Uncertainty Prioritized Experience Replay

Rodrigo Carrasco-Davis, Sebastian Lee, Claudia Clopath, Will Dabney

Keywords: Experience Replay, Uncertainty Estimation, Information Gain

Summary

Prioritized experience replay, which improves sample efficiency by selecting relevant transitions to update parameter estimates, is a crucial component of contemporary value-based deep reinforcement learning models. Typically, transitions are prioritized based on their temporal difference error. However, this approach is prone to favoring noisy transitions, even when the value estimation closely approximates the target mean. This phenomenon resembles the *noisy TV* problem postulated in the exploration literature, in which exploration-guided agents get stuck by mistaking noise for novelty. To mitigate the disruptive effects of noise in value estimation, we propose using epistemic uncertainty to guide the prioritization of transitions from the replay buffer. Epistemic uncertainty quantifies the uncertainty that can be reduced by learning, hence reducing transitions sampled from the buffer generated by unpredictable random processes. We first illustrate the benefits of epistemic uncertainty prioritized replay in two tabular toy models: a simple multi-arm bandit task, and a noisy gridworld. Subsequently, we evaluate our prioritization scheme on the Atari suite, outperforming quantile regression deep Q-learning benchmarks; thus forging a path for the use of epistemic uncertainty prioritized replay in reinforcement learning agents.

Contribution(s)

1. We introduce a new decomposition of uncertainties in reinforcement learning extending previous formulations of epistemic and aleatoric uncertainty estimators (Clements et al., 2020) to include a distance-to-target term. This decomposition better accounts for bias-variance trade-offs in the underlying estimator.
   Context: While Clements et al. (2020) start by defining total uncertainty estimator as the variance over distributional and ensemble dimensions of the value estimate, we start instead from the average square error to the target over distributional and ensemble dimensions. Under the definitions given by Lahlou et al. (2022) in their Direct Epistemic Uncertainty Prediction (DEUP) framework, this yields a modified epistemic uncertainty estimator that we term the *target epistemic uncertainty*.
2. We propose using these measures of epistemic and aleatoric uncertainty in an *information gain* criterion to prioritize experience replay in reinforcement learning. We call this prioritization scheme Uncertainty Prioritized Experience Replay (UPER).
   Context: The de facto method for prioritizing replay in reinforcement learning has been the absolute value of the temporal difference error since its introduction by Schaul et al. (2016). However we argue that this can lead to sub-optimal behavior in noisy environments. We go on to derive the information gain prioritization criterion from principled treatment of a toy Bayesian problem.
3. We demonstrate the effectiveness of this prioritization scheme in two toy models (a bandit and gridworld), as well as in a deep learning model on the Atari test suite. In the latter we use an ensemble of distributional QR agents (Dabney et al., 2017) to estimate the relevant uncertainty quantities.
   Context: We provide a series of ablation studies in Atari that isolate the effect of the prioritization variable (from architectural changes such as adding an ensemble), showing that UPER could be a promising alternative to PER and other uncertainty measures like plain ensemble disagreement.

Uncertainty Prioritized Experience Replay

Rodrigo Carrasco-Davis[1,†], **Sebastian Lee**[1,2,4,†], **Claudia Clopath**[3,4], **Will Dabney**[5]
rodrigo.cd.20@ucl.ac.uk,
sebastianlee@flatironinstitute.org

[1]**Gatsby Computational Neuroscience Unit, University College London**
[2]**Center for Computational Neuroscience, Flatiron Institute**
[3]**Sainsbury Wellcome Centre for Neural Circuits and Behaviour, University College London**
[4]**Bioengineering Department, Imperial College London**
[5]**Google DeepMind**

[†] Equal contribution

Abstract

Prioritized experience replay, which improves sample efficiency by selecting relevant transitions to update parameter estimates, is a crucial component of contemporary value-based deep reinforcement learning models. Typically, transitions are prioritized based on their temporal difference error. However, this approach is prone to favoring noisy transitions, even when the value estimation closely approximates the target mean. This phenomenon resembles the *noisy TV* problem postulated in the exploration literature, in which exploration-guided agents get stuck by mistaking noise for novelty. To mitigate the disruptive effects of noise in value estimation, we propose using epistemic uncertainty estimation to guide the prioritization of transitions from the replay buffer. Epistemic uncertainty quantifies the uncertainty that can be reduced by learning, hence reducing transitions sampled from the buffer generated by unpredictable random processes. We first illustrate the benefits of epistemic uncertainty prioritized replay in two tabular toy models: a simple multi-arm bandit task, and a noisy gridworld. Subsequently, we evaluate our prioritization scheme on the Atari suite, outperforming quantile regression deep Q-learning benchmarks; thus forging a path for the use of uncertainty prioritized replay in reinforcement learning agents.

1 Introduction

Deep Reinforcement Learning (DRL) has proven highly effective across a diverse array of problems, consistently yielding state-of-the-art results in control of dynamical systems (Nian et al., 2020; Degrave et al., 2022; Weinberg et al., 2023), abstract strategy games (Mnih et al., 2015; Silver et al., 2016), continual learning (Khetarpal et al., 2022; Team et al., 2021), and multi-agent learning (OpenAI et al., 2019; Baker et al., 2020). It has also been established as a foundational theory for explaining phenomena in cognitive neuroscience (Botvinick et al., 2020; Subramanian et al., 2022). Nonetheless, a significant drawback of these methods pertains to their inherent *sample inefficiency* whereby accurate estimations of value and policy necessitate a substantial demand for interactions with the environment.

Sample inefficiency has been mitigated through the use of, among other methods, Prioritized Experience Replay (PER) (Schaul et al., 2016). PER is an extension of Experience Replay (Lin, 1992), which uses a memory buffer populated with past agent transitions to improve training stability through the temporal de-correlation of data used in parameter updates. Subsequently, PER extends

this approach by sampling transitions from the buffer with probabilities proportional to their absolute Temporal Difference (TD) error, thereby allowing agents to *prioritize* learning from pertinent data. PER has been widely adopted as a standard technique in DRL; however, despite significantly better performance over uniform sampling in most cases, it is worth noting that PER can encounter limitations under specific task conditions and agent designs. The most prominent example of such a limitation is related to the so-called *noisy TV* problem (Burda et al., 2018), a thought experiment at the heart of the literature around exploration in RL. Just as novelty-based exploration bonuses can trap agents in noisy states, PER is susceptible to frequently replaying transitions involving high levels of randomness (e.g. in reward or transition dynamics) even if they do not translate to meaningful learning and thus are not useful for solving the task.

To combat this issue, we propose combining epistemic and aleatoric uncertainty estimators (Clements et al., 2020; Alverio et al., 2022; Lahlou et al., 2022; Liu et al., 2023; Jiang et al., 2023), originally used to promote exploration, under an information gain criterion for use in replay prioritization. Epistemic uncertainty, the uncertainty reducible through learning, is the key quantity of interest. However this need to be appropriately 'calibrated', which we show-both empirically, and with justification from Bayesian inference-can be done effectively by dividing the epistemic uncertainty estimate by an aleatoric uncertainty estimate (and taking the logarithm, i.e. the information gain). Intuitively the need for this kind of calibration can be seen by considering the following game: the aim is to estimate the mean of two distributions; the ground truth is that both distributions have identical mean but different variance, and your current estimates for both distributions are the same i.e. your epistemic uncertainty on the mean is the same for both distributions. However if I offer you a new sample from either distribution to refine your estimate you would choose to sample the distribution with lower variance since this is more likely to be informative. In addition to arguing for this novel prioritization variable, we also provide candidate methods involving distributions of ensembles (in the vein of Clements et al. (2020)) to estimate these quantities. A comprehensive review of related literature is provided in App. B, with further details in SM 1.

