QUERY THE AGENT: IMPROVING SAMPLE EFFICIENCY THROUGH EPISTEMIC UNCERTAINTY ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Curricula for goal-conditioned reinforcement learning agents typically rely on poor estimates of the agent's competency and epistemic uncertainty, or fail to consider the agents' epistemic uncertainty altogether, resulting in poor sample efficiency. We propose a novel algorithm, Query The Agent (QTA), which significantly improves sample efficiency by estimating the agent's epistemic uncertainty and setting goals in highly uncertain areas. Encouraging the agent to visit and collect data in highly uncertain states allows the agent to improve its estimation of the value function rapidly. We also propose a novel technique for estimating epistemic uncertainty, Predictive Uncertainty Estimation (PUE), to allow QTA to assess the agent's uncertainty in all previously observed states. We demonstrate that QTA offers decisive sample efficiency improvements over preexisting methods.

1 INTRODUCTION

Deep reinforcement learning has been demonstrated to be highly effective in a diverse array of sequential decision-making tasks (Silver et al., 2016; Berner et al., 2019). However, deep reinforcement learning remains challenging to implement in the real world, in part because of the massive amount of data required for training. This challenge is acute in robotics (Sünderhauf et al., 2018; Dulac-Arnold et al., 2019), in tasks such as manipulation (Liu et al.), and in self-driving cars (Kothari et al.).

Existing curriculum methods for training goal-conditioned reinforcement learning (RL) agents suffer from poor sample efficiency (Dulac-Arnold et al., 2019) and often fail to consider agents' specific competency and epistemic uncertainty when selecting goals. Instead, they rely on poor proxy estimates of epistemic uncertainty or high-level statistics from rollouts, such as success rate. Without customizing learning according to agents' epistemic uncertainties, existing methods suffer from poor sample efficiency, due to three scenarios. Firstly, a given curriculum may not be sufficiently challenging for an agent, thus using timesteps inefficiently. Secondly, a given curriculum may be too challenging to an agent, causing the agent to learn more slowly than it could otherwise. Thirdly, a curriculum may fail to take into account an agent catastrophically forgetting the value manifold in a previously learned region of the state space. Without knowing this, the curriculum will not encourage the agent to re-learn that region of the state space correctly. Only by having a detailed estimate of the agent's epistemic uncertainty throughout the state space and adapting to the agent's needs, can a curriculum maximize an agent's learning. Setting appropriate goals through an adaptive curriculum allows the agent to collect data in poorly learned regions of the state space, allowing the agent to learn more quickly (Kaelbling, 1993; Plappert et al., 2018).

We propose a novel curriculum, Query The Agent (QTA), to accelerate learning in goal-conditioned settings. The underlying mechanism driving QTA's success is a novel technique for estimating the agent's epistemic uncertainty, Predictive Uncertainty Estimation (PUE), and is applied to all previously observed states. The curriculum algorithm drives the agent to reduce its epistemic uncertainty as quickly as possible by forcing it into states with high epistemic uncertainty.

In order to effectively estimate an agent's epistemic uncertainty, PUE utilizes a novel neural architecture. PUE augments a standard DDPG (Lillicrap et al., 2016) agent with three critics, similar to in Lakshminarayanan et al. (2017). Additionally, each critic has a designated *predictive network* that is responsible for predicting the next state the agent will observe given a current state, action, and goal. The predictive networks take as input the critics' final latent representation to predict the agent's next observed state. Predicting directly from the critic's latent representation allows the predictive

