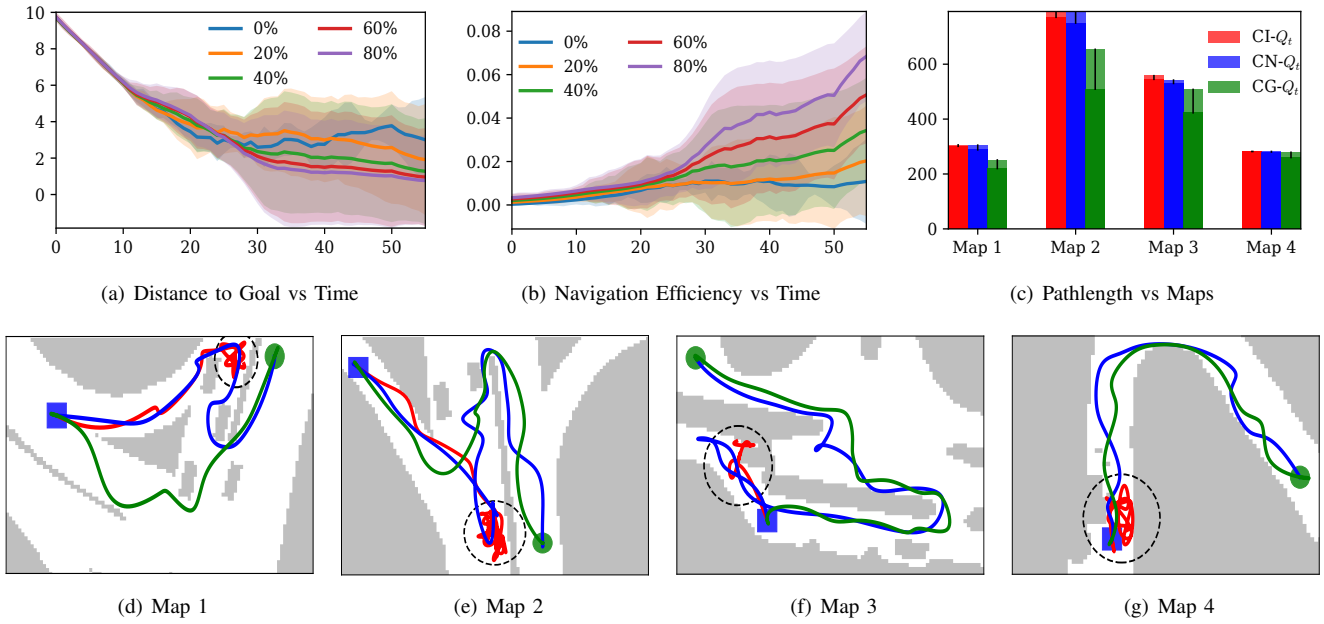
(f) CN-Q_t (t_1) (g) CN-Q_t (t_2) (h) CN-Q_t (t_3) (i) CN-Q_t (t_4) (j) CN-Q_t (t_5)

(k) CI-Q_t (t_1) (l) CI-Q_t (t_2) (m) CI-Q_t (t_3) (n) CI-Q_t (t_4) (o) CI-Q_t (t_5)

Fig. 3. Illustration of a grid environment with the starting point (blue square), robot's current position (blue circle), and the sensor field of view (cyan circle). The path traveled by the robot is shown in green, the yellow paths represent the default distribution with predicted mean path. (a) to (e) showcase the performance of the proposed method on the predicted map. (f) to (j) illustrate the performance of the CN-Q_t method on the observed map. (k) to (o) present the performance of the CI-Q_t method, which only considers sensor's field of view, the paths are overlaid on the ground truth map. All snapshots were captured simultaneously from the starting point to facilitate direct comparison.