Our primary contributions are as follows: (1) In Sec. 3, we present a novel approach for estimating epistemic uncertainty, building upon an existing uncertainty formalization introduced by Clements et al. (2020) & Jiang et al. (2023). This extension incorporates information about the target value that the model aims to estimate thereby accounting for bias in the estimator; (2) We derive a prioritization variable using estimated uncertainty quantities, finding a specific functional form derived from a concept called *information gain*, showing that both, epistemic and aleatoric uncertainty should be considered for prioritization; (3) In Sec. 4, we illustrate the advantages of our proposed epistemic uncertainty prioritization scheme through two interpretable toy models, a bandit task and a grid world; (4) In Sec. 5, we demonstrate the effectiveness of this method on the Atari-57 benchmark (Bellemare et al., 2013), where it significantly outperforms baseline models based on a combination of PER, QR-DQN and ensemble agents.

2 Background

2.1 Reinforcement Learning

Consider an environment modelled by a Markov Decision Process (MDP), defined by (S, A, R, P, γ) with state space S, action space A, reward function R, state-transition function P, and discount factor $\gamma \in (0, 1)$. Given the agent policy $\pi : S \to \Delta(A)$, where $\Delta(A)$ denotes the probability simplex over A, the cumulative discounted future reward is denoted by $G^\pi(s, a) = \sum_t \gamma^t R(s_t, a_t)$ with $s_0 = s$ and $a_0 = a$, and transitions sampled according to $a_t \sim \pi(a|s_t)$ and $s_{t+1}, r_t \sim P(s, r|s_t, a_t)$. We denote the action-value function as $Q^\pi(s, a) = \mathbb{E}\left[G^\pi(s, a)\right]$, and the corresponding state-action return-distribution function as $\eta^\pi(s, a)$; and we recall that $Q^\pi(s, a) = \mathbb{E}_{G\sim\eta^\pi(s,a)}\left[G\right]$. In general, the action value function is parameterized by ψ, such that Q_ψ can be trained by minimizing a mean-squared temporal difference (TD) error $\mathbb{E}[\delta_t^2]$. For example, in Q-Learning the error is given

by

$$\delta_t = r_t + \gamma \max_{a' \in A} Q_{\bar{\psi}}(s_{t+1}, a') - Q_{\psi}(s_t, a_t), \tag{1}$$

for the *transition* at time t (s_t, a_t, r_t, s_{t+1}), and where $\bar{\psi}$ denotes the possibly time-lagged *target parameters* (Watkins & Dayan, 1992; Mnih et al., 2015). Additionally, we will use policies that are ϵ-greedy with respect to the currently estimated action-value function, that is for some $\epsilon \in [0, 1]$, the selected action from any state s is drawn as $\arg\max_{a \in A} Q_{\psi}(s, a)$ with probability $1 - \epsilon$ and uniformly over A otherwise. See Sutton & Barto (2018) for a more in-depth overview of RL methods.

2.2 Prioritized Experience Replay

Reinforcement learning algorithms are notoriously sample inefficient. A widely adopted practice to mitigate this issue is the use of an experience replay buffer, which stores transitions in the form of (s_t, a_t, r_t, s_{t+1}) for later learning (Mnih et al., 2015). Loosely inspired by hippocampal replay to the cortex in mammalian brains (Foster & Wilson, 2006; McNamara et al., 2014), its primary conceptual motivation is to reduce the variance of gradient-based optimization by temporally decorrelating updates, thereby improving sample efficiency. It can also serve to prevent catastrophic forgetting by maintaining transitions from different time scales. The effectiveness of this buffer can often be improved further by *prioritising* some transitions at the point of sampling rather than selecting uniformly. Formally, when transition i is placed into replay, it is given a priority p_i. The probability of sampling this transition during training is given by:

$$P(i) = \frac{p_i^{\alpha}}{\sum_k p_k^{\alpha}}, \tag{2}$$

where α is a hyper-parameter called *prioritisation exponent* ($\alpha = 0$ corresponds to uniform sampling). Schaul et al. (2016) introduced *prioritized experience replay*, which most often uses the absolute TD-error $|\delta_i|$ of transition i, as $p_i = |\delta_i| + \epsilon$ where a small ϵ constant ensures transitions with zero error still have a chance of being sampled[1]. Sampling transitions non-uniformly from the replay buffer will change the observed distribution of transitions, biasing the solution of value estimates. To correct this bias, the error used for each update is re-weighted by an importance weight of the form $w_i \propto (NP(i))^{-\beta}$, where N is the size of the buffer and β controls the correction of bias introduced by important sampling ($\beta = 1$ corresponds to a full correction).

The key intuition behind PER is that transitions on which the agent previously made inaccurate predictions should be replayed more often than transitions on which the agent already has low error. While this heuristic is reasonable and has enjoyed empirical success, TD-errors can be insufficiently distinct from the irreducible aleatoric uncertainty; considering instead uncertainty measures more explicitly, this form of prioritisation can be significantly improved.

2.3 Uncertainty Estimation in RL

Uncertainty is a fundamental concept in statistics, and a natural way to frame it is through the lens of Bayesian methods (Lahlou et al., 2022; Narimatsu et al., 2023). In the context of reinforcement learning (RL), uncertainty has been extensively studied in relation to exploration. This often involves using proxies to encourage the agent to explore more uncertain states, through methods such as optimistic reward estimates for unexplored states or actions, i.e., upper confidence bounds (Auer, 2002a; Lattimore & Szepesvári, 2020; Antonov et al., 2022), or by assigning intrinsic rewards for visiting novel states (Bellemare et al., 2016b; Lobel et al., 2023). Various characteristics of the RL setting, such as bootstrapping and non-stationarity, make accurate uncertainty estimation particularly challenging. Nevertheless, uncertainty has played an increasingly important role in recent research, including applications in generalization (Jiang et al., 2023), reward bonuses for exploration (Nikolov

[1] Another form of prioritization, known as rank-based prioritisation, is to use $p_i = 1/\text{rank}(i)$ where $\text{rank}(i)$ is the rank of the experience in the buffer when ordered by $|\delta_i|$.

et al., 2019), and guiding safe actions (Lütjens et al., 2019; Kahn et al., 2017). We discuss here some of the key concepts around uncertainty relevant to this work, particularly those that address the delineation between aleatoric and epistemic uncertainty. A more comprehensive overview of related work around uncertainty in RL can be found in SM 1.

2.3.1 Bootstrapped DQN

The concept behind bootstrapping is to approximate a posterior distribution by sampling a prediction from an ensemble of estimators, where each estimator is initialized randomly and observes a distinct subset of the data (Tibshirani, 1994; Bickel & Freedman, 1981). In RL, Osband et al. (2016) introduced a protocol known as *bootstrapped DQN* for deep exploration, whereby bootstrapping is used to approximate the posterior of the action-value function, from which samples can be drawn. Each agent within an effective ensemble, parameterized by ψ, is randomly initialized and trained using a different subset of experiences via random masking. A sample estimate of the posterior distribution, denoted as $\psi \sim P(\psi|D)$ (D being training data), is obtained by randomly selecting one of the agents from the ensemble. In this work, we use extensions of the bootstrapped DQN idea in our epistemic uncertainty measurements—notably the ensemble disagreement.