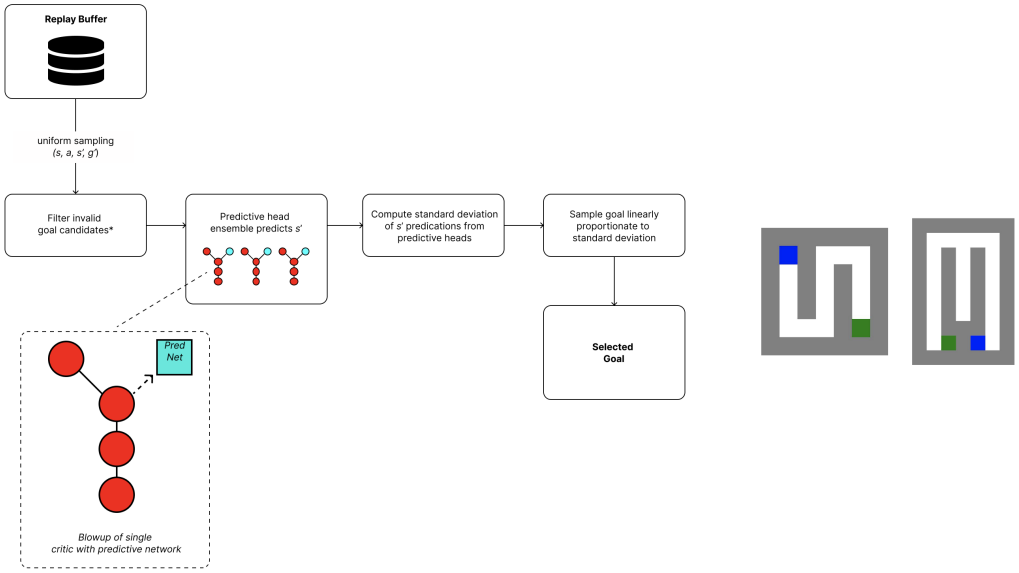

Figure 1: Left: An overview of our goal-selection method. *Threshold filter is only applied when sampling the first goal of an episode. Right: An example of a 1-period square-wave maze, and an M-maze. Starting locations are sampled uniformly in the blue regions and goal locations are sampled uniformly in the green regions. Environment described in detail in section 4

networks to simply extract the predicted next state, which we argue must already be present in the critic's latent representation in order for the agent to make a Q-value prediction. This also ensures that the quality of the predictive networks' predictions will be representative of the quality of the critics' approximation of the optimal value function. The level of disagreement indicates the variance of the quality of the critics' latent representations, thus serving as an effective indicator of the critics' epistemic uncertainty about a given state in the state space.

After estimating the agent's epistemic uncertainty throughout the state space using PUE, QTA's then stochastically samples a goal proportionately to the agent's epistemic uncertainty about all previously observed goal candidates. This encourages the agent to explore the regions that it's most uncertain about and rapidly drive down its uncertainty. QTA may sample the newest states that the agent has visited the least, or in the case of catastrophic forgetting QTA may sample a goal in a highly visited region that has increased in epistemic uncertainty due to an unstable gradient update. The agent discovers new, highly uncertain regions of the state space by frequently taking random actions throughout training. This goal selection procedure is shown in figure 1.

QTA can also be extended to further maximize learning by ensuring that the agent samples the most important transitions when sampling batches for gradient updates. We present a modified version of prioritized experience replay (Schaul et al., 2016) (PER), where the priorities are set according to the estimated epistemic uncertainty measure, as opposed to using the standard temporal difference error (Hasselt, 2010) prioritization scheme. Using epistemic uncertainty ensures that the most informative transitions will be sampled more frequently, and our analysis shows that this helps accelerate learning further. Additionally, to facilitate learning in a sparse reward environment, transitions' goals are probabilistically replaced, as in Andrychowicz et al. (2017).

Our contributions are:

1. Predictive Uncertainty Estimation (PUE): A new technique, broadly applicable to all Q learning approaches, for measuring an agent's epistemic uncertainty pertaining to a particular region of the state space by using the agent's own latent representation. We demonstrate an implementation with DDPG.

2. Query the Agent (QTA): A novel algorithm that adapts to the agent's epistemic uncertainty over the distribution of observed states, and pushes an agent to collect and rapidly integrate maximally informative data from the environment and drive down the agent's epistemic uncertainty.
3. An analysis, including ablation experiments, on how PUE estimates epistemic uncertainty throughout state space and how QTA evolves an agent's understanding of its environment over time.

2 Related Work

Recent advances build upon Hindsight Experience Replay (HER) (Andrychowicz et al., 2017) by designing curricula that allow the agent to cleverly select goals to accelerate learning. None of these appropriately take into account the agent's epistemic uncertainty. We will categorize methods according to which information about the agent they utilize to select goals, then discuss them.

Pitis et al. (2020) does not directly take into account any information regarding the agent's epistemic uncertainty or performance. Their method simply fits a kernel density estimate over the set of previously visited goals, then draws samples as goal candidates and selects the sample with the lowest likelihood under the kernel density estimate. This method benefits from reliably producing increasingly difficult goals for the agent and being immune to problems such as a goal generator network destabilizing. There are multiple drawbacks. Firstly, if an agent suffers from catastrophic forgetting, the curriculum won't be able to correct the problem. Secondly, if the incremental goals are insufficiently/overly challenging for the agent, the curriculum won't properly adapt. Pitis et al. (2020) shares some commonalities with pseudocount techniques, setting goals exclusively based on visitation frequency.

Similarly, count-based and pseudocount-based algorithms (Tang et al.; Bellemare et al.) rely on counting the number of times that a state or a region of state space has been visited. In high dimensional or continuous environments, these algorithms rely on various compression or learned similarity metrics (Machado et al.; Schmidhuber, 2008) or basic statistics in order to identify which regions of the state space are considered similar and adjacent to one another. Intrinsic exploration bonuses can be awarded and goals can be set based on visitation frequencies to accelerate exploration. These methods, taking into account exclusively visitation frequency, do not directly to into account the agent's epistemic uncertainty or performance.