C. The Crucial Role of Initial Context

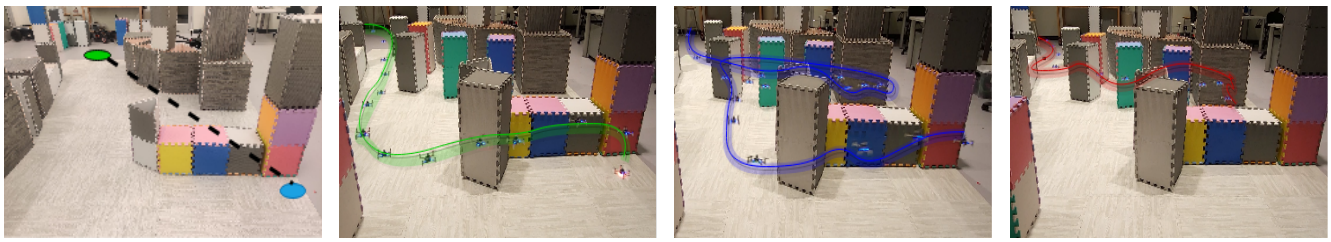
In the domain of map prediction, the significance of the initial context cannot be overstated, as it serves as a critical determinant of reliability. To underscore this, we conducted a comprehensive experiment, wherein we progressively revealed the map's details from 0% to 80%. Our observations of the resulting impact on both the distance to the goal and navigation efficiency revealed compelling insights. As depicted in Fig. 4(a), the distance to the goal exhibited a consistent improvement with each increment in the initial context provided to the proposed method. This, in turn, translated into enhanced navigation efficiency, as illustrated in Fig. 4(b). To fortify our findings, we conducted this experiment across four distinct maps as shown in Fig. 4(c), consistently yielding shorter path lengths. Note that the path length for a failed trial refers to the total distance traveled before the termination of that particular trial. The path lengths further reduced given more initial context (visualised by the vertical black line on each bar in Fig. 4(c)). The path followed by the all the methods are visualised on the maps for better understanding of the performance comparison. In absence of complex obstacles the performance of CN-Q_t and CG-Q_t are the same as shown in the Fig. 4(g). The CI-Q_t fails to achieve the goal in all scenarios as they completely ignored the context. In general, the proposed methods avoids complex obstacles as shown in Fig. 4(d) to Fig. 4(g), however, its performance is influenced by the inital context. This type of initial context is often encapsulated in the form of well-known landmarks - a resource readily available in real-world scenarios. Our work, therefore, offers a systemic approach to incorporating landmark information into planning within the bounds of rationality. This practical approach to navigation showcases



(a) Distance to Goal vs Time (b) Navigation Efficiency vs Time (c) Pathlength vs Maps

(d) Map 1 (e) Map 2 (f) Map 3 (g) Map 4

Fig. 4. Illustration of the impact of providing initial context on the navigation task and the performance of baselines and proposed method in four different environments. (a) and (b) shows the performance of proposed method on increasing initial context where x-axis represents time. Y-axis represents the distance to goal in (a) and navigation efficiency in (b). (c) demonstrates the path length on y-axis for 4 different maps. Note that the vertical black line shows the difference in the improvement for path length when given more initial context. The path travelled by all the methods on the 4 maps are visualised in (d) to (g) in which the start position is represented by blue square and goal is shown by the green circle. The black dotted circles on the map highlights the area where baselines encounter difficulties. Red path represents CI-Q_t, blue path represents CN-Q_t and green path represents CG-Q_t.



(a) Initial Map (b) CG-Q_t Path (c) CN-Q_t Path (d) CI-Q_t Path (Failed)

Fig. 5. Illustrates the experimental setup and snapshots of drones in action. (a) Depicts the initial environment. Drone starts from green circle and goal is represented by light blue circle. (b) Shows the snapshots drone following the path by our proposed method. (c) Demonstrates the path for CN-Q_t. (d) shows the path for CI-Q_t. Note that the testing environment is different than the initial environment.

the relevance of the proposed work to real-world applications and sets the stage for future advancements. The readers are encouraged to watch the experimental videos at `https://www.youtube.com/watch?v=Ial-EDNPSe0`.

V. Physical Experiments

In this section, we show the results of our physical experiments, designed to validate the adaptability and robustness of our proposed method, leveraging the agile Crazyflie 2.1 nano-drones within the motion capture system [31]. We incorporated an observation model with a restricted sensing range to replicate navigation within an unfamiliar map scenario. Our experimental environment recreates real-world challenges by introducing additional obstacles to the initial map that was similar to the maps in the training set. We deliberately added obstacles along the path to the goal. This deliberate perturbation of the environment serves as a litmus test for the adaptability and resilience of our planner, particularly when confronted with maps that slightly deviate from the training distribution. As shown in Fig. 5, all methods adeptly navigated around the obstacles, a testament to the adaptiveness inherent in the bounded rationality framework. However, our approach takes the optimal path to the goal due to a more informative default policy designed to consider predicted map. This transition from simulation to real-world demonstrate the consistency and robustness of our method.

VI. Conclusion

This paper presents a 'context-generative default policy' to address autonomous navigation in unknown environments by leveraging map prediction to anticipate and proactively avoid unseen obstacles. The adaptive nature of the bounded-rationality framework allows the robot to effectively handle unreliable predictions by selectively sampling trajectories in the vicinity of the default policy. Extensive evaluations have demonstrated the superiority of the proposed work. This work presents a systematic approach to integrate context into the planning process, paving the way for future research in which visual inputs in the form of context can be incorporated into the map prediction module.

REFERENCES

[1] H. A. Simon, "Satisficing," in *The New Palgrave: A Dictionary of Economics* (J. Eatwell, M. Milgate, and P. Newman, eds.), vol. 4, pp. 243–245, New York: Palgrave, 1987.

[2] P. A. Ortega, D. A. Braun, J. Dyer, K.-E. Kim, and N. Tishby, "Information-theoretic bounded rationality," *arXiv preprint arXiv:1512.06789*, 2015.

[3] P. A. Ortega and D. A. Braun, "Information, utility and bounded rationality," in *Artificial General Intelligence* (J. Schmidhuber, K. R. Thórisson, and M. Looks, eds.), (Berlin, Heidelberg), pp. 269–274, Springer Berlin Heidelberg, 2011.