2.3.2 Distributional RL

Learning quantities beyond the mean return has been a long-standing programme of RL research, with particular focus on the return variance (Sobel, 1982). A yet richer representation of the return is sought by more recent methods known collectively as *distributional RL* (Bellemare et al., 2017), which aims to learn not just the mean and variance, but the entire return distribution. We focus here on one particular class of distributional RL methods: those that model the quantiles of the distribution, specifically QR-DQN (Dabney et al., 2017). A broader treatment of the distributional RL literature can be found in Bellemare et al. (2023).

In QR-DQN, the distribution of returns, for example from taking action a in state s and subsequently following policy π, $\eta^\pi(s,a)$ is approximated as a *quantile representation* (Bellemare et al., 2023), that is, as a uniform mixture of Diracs, and trained through *quantile regression* (Koenker & Hallock, 2001). For such a distribution, $\hat{\nu} = \frac{1}{m}\sum_{i=1}^{m}\delta_{\theta_{\tau_i}}$, with learnable quantile values θ_{τ_i} and corresponding quantile targets $\tau_i = \frac{2i-1}{2m}$, the quantile regression loss for target distribution ν is given by

$$\mathcal{L}_{\text{QR}} = \sum_{i=1}^{m} \mathbb{E}_{Z\sim\nu}[\rho_{\tau_i}(Z - \theta_{\tau_i})], \tag{3}$$

where $\rho_\tau(u) = u(\tau - \mathbb{1}_{u<0})$ and $\mathbb{1}$ is the indicator function. By leveraging the so-called distributional Bellman operator and the standard apparatus of a DQN model, QR-DQN prescribes a temporal difference deep learning method for minimizing the above loss function and learning an approximate return distribution function via quantile regression.

Distributional RL in itself does not (so far) permit a natural decomposition of uncertainties into epistemic and aleatoric (Clements et al., 2020; Chua et al., 2018; Charpentier et al., 2022); rather the variance of the learned distribution will converge on what can reasonably be thought of as the aleatoric uncertainty. In Sec. 3.1 we extend previous techniques that combine distributions with ensembles to construct estimates of both epistemic and aleatoric uncertainties. Both of these techniques to characterize epistemic uncertainty can be understood under an excess risk framework, which we outline below.

2.3.3 Direct Epistemic Uncertainty Prediction

We employ a clear and formal representation of uncertainty, where total uncertainty is defined as the sum of epistemic and aleatoric components such that the epistemic uncertainty can be interpreted as the excess risk. This notion was introduced by Xu & Raginsky (2022) and later extended by Lahlou et al. (2022); we adapt their framing to our setting here. Consider the **total uncertainty** $\mathcal{U}(s,a)$ of

an action-value predictor $Q_\psi(s,a)$, for a given state s and action a as:

$$\mathcal{U}(Q_\psi, s, a) = \int \left(\Theta(s', r) - Q_\psi(s,a)\right)^2 P(s', r|s, a) ds' dr, \tag{4}$$

where $\Theta(s', r)$ is the Q-learning target as in equation Eq. 1. Then, the **aleatoric uncertainty** $\mathcal{A}(s,a)$, is given by the total uncertainty (as defined above) of a Bayes-optimal predictor Q^*_ψ (see Lahlou et al. (2022)):

$$\mathcal{A}(s,a) = \mathcal{U}(Q^*_\psi, s, a). \tag{5}$$

Note that this quantity is independent of any learned predictor and is a function of the data only. The **epistemic uncertainty** $\mathcal{E}(Q_\psi, s, a)$, which is computed for a given predictor, is defined as the total uncertainty of the predictor minus the aleatoric uncertainty:

$$\mathcal{E}(Q_\psi, s, a) = \mathcal{U}(Q_\psi, s, a) - \mathcal{A}(s,a), \tag{6}$$

where $\mathcal{E}(Q_\psi, s, a)$ is the squared distance between the true mean and estimate mean as shown in SM 2. Concretely, this decomposition can be useful in instances where you want to estimate epistemic uncertainty, but doing so directly is significantly more difficult than estimating total and aleatoric uncertainty, which is often the case. In Sec. 3, we provide a way to estimate quantities in this manner, which later we use to prioritize transitions in the replay buffer.

2.3.4 Ensembles of Distributions

Using an ensemble of distributional RL agents gives us a concrete prescription for computing epistemic uncertainty as well as aleatoric uncertainty. This approach was first formalized by Clements et al. (2020), who define learned aleatoric and epistemic uncertainty quantities as a decomposition of the variance of the estimation from the ensemble (here defined as total uncertainty $\hat{\mathcal{U}}$) of distributional RL agents:

$$\hat{\mathcal{U}}(s,a) = \mathbb{V}_{\tau,\psi}\left[\theta_\tau(s,a;\psi)\right] = \hat{\mathcal{E}}(s,a) + \hat{\mathcal{A}}(s,a) \tag{7}$$

where

$$\hat{\mathcal{A}}(s,a) = \mathbb{V}_\tau[\mathbb{E}_\psi(\theta_\tau(s,a;\psi))], \quad \hat{\mathcal{E}}(s,a) = \mathbb{E}_\tau[\mathbb{V}_\psi(\theta_\tau(s,a;\psi))], \tag{8}$$

and s, a are state and action, $\psi \sim P(\psi|D)$ are the model parameters of each agent in the ensemble, D denotes the data distribution, and θ_τ is the value of the τth quantile. $\mathbb{V}$ and $\mathbb{E}$ are variance and expectation operators respectively. The term $\hat{\mathcal{E}}$ measures epistemic uncertainty as the expected disagreement (variance) in the parameters (quantiles) across the ensemble, that is, it approximates the variance of the posterior as sampled using the ensemble, which can be reduced by learning. The aleatoric uncertainty $\hat{\mathcal{A}}$ is computed by first averaging the predictions across the ensemble to obtain the average distribution estimated via quantile regression, and then taking the variance of this average distribution, which cannot be reduced by learning. In other words, it captures the variance of the estimated data distribution. The initial implementation by Clements et al. (2020) used a two-sample approximation for the posterior (i.e., an ensemble of two agents). However, Jiang et al. (2023) subsequently employed more explicit ensemble methods, which we also adopt in this work.

3 Uncertainty Prioritized Experience Replay

In this section we will introduce a new method for estimating epistemic uncertainty, which arises from a decomposition of the total uncertainty as defined by the average error over both the ensemble and quantiles. This decomposition is in the vein of Clements et al. (2020); however, it considers distance from the target in addition to the disagreement within the ensemble, thereby allowing us to handle, among others, model bias. We go on to derive an expression for prioritisation variables based

on the concept of *information gain*, which trades off epistemic and aleatoric uncertainty with a view to maximizing learnability from each sampled transition. We name this method Uncertainty Prioritized Experience Replay (UPER). Importantly, we are not changing the prioritize replay algorithm itself, but just the variable p_i used to prioritize in Eq. 2, replacing the TD-error by the information gain.