Other techniques, such as Bharadhwaj et al. (2020); Portelas et al. (2019); Florensa et al. (2018), only indirectly take into account the agent's understanding of the state space through approximations based on statistics regarding agents' historical ability to reach goals. Since such metrics offer minimal insight into the agent's understanding of the task, these methods select goals that may not be maximally informative to the agent. The major weakness of these methods is that the goal-generation networks can destabilize, and are required to set goals with only very abstract data about the agent's understanding of the state space.

Another class of algorithms, such as Pathak et al.; Zhang et al. (2020), rely on an external ensemble of networks and estimate epistemic uncertainty throughout the state space. Pathak et al. maintains an external network of forward models, which observe the same data as the agent, and uses the disagreement among the forward models as an epistemic uncertainty estimate. Zhang et al. (2020) does the same, but with three external critics. The major problem with these algorithms is that they are estimating the epistemic uncertainty of *other* networks, not the agent's network. Because these external networks' understanding of the state space will drift away from the agent's over time, the epistemic uncertainty estimate of the external networks is not representative of the agent's epistemic uncertainty estimate. Additionally, Zhang et al. (2020) requires privileged access to the entire state space in order to stabilize.

Other curiosity-related methods, such as random network distillation (Burda et al.), use different formulations of a similar mechanism; estimating the uncertainty of networks disconnected from the agent's computation graph. Random network distillation relies on using the agent's experiences to train an external network to predict the output of a frozen, randomly initialized network known as the target network. As an agent visits a particular state more, the external network improves its prediction of the output of the target network, making it an effective exploration technique to provide

intrinsic reward bonuses proportionate to the size of the external network's prediction error. However, this approach relies on estimating the agent's error in the random network prediction task to gauge uncertainty in the Q-learning task. These are two very different tasks with different difficulties and will learn at different speeds, thus this technique provides a poor estimate of epistemic uncertainty.

A number of techniques (Chane-Sane et al., 2021; Nachum et al., 2018; Christen et al., 2021; OpenAI et al., 2021) have used multiple reinforcement learning agents in order to solve long-horizon tasks, by either using the agents in an adversarial fashion or in a hierarchy. Using multiple reinforcement learning agents incurs a significant computation expense, and comes along with significant technical challenges, such as nonstationarity between agents. Consideration of these techniques is outside the scope of this work, as this work focuses on maximizing the learning of a single agent.

3 METHODS

When learning to navigate a goal-conditioned reinforcement learning environment, we seek to collect as much useful information as possible with each rollout, based on which areas of the state space the agent is least certain about. We measure the agent's uncertainty using PUE and sample a goal in a highly uncertain region of the state space in order to gather transitions that will be maximally informative to the agent when performing gradient updates. We ensure that the most important transitions are sampled for gradient updates through a modified version of prioritized experience replay, in which the priorities are determined by the level of disagreement among the predictive heads. In this section, we first introduce the critic ensemble and predictive networks, which are essential for facilitating epistemic uncertainty estimation. Next, we introduce QTA's mechanism for using epistemic uncertainty estimates to select a goal. Lastly, we introduce our customized replay buffer based on prioritized experience replay, which allows for further sample efficiency improvements.

3.1 UNCERTAINTY ESTIMATION

Critic Ensemble We utilize a critic ensemble to enable effective epistemic uncertainty estimation. QTA utilizes G critics $\phi_{i\in\{1,...,G\}}$, G target critics $\phi_{targ,i\in\{1,...,G\}}$, and a standard policy π_θ. In our QTA implementation we fix $G = 3$. Having multiple critics providing estimates is essential for assessing the epistemic uncertainty of the network. In order to ensure the stability of an ensemble of critics, we train them with a joint loss function. Given a mini-batch $B = \{(s, a, r, s', g, g')\}$, similar to Chen et al. (2021) we sample a set of two indices M from $\{1, ..., G\}$ then compute the target:

$$y_t = r_t + \gamma(\min_{i\in M} Q_{\phi_{targ,i}}(s', a', g')), \quad a' \sim \pi_\theta(\cdot|s')$$

Then update ϕ_i with gradient descent using

$$\nabla_\phi \frac{1}{|B|} \sum_{(s,a,r,s',g,g')\in B} (Q_{\phi_i}(s, a, g) - y)^2$$

We update target networks using polyak averaging coefficient ρ:

$$\phi_{targ,i} \leftarrow \rho\phi_{targ,i} + (1 - \rho)\phi_i$$

Lastly, the policy is updated as in DDPG.