[4] H. J. Kappen, V. Gómez, and M. Opper, "Optimal control as a graphical model inference problem," *Machine learning*, vol. 87, no. 2, pp. 159–182, 2012.

[5] J. Xu, D. Pushp, K. Yin, and L. Liu, "Decision-making among bounded rational agents," 2022.

[6] D. Pushp, J. Xu, and L. Liu, "Coordination of bounded rational drones through informed prior policy," 2023.

[7] C. H. Papadimitriou and J. Tsitsiklis, "Intractable problems in control theory," *SIAM journal on control and optimization*, vol. 24, no. 4, pp. 639–654, 1986.

[8] T. L. Dean, "Intractability and time-dependent planning," in *Proceedings of the 1986 workshop on Reasoning about Actions & Plans*, pp. 245–266, 1986.

[9] S. Zilberstein, "27Metareasoning and Bounded Rationality," in *Metareasoning: Thinking about Thinking*, The MIT Press, 03 2011.

[10] S. Zilberstein, "Metareasoning and bounded rationality.," 2011.

[11] S. Russell and E. Wefald, "Principles of metareasoning," *Artificial intelligence*, vol. 49, no. 1-3, pp. 361–395, 1991.

[12] C. Boutilier, T. Dean, and S. Hanks, "Decision-theoretic planning: Structural assumptions and computational leverage," *Journal of Artificial Intelligence Research*, vol. 11, pp. 1–94, 1999.

[13] C. Camerer, "Bounded rationality in individual decision making," *Experimental economics*, vol. 1, pp. 163–183, 1998.

[14] R. Selten, "Bounded rationality," *Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft*, vol. 146, no. 4, pp. 649–658, 1990.

[15] V. Pacelli and A. Majumdar, "Robust control under uncertainty via bounded rationality and differential privacy," 2023.

[16] T. Genewein, F. Leibfried, J. Grau-Moya, and D. A. Braun, "Bounded rationality, abstraction, and hierarchical decision-making: An information-theoretic optimality principle," *Frontiers in Robotics and AI*, vol. 2, p. 27, 2015.

[17] P. A. Ortega and D. A. Braun, "Thermodynamics as a theory of decision-making with information-processing costs," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 469, no. 2153, p. 20120683, 2013.

[18] D. Fox, W. Burgard, and S. Thrun, "The dynamic window approach to collision avoidance," *IEEE Robotics & Automation Magazine*, vol. 4, no. 1, pp. 23–33, 1997.

[19] S. Quinlan and O. Khatib, "Elastic bands: Connecting path planning and control," in *[1993] Proceedings IEEE International Conference on Robotics and Automation*, pp. 802–807, IEEE, 1993.

[20] G. Williams, A. Aldrich, and E. A. Theodorou, "Model predictive path integral control: From theory to parallel computation," *Journal of Guidance, Control, and Dynamics*, vol. 40, no. 2, pp. 344–357, 2017.

[21] B. Yamauchi, "A frontier-based approach for autonomous exploration," in *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97.'Towards New Computational Principles for Robotics and Automation'*, pp. 146–151, IEEE, 1997.

[22] D. Pagac, E. M. Nebot, and H. Durrant-Whyte, "An evidential approach to map-building for autonomous vehicles," *IEEE Transactions on Robotics and Automation*, vol. 14, no. 4, pp. 623–629, 1998.

[23] J.-A. Meyer and D. Filliat, "Map-based navigation in mobile robots:: Ii. a review of map-learning and path-planning strategies," *Cognitive Systems Research*, vol. 4, no. 4, pp. 283–317, 2003.

[24] A. Elhafsi, B. Ivanovic, L. Janson, and m. Pavone, "Map-predictive motion planning in unknown environments," 2019.

[25] K. D. Katyal, A. Polevoy, J. Moore, C. Knuth, and K. M. Popek, "High-speed robot navigation using predicted occupancy maps," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5476–5482, IEEE, 2021.

[26] L. Wang, H. Ye, Q. Wang, Y. Gao, C. Xu, and F. Gao, "Learning-based 3d occupancy prediction for autonomous navigation in occluded environments," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4509–4516, IEEE, 2021.

[27] G. Gigerenzer and P. M. Todd, *Simple heuristics that make us smart*. Oxford University Press, USA, 1999.

[28] G. Gigerenzer and H. Brighton, "Homo heuristicus: Why biased minds make better inferences," *Topics in cognitive science*, vol. 1, no. 1, pp. 107–143, 2009.

[29] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

[30] Z. Chen, D. Duggirala, D. Crandall, L. Jiang, and L. Liu, "Sepaint: Semantic map inpainting via multinomial diffusion," *arXiv preprint arXiv:2303.02737*, 2023.

[31] J. A. Preiss*, W. Hönig*, G. S. Sukhatme, and N. Ayanian, "Crazyswarm: A large nano-quadcopter swarm," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3299–3304, IEEE, 2017. Software available at `https://github.com/USC-ACTLab/crazyswarm`.