3.1 Uncertainty from Distributional Ensembles

The definitions given in Eq. 7 arise from a decomposition of $\mathbb{V}_{\psi,\tau}[\theta_\tau(s,a;\psi)]$ considered as total uncertainty in the original work (see Clements et al. 2020 for details). This quantity does not explicitly consider how far estimates are from targets, but rather how consistent the estimates are among the quantiles and members of the ensemble. We propose a modified concept of total uncertainty $\hat{\mathcal{U}}_\delta$ named *target total uncertainty*, simply defined as the average squared error to the target Θ over the quantiles and ensemble, which can be decomposed as:

$$\hat{\mathcal{U}}_\delta = \mathbb{E}_{\tau,\psi}[(\Theta(s',r) - \theta_\tau(s,a;\psi))^2] = \underbrace{\delta^2_\Theta(s,a) + \hat{\mathcal{E}}(s,a)}_{\hat{\mathcal{E}}_\delta(s,a)} + \hat{\mathcal{A}}(s,a); \tag{9}$$

where $\delta^2_\Theta(s,a) = (\Theta(s',r) - \mathbb{E}_{\tau,\psi}[\theta_\tau(s,a;\psi)])^2$, and we introduce the *target epistemic uncertainty* $\hat{\mathcal{E}}_\delta(s,a) = \delta^2_\Theta(s,a) + \hat{\mathcal{E}}(s,a)$ (see proof of this decomposition in App. A). Note that in order to construct ensemble disagreement estimates or estimates of the total uncertainty $\hat{\mathcal{U}}_\delta$, we assume independence among the ensemble, which is facilitated by masking and random initialization akin to bootstrapped DQN. Through the lens of the DEUP formulation from Sec. 2.3.3, this decomposition suggests a modified definition of epistemic uncertainty that considers the distance to the target δ^2_Θ as well as the disagreement in estimation within the ensemble $\hat{\mathcal{E}}$ from Clements et al. (2020) and Jiang et al. (2023). To see why this extra term can be useful, consider the following pathological example: all members of an ensemble are initialized equally, the variance among the ensemble, and the resulting epistemic uncertainty estimate without this additional error term, will be zero. A more subtle generalization of this would be if inductive biases from other parts of the learning setup (architecture, learning rule etc.) lead to characteristic learning trajectories in which individual members of the ensemble effectively collapse with no variance. In essence, $\hat{\mathcal{E}}$ assesses ensemble disagreement without including the estimation offset. The use of pseudo-counts (Lobel et al., 2023) presents a similar problem: while epistemic uncertainty does scale with the number of visits to a state, it does not necessarily encode the true distance between the estimation and target values. Pseudo-counts bear the additional disadvantage of being task agnostic, i.e. ignoring context, which makes them brittle under any change in the underlying MDP. We provide a simulation where we show the advantage of using $\hat{\mathcal{E}}_\delta$ instead of $\hat{\mathcal{E}}$ to prioritize replay in Sec. 4. Note that given a target and an estimate from an ensemble of distributional agents, the uncertainty decomposition in Eq. 9, and thus the information gain, can be computed. This extends to other learning settings beyond RL, where even if squared loss is not the optimization objective, the total uncertainty can serve as a surrogate for uncertainty estimation.

3.2 Prioritizing using Information Gain

Having arrived at suitable methods for estimating both epistemic and aleatoric uncertainty, it remains to establish a functional form for the prioritization variable, denoted $p_i = h(\mathcal{E}(s_i,a_i), \mathcal{A}(s_i,a_i))$. The most straightforward approach is to directly use $p_i = \hat{\mathcal{E}}_\delta$; however, in practical applications, this does not yield satisfactory results. One intuition for this, which will be made more concrete in later passages, is that the magnitude of epistemic uncertainty does not in itself determine how easily reducible that uncertainty is. It is informative therefore to also consider the aleatoric uncertainty, since this indicates the fidelity of the data, and hence how readily it can be used to reduce the epistemic uncertainty (this is demonstrated experimentally in Sec. 4.1 and SM 4, and expounded upon in SM 3).

We take inspiration from the idea of *information gain* to determine h. For the purpose of this explanation, consider a hypothetical dataset of points $x_i \sim \mathcal{N}(\mu_x, \sigma_x^2)$. Our objective is to estimate the

posterior distribution $P(\nu|x_i) \propto P(x_i|\nu)P(\nu)$ with a prior distribution $\nu \sim \mathcal{N}(\mu, \sigma^2)$. Following the observation of a single sample x_i, the posterior distribution becomes a Gaussian with variance $\sigma_\nu^2 = \frac{\sigma^2\sigma_x^2}{\sigma_x^2+\sigma^2}$. To quantify the information gained by incorporating the sample x_i when computing the posterior, we measure the difference in entropy between the prior distribution and the posterior as

$$\Delta\mathcal{H} = \mathcal{H}\left(P(\nu)\right) - \mathcal{H}\left(P(\nu|x_i)\right), \tag{10}$$

From here, we consider $\sigma^2 = \hat{\mathcal{E}}_\delta$ as a form of epistemic uncertainty, since the ensemble disagreement is reduced by sampling more points, and $\sigma_x^2 = \hat{\mathcal{A}}$ as aleatoric uncertainty corresponding to the variance of the ensemble average distribution, giving the irreducible noise of the data, obtaining a prioritization variable

$$p_i = \Delta\mathcal{H}_\delta = \frac{1}{2}\log\left(1 + \frac{\hat{\mathcal{E}}_\delta(s,a)}{\hat{\mathcal{A}}(s,a)}\right). \tag{11}$$

For a detailed derivation of the information gain, an illustrative simulation demonstrating the use of variance as an uncertainty estimate, and a comprehensive exploration of other functional forms of prioritization variables based on uncertainty, please refer to SM 3.

4 Motivating Examples

We proceed to employ epistemic uncertainty estimators and the information gain criterion in simple and interpretable toy models to highlight their potential as experience replay prioritization variables.

4.1 Conal Bandit

We devise a multi-armed bandit task in which each arm has the same expected reward but with increasing noise level as per arm, forming a *cone* as shown from left to right in Fig. 1a. The memory buffer in this experiment has one transition per arm, and after sampling one arm, the observed reward is replaced in the buffer for the respective transition (as done in the toy example in the original PER paper, Schaul et al. 2016). Specifically, let n_a denote the number of arms; then the reward distribution r for arm a is defined as:

$$r(a) = \bar{r} + \eta \cdot \sigma(a), \quad \sigma(a) = a \cdot \sigma_{\text{max}}/(n_a - 1) + \sigma_{\text{min}}; \tag{12}$$

where $\bar{r}$ represents the expected reward, $\sigma(a)$ is the reward standard deviation associated with arm a, σ_{max} and σ_{min} are constant, and η is sampled from a centered, unit-variance Gaussian.

The choice of employing noisy arms serves the purpose of demonstrating that the TD-errors will inherently include the sample noise, regardless of whether the reward estimation for each arm $Q(a) = \mathbb{E}_j[\theta_j(a)]$ approximates the target value $\bar{r}$. We depict results for the bandit task using different variables to prioritize learning in Fig. 1b for $n_a = 5$, $\bar{r} = 2$, $\sigma_{\max} = 2$ and $\sigma_{\min} = 0.1$ (details in SM 4).

Four relevant prioritization schemes are shown in this section (see SM 4 for other prioritization schemes): TD-error (standard PER): $p_i = \frac{1}{N_e}\sum_\psi |r_i - Q(a_i;\psi)|$; Inverse count: $p_i = 1/\sqrt{1+C}$, where C denotes the number of times an arm has been sampled to update the reward estimate; Information gain (UPER): $p_i = \Delta\mathcal{H}_\delta$; True distance to target: $p_i = \mathcal{E}^* = |\bar{r} - Q(a_i)|$.