Predictive Network We assign a predictive network P_i to each critic network, where each predictive network is a single feed-forward layer that takes as input the final latent state of the critic and predicts the next hidden state the agent will observe. More specifically $Q_{backbone,i}(s, a, g)$ yields a final hidden state ξ_i, and the final layer of Q_i maps $Q_{i,final}(\xi_i) \rightarrow Q_i(s, a, g)$ while the predictive network maps $P_i(\xi_i) \rightarrow \tilde{s}'$. We train the predictive network by minimizing the objective:

$$L = \frac{1}{|B|} \sum_{\{(s,a,r,s',g,g')\}} (s' - \tilde{s}')^2$$

The critic's epistemic uncertainty regarding a particular state is simply the standard deviation of the three predictive heads' predictions. Note that we intentionally use the same ξ_i as input to both the final head of the critic and as input to the predictive network so that the quality of the two predictions will be tightly coupled, which allows the disagreement among the predictive networks to be highly indicative of the critic's epistemic uncertainty. Further, we intentionally use an identical architecture for the predictive network as with the critic network to ensure they have the same learning capacity and remain tightly coupled.

3.2 GOAL SELECTION

QTA samples previously seen goals from the replay buffer to serve as goal candidates. In order to get an unbiased estimate of epistemic uncertainty ϵ for a particular state s, we uniformly sample d random actions from the action space and compute the epistemic uncertainty of a state to be the mean of d uncertainty estimates for that state. More precisely:

$$\epsilon_k = \frac{1}{d}\sum_{j=0}^{d}\frac{1}{G}\sum_{i=0}^{G}P_i(Q_{backbone,i}(s, a_j, g_{desired}))$$

We use the desired goal sampled from the environment for these estimates, though the prediction is learned to be agnostic to the choice of goal. The distribution of estimated epistemic uncertainties is normalized between 0 and 1 before proceeding to goal selection. Each goal is sampled with a probability linearly proportionate to its estimated epistemic uncertainty. This very simple approach encourages QTA to sample goals with high epistemic uncertainty, allowing the agent to gather maximally informative transitions in the regions of the state space where it is least certain. Given a slope m and y-intercept b, a goal candidate with epistemic uncertainty ϵ_k is selected as the agent's next goal with probability:

$$\frac{max(m\epsilon_k + b, 0)}{\sum_{k=0}^{N} max(m\epsilon_k * b, 0)}$$

Whenever the agent successfully reaches its goal, a new goal is sampled using this same procedure. Estimating the agent's epistemic uncertainty only requires a forward pass through the critic backbone and predictive head, so QTA is able to efficiently consider an arbitrarily large number of goal candidates by utilizing batch operations.

3.3 PRIORITIZED EXPERIENCE REPLAY

We implement a modified version of prioritized experience replay (PER) in order to ensure that our agent samples the transitions with greatest epistemic uncertainty when performing gradient updates. We make two key modifications. Firstly, rather than using temporal difference error to establish priorities in the replay buffer, we utilize the estimated epistemic uncertainty according to PUE to provide a more accurate measure of which transitions are most beneficial to sample. Secondly, once a transition is sampled its goal is probabilistically replaced with the goal of a transition later in its trajectory and its reward is recomputed, as in the standard hindsight experience replay (HER) (Andrychowicz et al., 2017) mechanism. That is to say, that in the sparse reward setting we use PER to sample from a HER replay buffer instead of a standard replay buffer. This maintains the benefits of the HER replay buffer, while allowing us to sample the most important transitions through the PER mechanism. Similarly to in Schaul et al. (2016), we sample a transition with priority p_j with probability:

$$\frac{p_j^\alpha}{\sum_j p_j^\alpha}$$

Schaul et al. (2016) demonstrated that oversampling high-priority transitions introduces a bias that can destabilize gradient updates. In order to anneal the bias created by oversampling transitions with high epistemic uncertainty, we multiply by importance sampling weights:

$$w_j = (\frac{1}{N}\frac{1}{p_j})^\beta$$

Where β is initialized to a set value and gradually increased throughout training until $\beta = 1$, at which point the bias is fully compensated for.

Finally, it is important that the goals selected by QTA are achievable. To promote achievability, QTA eliminates invalid goal candidates from being selected as the first goal in a rollout through a simple threshold mechanism. Any goal candidates with a mean Q value larger than cT are eliminated, where T is the maximum length of the rollout. This heuristic is not core to the algorithm, and could be replaced with, e.g., a generative model designed to generate achievable goals (Florensa et al., 2018) or a discrminative model to quantify the achievability of goals.