Prioritizing with epistemic uncertainty measures, such as UPER or inverse counts (a proxy for epistemic uncertainty), leads to improved training speed and final true Mean Squared Error (true MSE, averaged across all arms, between the estimated reward and the true mean reward), compared to $p_i = |\delta_i|$ (PER), as illustrated in Fig. 1b. Throughout the paper, we highlight that the TD-error includes aleatoric uncertainty, corresponding to the arm variance in this scenario, which is irreducible through learning (see SM 2.1 for more details). Therefore, the TD-error tends to over-sample arms

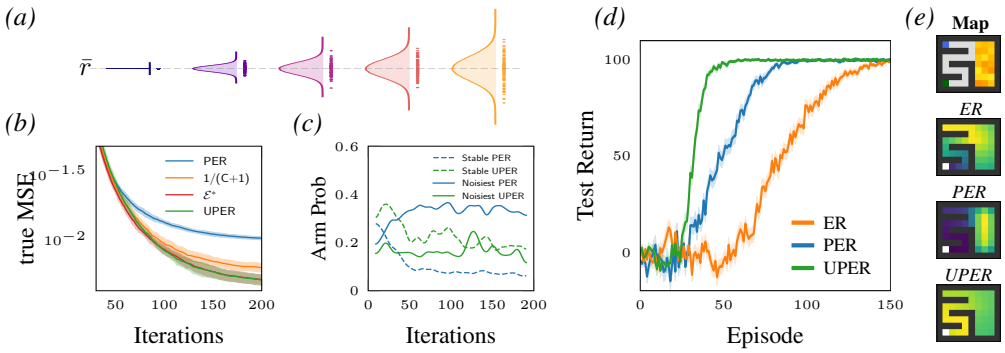

Figure 1: **Conal Bandit.** *(a)* multi-armed bandit task constructed such that each arm has identical mean payoff but increasing variance. *(b)* true MSE (average error across arms, between estimated reward and the true reward mean) over 200 iterations (each of 1000 steps) using different quantities to prioritise transitions from the replay buffer: absolute value of the TD error $|\delta|$ (PER), inverse counts (C being the number of visits to the respective arm), information gain $\Delta\mathcal{H}_\delta$ (UPER), and an oracle epistemic uncertainty $\mathcal{E}^*$ measured as the distance from the estimated mean to the true mean, shaded area is one standard deviation. *(c)* arm replaying selection probabilities for the stablest (dashed) and noisiest (solid) arms in the conal bandit; the key intuition is that prioritising by TD-error over-samples noisier arms, while prioritising using UPER places importance on learn-ability and leads to greater selection of stable arms. Results averaged across 10 seeds. **Noisy Gridworld.** *(d)* 300 seeds return on a test episode throughout training of an agent on the noisy gridworld, with the shaded region being one standard error on the mean. *(e)* in the **Map**, blue denotes the starting state, green is the goal state, and yellow are the non-zero variance immediate rewards. Below, sampling heatmaps where yellows are highly sampled and blues are scarcely sampled: uniform experience replay (ER) leads to sampling more from early parts of a trajectory since these fill the buffer first; replay based on TD error (PER) leads to a pathological sampling of the noisy part of the gridworld; replay using UPER leads to greater sampling of later parts of the trajectory.

with high variance compared with UPER, to the cost of not sampling the low variance arm. This is demonstrated in Fig. 1c.

Using inverse counts as the prioritization variable (similar to Lobel et al. (2023)), outperforms TD-error (as designed in the task) but not UPER. The reason is the fact that, although each initial estimated Q-values per arm are equidistant from the true mean, the learning speed for each arm diminishes with the variance of the respective arm. Inverse counts do not account for this variance-dependent decay in learning speed, so the number of updates per arm will not reflect the distance of the estimation to the true target, whereas UPER (prioritizing by $\hat{\mathcal{E}}_\delta$ and inverse $\hat{\mathcal{A}}$) tends to sample arms with high aleatoric uncertainty less frequently, and is also based on the distance to the targets as defined in Eq. 9.

The distance between the estimated mean and the true mean, denoted as $\mathcal{E}^*$ (accessible due to the task design), is equivalent to the epistemic uncertainty in the DEUP formulation, as derived in SM 2. This distance is the ideal prioritization variable to which we do not have access in general. Notably, using UPER, which prioritizes based on information gain, yields results comparable to prioritizing directly based on the true distance. These results show UPER as a promising modification to TD-error-based prioritized replay.

To emphasize the significance of incorporating the target value when utilizing the target epistemic uncertainty $\hat{\mathcal{E}}_\delta$ for replay prioritization, we introduced modifications to the conal bandit task by assigning distinct mean rewards per arm, denoted as $\bar{r} \to \bar{r}(a)$ (see simulation details in SM 4, Fig. 6). In the original conal bandit task, all arms shared the same mean reward $\bar{r}$, resulting in an equal initial distance expectation from $Q(a)$ to each arm. This uniformity dampened the performance improvement when considering the target distance $\delta_\ominus$ in $\hat{\mathcal{E}}_\delta$ with respect to $\hat{\mathcal{E}}$. By introducing varying

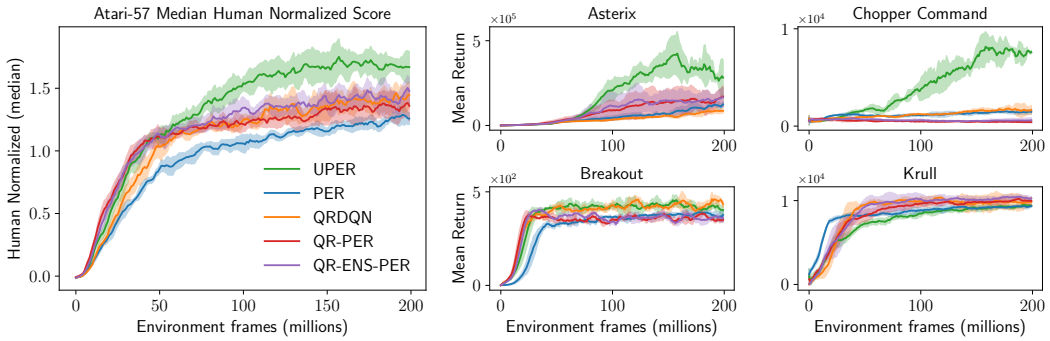

Figure 2: (**Left**) Comparing Uncertainty Prioritized Experience Replay (UPER) with Prioritized Experience Replay (PER) and QR-DQN on the full Atari-57 benchmark. Median human normalized score for UPER is significantly higher than baselines throughout the learning trajectory. (**Right**) Example of per-game performance, with vastly superior performance on e.g. Asterix and Chopper Command; cases in which UPER is worse are far less extreme, for instance Breakout and Krull (this is shown graphically in Fig. 14 and Fig. 15). All results are averages over 3 seeds. The shaded region indicates two standard deviations..

mean rewards per arm, denoted as $r(a)$, the relevance of information about the target value becomes important. This adjustment highlights the advantage of employing our proposed target epistemic uncertainty $\hat{\mathcal{E}}_\delta$ over merely considering ensemble disagreement $\hat{\mathcal{E}}$.

4.2 Noisy Gridworld

In order to move toward the full RL problem, we consider in this section a tabular gridworld. We take inspiration from ideas in planning within dynamic programming methods (Moore & Atkeson, 1993) to probe uncertainty-guided prioritized replay. Typically under this framework, 'direct' reinforcement learning on interactions with the environment (sometimes referred to as control) is supplemented with 'indirect' learning of a model from stored experiences (sometimes referred to as planning). In our case, we learn purely model-free but retain these ideas of offline vs. online learning. In some ways these methods are the precursor to the use of experience replay buffers in DRL. When making updates on stored data offline (for planning or otherwise), the same questions around criteria for prioritization arise. Notably, prioritized sweeping (preference over high error samples in memory) was an early extension to the Dyna models that exemplify this learning protocol (Sutton, 1991). In Fig. 1e Map, we construct a gridworld where the agent can encounter a set of very noisy states with random rewards early on in the episode while a single deterministic state with a much larger reward is at the end of the maze. Fig. 1d shows that this simple task can be solved without the additional planning steps, but ER (sampling uniformly) helps improve sample efficiency. This is improved further by PER (prioritizing using TD), but even more so by UPER where we prioritize using the information gain criterion and the inverse of state visitation counts (a good proxy for epistemic uncertainty in this tabular setting). As shown by the heatmaps in Fig. 1e, PER over-samples noisy states while UPER prioritizes on novel states towards the end of the trajectory. Full details of the experimental setup and hyper-parameters can be found in SM 5.

5 Deep RL: Atari

In our final set of experiments we apply our insights in a DRL setting, specifically the Atari benchmark (Bellemare et al., 2013). Our agent is an ensemble of QR-DQN distributional predictors (N=10), in which experience replay is prioritized using the information gain (UPER in Sec. 3.2). We compare this method to a vanilla QR-DQN agent (Dabney et al., 2017) with uniform prioritization and the original PER agent (Schaul et al., 2016). To show that the gain in performance is

not due to either the quantile regression method, nor the ensemble, we trained a QR-DQN agent with TD-error prioritization (QR-PER), and an ensemble of QR agents with TD-error prioritization (QR-ENS-PER). A summary of our empirical results is shown in Fig. 2, with further ablations and details in SM 6.

Except for the additional hyper-parameters associated with the ensemble of distributional prediction heads and a more commonly used configuration for the Adam optimizer ($\epsilon = 0.01/(\text{batch_size})^2$), the network architecture and all hyper-parameters in UPER are identical to QR-DQN (Dabney et al., 2017). Likewise PER, QRDQN, and QR-PER baselines follow the implementations of Dabney et al. (2017) and Schaul et al. (2016) respectively, while QR-ENS-PER is identical to UPER except for the prioritization variable which is TD-error. Concretely for the UPER agent, we compute the target epistemic uncertainty using $\hat{\mathcal{E}}_\delta(s,a) = \hat{\mathcal{U}}_\delta(s,a) - \hat{\mathcal{A}}(s,a)$. Then for a given transition i the total uncertainty is given by

$$\hat{\mathcal{U}}_\delta = \mathbb{E}_{\tau,\tau',\psi}\left[\left(r_i + \gamma\theta_{\tau'}(s_i', a_i'; \bar{\psi}) - \theta_\tau(s_i, a_i; \psi)\right)^2\right], \tag{13}$$

where τ (τ') are the quantiles of the online (target) network ψ ($\bar{\psi}$). The aleatoric uncertainty estimate is given by $\hat{\mathcal{A}}(s,a)$ in Eq. 8. From these estimates we construct UPER priority variable using the uncertainty ratio discussed in Sec. 3.2, i.e. Eq. 11. Since UPER and QR-ENS-PER are ensemble agents, we store a random mask $m \in \mathbb{R}^N$ for each transition in the buffer where $m_i \sim \mathcal{B}(0.5)$. When the transition is sampled for learning, gradients are only propagated for heads whose corresponding element in the mask is 1. This follows the proposal of (Osband et al., 2016) and serves to de-correlate the learning trajectories of the ensemble members, which is integral to the validity of our uncertainty estimates.

As depicted in Fig. 2, the median UPER performance across games is better than other prioritization schemes, showing that the performance improvement is not due to either the quantile regression technique or the ensemble alone. Importantly, UPER demonstrates performance improvement compared to its closest comparison QR-ENS-PER, whose only difference with UPER is the prioritization using TD-error (see Fig. 14). In most games where UPER does not improve performance, such as Krull, Q*bert or H.E.R.O., the difference in performance is not significant. This is shown in the panels per game in Fig. 15 and the asymmetry of the bar plots in SM 6.

6 Discussion and Conclusions

In this study, we propose using epistemic uncertainty measures to guide the prioritization of transitions from the replay buffer. We demonstrate both via mathematical analysis and careful experiments that the typically applied TD-error criterion can include aleatoric uncertainty, and lead to over-sampling of noisy transitions. Prioritizing by a principled function of epistemic and aleatoric uncertainty in the form of the information gain mitigates these effects. To construct this function, we expand the concept of epistemic uncertainty from Clements et al. (2020) to incorporate the distance to the target, achieving performance advantages in toy settings and complex problems such as the Atari 57 benchmark. In estimating these auxiliary quantities, one concern may be the increased computational cost in the deep learning setting. However, sharing of the lower level representation over multiple heads alongside efficient implementations can significantly mitigate this burden. To demonstrate this, we conducted an experiment on a lower-capacity GPU comparing the training times of DQN, QR-DQN, and QR-DQN + ensemble networks in the Pong environment. The time per iteration is presented in Table 1. The comparable training times can be attributed to effective batch processing facilitated by GPU parallelization. In our implementation, each agent in the ensemble is represented by a distinct output head in the network architecture. By extending the batch dimension to (batch, action, quantiles, ensemble), we leverage the parallelization capacity of the GPU, which still operates within capacity for the QR-DQN ensemble network. Further details of this experiment and the computer architecture used are presented in SM 6.2. Note that this analysis does not aim to evaluate or compare the computational cost of sampling with a priority variable vs. uniform sampling. This is already addressed in the original PER paper and has negligible impact (Schaul et al., 2016).

Table 1: Computational Cost
(seconds per iteration with one standard deviation for 20 runs each)

Architecture	CPU	GPU
QR-DQN-ENS	28.40 ± 0.26	20.74 ± 0.43
QR-DQN	17.80 ± 0.13	18.49 ± 0.68
DQN	18.34 ± 0.09	18.39 ± 0.56

While we focus our implementation on distributional RL—a widely used set of methods, exploring other forms of uncertainty estimation in RL such as pseudo-counts (Lobel et al., 2023), in combination with different functional forms outside information gain, is a promising research path both for different prioritization schemes and related parts of the RL problem like exploration (see SM 3 and SM 4).

The framework of combining epistemic and aleatoric uncertainties in an information gain introduced in this work is not restricted to reinforcement learning. In principle, these concepts can be extrapolated to other learning systems. A substantial body of literature exists on the efficient selection of datapoints to enhance learning in other paradigms such as supervised (Hüllermeier & Waegeman, 2021; Zhou et al., 2022), continual (Henning et al., 2021; Li et al., 2021), or active learning (Nguyen et al., 2022). In addition, our work has the potential to offer alternative insights into replay events in biological agents (Daw et al., 2005; Mattar & Daw, 2018; Liu et al., 2019; Schiffer, 2019; Antonov et al., 2022; Wittkuhn et al., 2025).