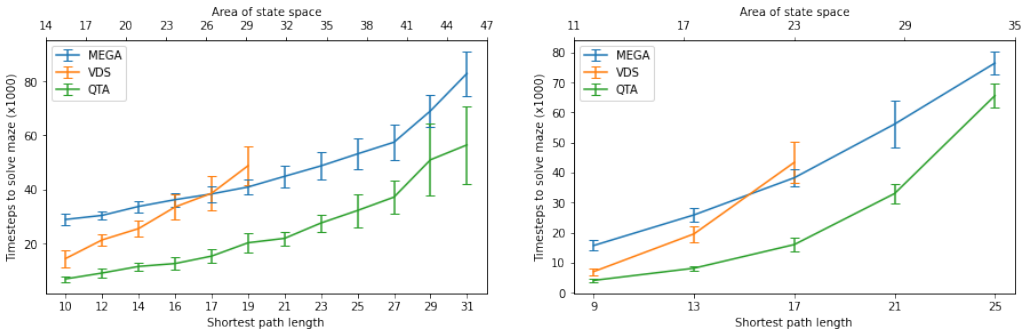

Figure 2: Left: M-maze performance. Right: Square-wave performance. Above we plot the number of timesteps required for various algorithms to attain evaluation performance of 1.0 across 12 seeds on various environments. Performance omitted for when seeds failed to converge for a particular environment. Lower on the graph represents better sample efficiency. Confidence intervals represent one standard deviation.

4 Results and Analysis

Environments We utilize custom a suite of procedurally generated 2D mazes for a dot agent with continuous action and state spaces. In all environments the agent receives a reward of -1 unless it reaches the goal, in which case it receives a reward of 0. These mazes are carefully designed to have a minimum solving time T such that an optimal agent cannot solve the task in fewer than T timesteps. This allows us to evaluate the performance of QTA compared to baseline methods in a series of increasingly challenging environments, where each environment is quantifiably more difficult than the last. Our first set of environments is a set of square-wave mazes, in which the agent must traverse a long path with a single route forward. The square-wave mazes allow us to directly evaluate the most critical skills pertaining to sample efficiency: how quickly an agent is able to explore new regions of the state space and integrate information, while reintegrating information about any regions of the state space that are catastrophically forgotten. Our second set of environments is an M-maze, which features a large, useless dead end that takes up a large fraction of the state space and is designed to waste that agent's time. While evaluating many of the same challenges, it additionally allows us to evaluate agents' sample efficiency when faced with an artificially enlarged state space.

Baseline algorithms We compare QTA's performance with two competitive curriculum baselines: MEGA (Pitis et al., 2020), VDS (Zhang et al., 2020). HER was not able to reach the goal in any of our environments within one million timesteps. We used the authors' code and hyperparameters for MEGA and VDS, and tried to tune the hyperparameters to our environment. We were unable to imprive performance further, likely because both algorithms were originally tuned for very similar 2D navigation environments.

Main results By accurately estimating the epistemic uncertainty of the agent, QTA is able to navigate mazes using significantly fewer timesteps faster than MEGA, as shown in figure 2. MEGA, which ignores the epistemic uncertainty of the model and largely relies on high-level statistics about visited states, did not induce learning as quickly as QTA's methodology of taking into account the epistemic uncertainty. Interestingly, we see that as the time horizon continues to expand to very long horizons and state spaces, MEGA begins to approach the performance of QTA, which may be due to the predictive network overfitting when predicting the agent's epistemic uncertainties.

We find that VDS is competitive at first, but quickly becomes less sample efficient and then fails to solve many environments. We suspect that it's only competitive at the beginning because VDS has access to privileged information about the entire state space, including states not yet observed.

We also observe that both algorithms perform much worse and scale more poorly on square-wave-shaped environments, despite having much shorter mazes to solve with much smaller state spaces. The greater number of sharp turns required to solve the square-wave mazes hurts sample efficiency, as the optimal value function is harder to approximate.

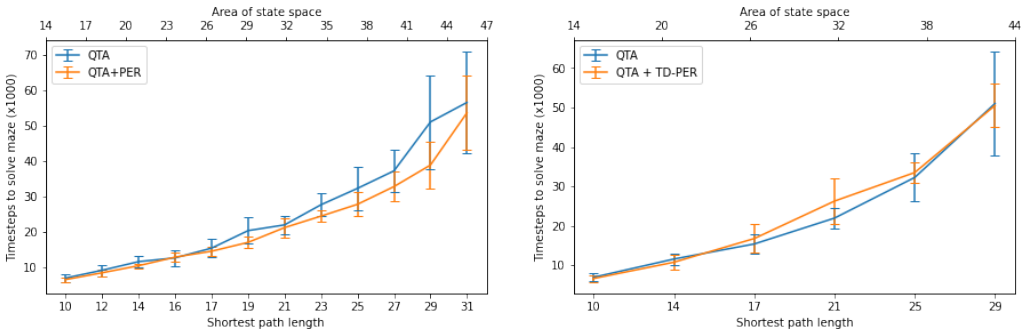

Figure 3: Left: QTA with and without PER. Right: QTA vs standard PER. We plot the number of timesteps needed to attain evaluation performance of 1.0 across 12 seeds in M-mazes of various lengths.

We further demonstrate the importance of estimating the agent's epistemic uncertainty in figure 3 by showing that it can be used as a signal to prioritize certain replay transitions, as shown in figure 3. In the left figure, we show that introducing PER offers some performance benefits with longer mazes before the predictive network begins to overfit. In the right figure. We demonstrate that standard PER with temporal-difference error for prioritization offers nearly no benefit to performance. Prioritizing based on temporal difference was not helpful to the algorithm, but prioritizing with epistemic uncertainty benefited the algorithm in navigating longer mazes.

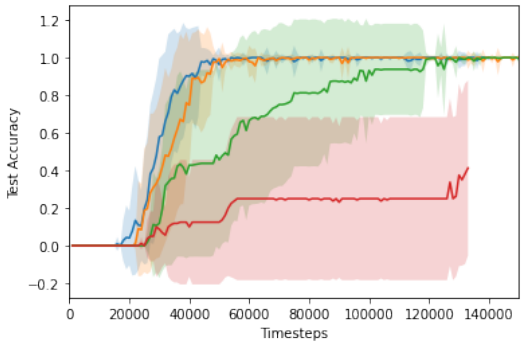

Figure 4: We add gaussian noise to PUE's uncertainty estimates. Blue: no noise. Orange: standard deviation=0.01. Green: standard deviation=0.03. Red: standard deviation=0.05

Lastly, we will verify that QTA is forcing the agent to minimize uncertainty and not simply choosing goals that are incrementally further away or following any other strategy. To demonstrate this, as shown in figure 4, we add small amounts of gaussian noise to PUE's uncertainty estimates and find that performance suffers significantly and ultimately collapses with even a small amount of added noise. We demonstrate that the performance of QTA hinges on following highly precise estimates of epistemic uncertainty.

Analysis of critic networks In figure 5 we visualize the critics of one of our agents learning over time in an M-maze. The agent starts in the bottom-left inlet, as shown in figure 1. We demonstrate that our neural networks are performing as we would expect. Firstly, upon the analysis of the value function, we see that the map generally gets lighter over time, indicating that the estimated Q values increase over time as the agent's competency increases. Naturally, some areas of the value manifold are better learned than others.

Comparing this learned value function with the optimal value function yields the absolute error map shown in the second column. We show that in general the error starts high and uniform, and decreases over time. As expected, the error is greater in states further away from the starting point, as those states have been visited less frequently and been sampled fewer times in gradient updates than those closer to the start.