A Total Error Decomposition

Here we extend the notion of uncertainty proposed by Clements et al. (2020) by replacing the variance of the estimation as total uncertainty, instead using the averaged square error to the target Θ over the quantiles and ensemble as total uncertainty, which we call *target total uncertainty*. Quantiles and ensemble are indexed by τ and ψ respectively, hence the target total uncertainty can be written as $\hat{\mathcal{U}}_\delta = \mathbb{E}_{\tau,\psi}[(\Theta(s',r) - \theta_\tau(s,a;\psi))^2]$, dropping the dependency on the transition (s', r, s, a) for simplicity, it can be decomposed into:

$$\mathbb{E}_{\tau,\psi}[(\Theta - \theta_\tau(\psi))^2] = \int_\psi \frac{1}{N} \sum_\tau^N (\Theta - \theta_\tau(\psi))^2 P(\psi|D) d\psi, \tag{14}$$

$$\tag{15}$$

we add and subtract $\mathbb{E}_\psi(\theta_\tau(\psi))$ in the squared term

$$\mathbb{E}_{\tau,\psi}[(\Theta - \theta_\tau(\psi))^2] = \int_\psi \frac{1}{N} \sum_\tau^N \left[\Theta - \theta_\tau(\psi) \pm \mathbb{E}_\psi(\theta_\tau(\psi))\right]^2 P(\psi|D) d\psi, \tag{16}$$

$$= \int_\psi \frac{1}{N} \sum_\tau^N \Big[\left(\Theta - \mathbb{E}_\psi(\theta_\tau(\psi))\right)^2 + \left(\mathbb{E}_\psi(\theta_\tau(\psi)) - \theta_\tau(\psi)\right)^2 \tag{17}$$

$$+2\left(\Theta - \mathbb{E}_\psi(\theta_\tau(\psi))\right)\left(\mathbb{E}_\psi(\theta_\tau(\psi)) - \theta_\tau(\psi)\right)\Big] P(\psi|D) d\psi, \tag{18}$$

$$= \int_\psi \frac{1}{N} \sum_\tau^N \left(\Theta - \mathbb{E}_\psi(\theta_\tau(\psi))\right)^2 P(\psi|D) d\psi \tag{19}$$

$$+ \underbrace{\frac{1}{N} \sum_\tau^N \int_\psi \left(\mathbb{E}_\psi(\theta_\tau(\psi)) - \theta_\tau(\psi)\right)^2 P(\psi|D) d\psi}_{\hat{\mathcal{E}} \text{ in Eq. 8}}, \tag{20}$$

the term in Eq. 18 is zero when integrating over ψ. Finally, the term in Eq. 19 is

$$\int_{\psi} \frac{1}{N} \sum_{\tau}^{N} \left(\Theta - \mathbb{E}_{\psi}(\theta_{\tau}(\psi))\right)^2 P(\psi|D) d\psi = \Theta^2 - 2\mathbb{E}_{\psi,\tau}\left(\theta_{\tau}(\psi)\right) + \mathbb{E}_{\tau}\left(\mathbb{E}_{\psi}\left[\theta_{\tau}(\psi)\right]^2\right) \tag{21}$$

$$= \underbrace{\left(\Theta - \mathbb{E}_{\psi,\tau}\left[\theta_{\tau}(\psi)\right]\right)^2}_{\text{Distance to the target } \delta^2_{\Theta}} + \underbrace{\mathbb{V}_{\tau}\left(\mathbb{E}_{\psi}\left[\theta_{\tau}(\psi)\right]\right)}_{\hat{\mathcal{A}} \text{ in equation 7}}, \tag{22}$$

obtaining our proposed uncertainty decomposition

$$\hat{\mathcal{U}}_{\delta} = \mathbb{E}_{\tau,\psi}[(\Theta(s',r) - \theta_{\tau}(s,a;\psi))^2] = \delta^2_{\Theta}(s,a) + \hat{\mathcal{E}}(s,a) + \hat{\mathcal{A}}(s,a); \tag{23}$$

B Related Work

Exploration. While UPER is not explicitly promoting exploration through a reward bonus to unexplored or uncertainty states, we borrow methods from this field to estimate epistemic and aleatoric uncertainty (Clements et al., 2020) to prioritize transitions from the replay buffer based on the information gain. A fundamental dilemma faced by RL agents is the exploration-exploitation trade-off (Osband et al., 2016; O'Donoghue, 2023), in which agents must balance competing objectives for action selection, between uncovering new information about the environment (exploration) and accumulating as much reward as they currently can (exploitation). Replay sampling and exploration strategies both affect the data used to enhance the estimation of the value function. The former controls the experiences used for value estimation updates, while the latter selects experiences that will end up populating the replay buffer. Many exploration strategies have been built around ideas of intrinsic reward (Oudeyer & Kaplan, 2007) and episodic memory (Savinov et al., 2019; Badia et al., 2020). These are susceptible to pathological behavior induced by the noisy TV, and later variants are designed partly with this problem in mind; as a result they are frequently concerned with reliable and meaningful estimates of counts and novelty (Ostrovski et al., 2017b; Bellemare et al., 2016b; Burda et al., 2018; Lobel et al., 2023), dynamics (Stadie et al., 2015; Pathak et al., 2017), uncertainty (Mavor-Parker et al., 2022), and related quantities, many of which are relevant to our problem of constructing suitable measures for replay prioritization.

PER. Various efforts have been made to understand and improve upon aspects of prioritized experience replay since its introduction by Schaul et al. (2016). Integration of information related to uncertainty has often been in conjunction with strategies for managing the exploration-exploitation trade-off. For instance, in Sun et al. (2020), frequently visited states are sampled more frequently to reduce uncertainty around known states. Conversely, Alverio et al. (2022) approach is prioritizing uncertain states to encourage exploration, utilizing epistemic uncertainty estimated as the standard deviation across an ensemble of next-state predictors. This technique is combined with other methods to enhance sample efficiency.

Another method, presented in Lobel et al. (2023), employs a pseudo-count approximation to gauge state visits, fostering exploration as an intrinsic reward. In training the pseudo-count network they prioritize transitions according to the counts themselves; they do not however go as far as performing this prioritization for learning the actual value network, as is the focus of our work. The method of Lobel et al. (2023) allows estimation of epistemic uncertainty independent of the sparsity or density of the reward signal, making it especially appealing in sparse-reward environments. However, using pseudo-counts for epistemic uncertainty can also be poorly aligned with uncertainty about the actual value estimation problem (Osband et al., 2018). As described in Sec. 4.1, the number of visits to a specific state-action does not necessarily describe the error between the mean estimates to the true one. In addition to this, as explained in Sec. 3.2 and shown by simulation in Sec. 4.1, both epistemic and aleatoric uncertainty should be considered to build a proper prioritization scheme.

References

Julian Alverio, Boris Katz, and Andrei Barbu. Query The Agent: Improving sample efficiency through epistemic uncertainty estimation, October 2022. URL http://arxiv.org/abs/2210.02585. arXiv:2210.02585 [cs].

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*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/453fadbd8a1a3af50a9df4df899537b5-Abstract.html.

Georgy Antonov, Christopher Gagne, Eran Eldar, and Peter Dayan. Optimism and pessimism in optimised replay. *PLOS Computational Biology*, 18(1):e1009634, January 2022. ISSN 1553-7358. DOI: 10.1371/journal.pcbi.1009634. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1009634. Publisher: Public Library of Science.