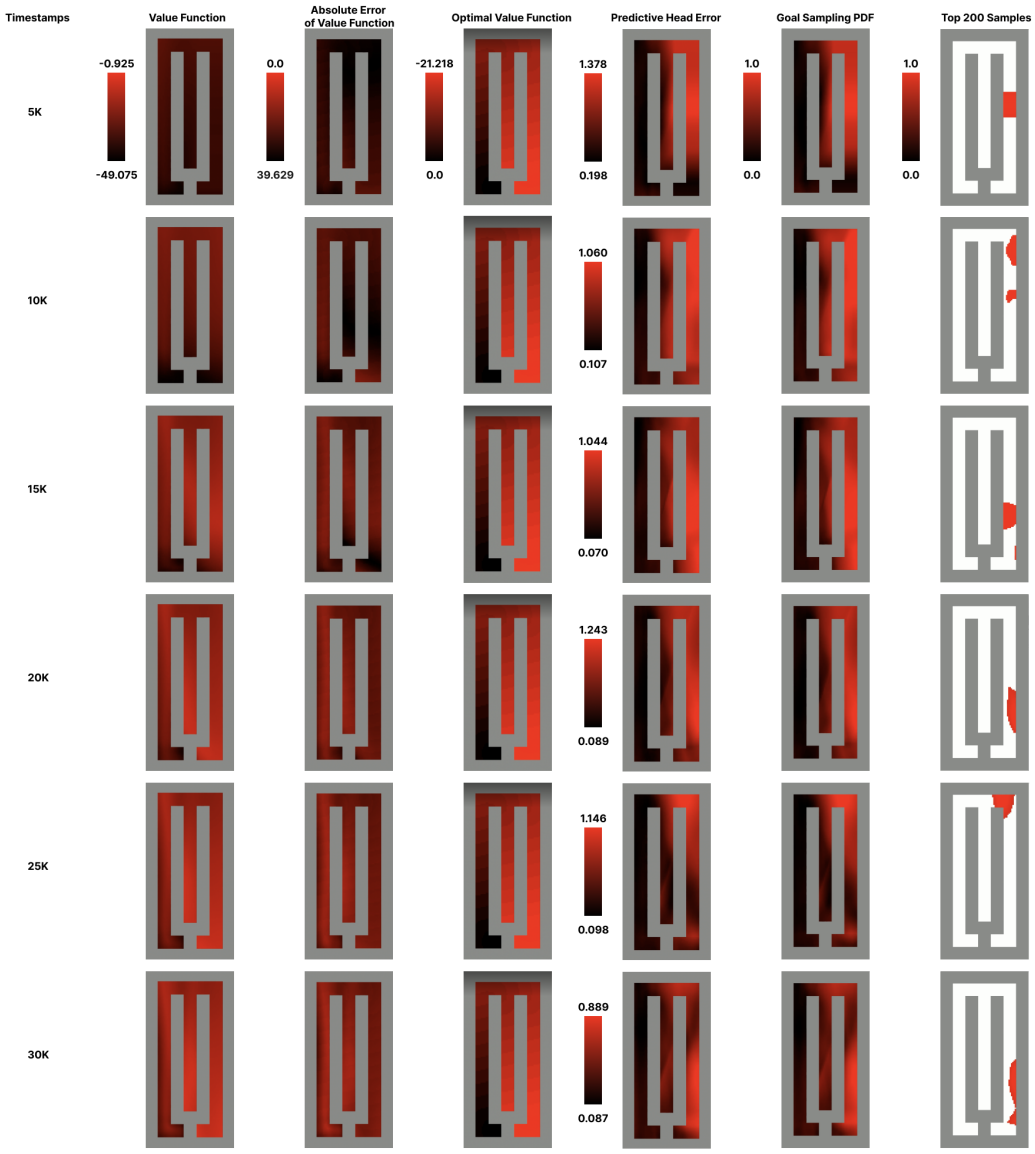

Figure 5: Analysis of the ensemble of critics for one seed on a 12x5-unit M-maze. From left to right, in each column we show: 1) the value function for all states averaged across all critics 2) the absolute value of the error between the average value function and the optimal value function, 3) the optimal value function 4) the average error among the three predictive heads 5) the probability of a state being sampled as the agent's next goal 6) 200 sampled states based on epistemic uncertainty. We emphasize to the reader that across time all color scales are the same across time except for the predictive head error, which changes over time to make the contrast more visible.

By analyzing the estimated epistemic uncertainty in the fourth column, we gain insight into how the agent selects goals. The fourth column contains a spatial representation of the agent's epistemic uncertainty where brighter colors indicate greater uncertainty. We simultaneously discuss the fifth column, showing the likelihood of any particular point being sampled as a goal for the agent. We see from the average predictive head errors that after only five-thousand iterations, the network already understands and has explored a large region of the maze. We can see that the network least understands the furthest regions of the maze. This is demonstrated by the network sampling goals only in the regions of the maze furthest from the start. As time progresses, the network overall becomes more certain and the magnitude of the average forward error decreases significantly but is

punctuated by occasional catastrophic forgetting events. We point out to the reader that the visual scale of the fourth column is unique to each image, as indicated to the left of the maze graphic, in order to make the contrast more visible to the reader. We see that the goal sampling likelihoods in the fifth column track very closely with the predictive head errors in the fourth column, demonstrating that QTA samples goals in the most epistemically uncertain locations for the agent.

In the final column we have a representative sampling of 200 goals, and see that they tend to be targeted toward areas of the maze that are far away from the origin, tracking very closely with the goal sampling likelihoods.

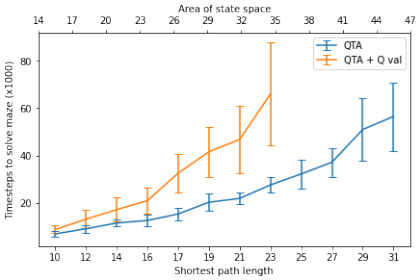

Figure 6: Performance of QTA using PUE uncertainty, vs using disagreement among Q values for PUE on various M-mazes.

Q value disagreement PUE utilizes the disagreement among predictive networks' outputs to estimate epistemic uncertainty. A more intuitive choice might be to use the disagreement among Q values. We experimented with using the disagreement among Q values rather than the predicted next states as in 6, but found it to be an inferior estimator of epistemic uncertainty, leading to dramatically worse performance. By analyzing the sources of noise, we see that we would anticipate the next state prediction to be less noisy and thus more stable than the prediction for the Q value. By the definition of the Q value $Q(s) = r + \gamma Q(s')$, there are three sources of noise δ in the estimation of the Q value: δ_{reward}, $\delta_{s'}$, and $\delta_{\gamma Q(s')}$. When predicting s', the predictive network's output has only one source of error: $\delta_{s'}$. Naturally, the signal-to-noise ratio will be much higher with the predictive heads' output. A very low signal-to-noise ratio is consistent with the analysis in Zhang et al. (2020). Further, we intuitively anticipate $\tilde{s}'$ to be a good predictor of the quality of the critic's latent representation since the accuracy $\tilde{s}'$ is essential for accurately predicting $Q(s)$.

REFERENCES

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, 2017.

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Christopher Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique Pondé de Oliveira Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. Dota 2 with large scale deep reinforcement learning. *CoRR*, abs/1912.06680, 2019.

Homanga Bharadhwaj, Animesh Garg, and Florian Shkurti. Leaf: Latent exploration along the frontier. *CoRR*, abs/2005.10934, 2020.

Yuri Burda, Harrison Edwards, Amos J. Storkey, and Oleg Klimov. Exploration by random network distillation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Elliot Chane-Sane, Cordelia Schmid, and Ivan Laptev. Goal-conditioned Reinforcement Learning with Imagined Subgoals. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Xinyue Chen, Che Wang, Zijian Zhou, and Keith W. Ross. Randomized ensembled double q-learning: Learning fast without a model. In *ICLR*, 2021.

Sammy Christen, Lukás Jendele, Emre Aksan, and Otmar Hilliges. Learning functionally decomposed hierarchies for continuous control tasks with path planning. *IEEE Robotics Autom. Lett.*, 6(2): 3623–3630, 2021.

Gabriel Dulac-Arnold, Daniel J. Mankowitz, and Todd Hester. Challenges of real-world reinforcement learning. *CoRR*, abs/1904.12901, 2019.

Carlos Florensa, David Held, Xinyang Geng, and Pieter Abbeel. Automatic goal generation for reinforcement learning agents. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1515–1528, 2018.

Hado Hasselt. Double q-learning. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

Leslie Pack Kaelbling. Learning to achieve goals. In *Proc. of IJCAI-93*, pp. 1094–1098, 1993.

Parth Kothari, Christian Perone, Luca Bergamini, Alexandre Alahi, and Peter Ondruska. Drivergym: Democratising reinforcement learning for autonomous driving.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf`.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *ICLR*, 2016.

Rongrong Liu, Florent Nageotte, Philippe Zanne, Michel de Mathelin, and Birgitta Dresp-Langley. Deep reinforcement learning for the control of robotic manipulation: A focussed mini-review.

Marlos C. Machado, Marc G. Bellemare, and Michael Bowling. Count-based exploration with the successor representation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*.

Ofir Nachum, Shixiang (Shane) Gu, Honglak Lee, and Sergey Levine. Data-efficient hierarchical reinforcement learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

OpenAI OpenAI, Matthias Plappert, Raul Sampedro, Tao Xu, Ilge Akkaya, Vineet Kosaraju, Peter Welinder, Ruben D'Sa, Arthur Petron, Henrique Ponde de Oliveira Pinto, Alex Paino, Hyeonwoo Noh, Lilian Weng, Qiming Yuan, Casey Chu, and Wojciech Zaremba. Asymmetric self-play for automatic goal discovery in robotic manipulation, 2021.

Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97, pp. 5062–5071. PMLR.

Silviu Pitis, Harris Chan, Stephen Zhao, Bradly Stadie, and Jimmy Ba. Maximum Entropy Gain Exploration for Long Horizon Multi-goal Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Matthias Plappert, Marcin Andrychowicz, Alex Ray, Bob McGrew, Bowen Baker, Glenn Powell, Jonas Schneider, Josh Tobin, Maciek Chociej, Peter Welinder, Vikash Kumar, and Wojciech Zaremba. Multi-goal reinforcement learning: Challenging robotics environments and request for research. *CoRR*, abs/1802.09464, 2018.

Rémy Portelas, Cédric Colas, Katja Hofmann, and Pierre-Yves Oudeyer. Teacher algorithms for curriculum learning of deep RL in continuously parameterized environments. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura (eds.), *3rd Annual Conference on Robot Learning, CoRL 2019, Osaka, Japan, October 30 - November 1, 2019, Proceedings*, volume 100, pp. 835–853, 2019.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

Jürgen Schmidhuber. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. *CoRR*, abs/0812.4360, 2008. URL `http://arxiv.org/abs/0812.4360`.

David Silver, Aja Huang, Christopher J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–503, 2016. URL `http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html`.

Niko Sünderhauf, Oliver Brock, Walter J. Scheirer, Raia Hadsell, Dieter Fox, Jürgen Leitner, Ben Upcroft, Pieter Abbeel, Wolfram Burgard, Michael Milford, and Peter Corke. The limits and potentials of deep learning for robotics. *CoRR*, abs/1804.06557, 2018.

Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. #exploration: A study of count-based exploration for deep reinforcement learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*.

Yunzhi Zhang, Pieter Abbeel, and Lerrel Pinto. Automatic Curriculum Learning through Value Disagreement. In *Advances in Neural Information Processing Systems*, volume 33, 2020.