John Asmuth and Michael L Littman. Learning is planning: near bayes-optimal reinforcement learning via monte-carlo tree search. *arXiv preprint arXiv:1202.3699*, 2012.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002a.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3, 2002b. ISSN 1533-7928. URL https://www.jmlr.org/papers/v3/auer02a.html.

Adrià Puigdomènech Badia, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, Bilal Piot, Steven Kapturowski, Olivier Tieleman, Martín Arjovsky, Alexander Pritzel, Andew Bolt, and Charles Blundell. Never Give Up: Learning Directed Exploration Strategies, February 2020. URL http://arxiv.org/abs/2002.06038. arXiv:2002.06038 [cs, stat].

Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. Emergent Tool Use From Multi-Agent Autocurricula, February 2020. URL http://arxiv.org/abs/1909.07528. arXiv:1909.07528 [cs, stat].

Yogesh Balaji, Mehrdad Farajtabar, Dong Yin, Alex Mott, and Ang Li. The Effectiveness of Memory Replay in Large Scale Continual Learning, October 2020. URL http://arxiv.org/abs/2010.02418. arXiv:2010.02418 [cs].

Andrew G Barto. Intrinsic motivation and reinforcement learning. *Intrinsically motivated learning in natural and artificial systems*, pp. 17–47, 2013.

M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling. The Arcade Learning Environment: An Evaluation Platform for General Agents. *Journal of Artificial Intelligence Research*, 47:253–279, June 2013. ISSN 1076-9757. DOI: 10.1613/jair.3912. URL https://jair.org/index.php/jair/article/view/10819.

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016a.

Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying Count-Based Exploration and Intrinsic Motivation, November 2016b. URL http://arxiv.org/abs/1606.01868. arXiv:1606.01868 [cs, stat].

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pp. 449–458. PMLR, 2017.

Marc G Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. MIT Press, 2023. URL http://www.distributional-rl.org.

Peter J. Bickel and David A. Freedman. Some Asymptotic Theory for the Bootstrap. *The Annals of Statistics*, 9(6):1196–1217, November 1981. ISSN 0090-5364, 2168-8966. DOI: 10.1214/aos/1176345637. URL https://projecteuclid.org/journals/annals-of-statistics/volume-9/issue-6/Some-Asymptotic-Theory-for-the-Bootstrap/10.1214/aos/1176345637.full. Publisher: Institute of Mathematical Statistics.

Matthew Botvinick, Jane X. Wang, Will Dabney, Kevin J. Miller, and Zeb Kurth-Nelson. Deep Reinforcement Learning and Its Neuroscientific Implications. *Neuron*, 107(4):603–616, August 2020. ISSN 0896-6273. DOI: 10.1016/j.neuron.2020.06.014. URL https://www.sciencedirect.com/science/article/pii/S0896627320304682.

Emma Brunskill. Bayes-optimal reinforcement learning for discrete uncertainty domains. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*, pp. 1385–1386, 2012.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by Random Network Distillation, October 2018. URL http://arxiv.org/abs/1810.12894. arXiv:1810.12894 [cs, stat].

Bertrand Charpentier, Ransalu Senanayake, Mykel Kochenderfer, and Stephan Günnemann. Disentangling Epistemic and Aleatoric Uncertainty in Reinforcement Learning, June 2022. URL http://arxiv.org/abs/2206.01558. arXiv:2206.01558 [cs].

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep Reinforcement Learning in a Handful of Trials using Probabilistic Dynamics Models, November 2018. URL http://arxiv.org/abs/1805.12114. arXiv:1805.12114 [cs, stat].

William R. Clements, Bastien Van Delft, Benoît-Marie Robaglia, Reda Bahi Slaoui, and Sébastien Toth. Estimating Risk and Uncertainty in Deep Reinforcement Learning, September 2020. URL http://arxiv.org/abs/1905.09638. arXiv:1905.09638 [cs, stat].

David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996.

Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional Reinforcement Learning with Quantile Regression, October 2017. URL http://arxiv.org/abs/1710.10044. arXiv:1710.10044 [cs, stat].

Nathaniel D. Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12):1704–1711, December 2005. ISSN 1546-1726. DOI: 10.1038/nn1560. URL https://www.nature.com/articles/nn1560. Number: 12 Publisher: Nature Publishing Group.

Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de las Casas, Craig Donner, Leslie Fritz, Cristian Galperti, Andrea Huber, James Keeling, Maria Tsimpoukelli, Jackie Kay, Antoine Merle, Jean-Marc Moret, Seb Noury, Federico Pesamosca, David Pfau, Olivier Sauter, Cristian Sommariva, Stefano Coda, Basil Duval, Ambrogio Fasoli, Pushmeet Kohli, Koray Kavukcuoglu, Demis Hassabis, and Martin Riedmiller. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, February 2022. ISSN 1476-4687. DOI: 10.1038/s41586-021-04301-9. URL https://www.nature.com/articles/s41586-021-04301-9. Number: 7897 Publisher: Nature Publishing Group.

William Fedus, Prajit Ramachandran, Rishabh Agarwal, Yoshua Bengio, Hugo Larochelle, Mark Rowland, and Will Dabney. Revisiting fundamentals of experience replay. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, pp. 3061–3071. JMLR.org, July 2020.

David J. Foster and Matthew A. Wilson. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, 440(7084):680–683, March 2006. ISSN 1476-4687. DOI: 10.1038/nature04587. URL https://www.nature.com/articles/nature04587. Number: 7084 Publisher: Nature Publishing Group.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Samuel J Gershman. Dopamine, inference, and uncertainty. *Neural Computation*, 29(12):3311–3326, 2017.

William H Greene. Econometric analysis 4th edition. *International edition, New Jersey: Prentice Hall*, pp. 201–215, 2000.

Christian Henning, Maria Cervera, Francesco D' Angelo, Johannes von Oswald, Regina Traber, Benjamin Ehret, Seijin Kobayashi, Benjamin F. Grewe, and João Sacramento. Posterior Meta-Replay for Continual Learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 14135–14149. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/761b42cfff120aac30045f7a110d0256-Abstract.html.

Matteo Hessel, Joseph Modayil, Hado van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining Improvements in Deep Reinforcement Learning, October 2017. URL http://arxiv.org/abs/1710.02298. arXiv:1710.02298 [cs].

Jan Humplik, Alexandre Galashov, Leonard Hasenclever, Pedro A Ortega, Yee Whye Teh, and Nicolas Heess. Meta reinforcement learning as task inference. *arXiv preprint arXiv:1905.06424*, 2019.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506, March 2021. ISSN 1573-0565. DOI: 10.1007/s10994-021-05946-3. URL https://doi.org/10.1007/s10994-021-05946-3.

Yiding Jiang, J. Zico Kolter, and Roberta Raileanu. On the Importance of Exploration for Generalization in Reinforcement Learning, June 2023. URL http://arxiv.org/abs/2306.05483. arXiv:2306.05483 [cs].

Gregory Kahn, Adam Villaflor, Vitchyr Pong, Pieter Abbeel, and Sergey Levine. Uncertainty-Aware Reinforcement Learning for Collision Avoidance, February 2017. URL http://arxiv.org/abs/1702.01182. arXiv:1702.01182 [cs].

Christos Kaplanis, Claudia Clopath, and Murray Shanahan. Continual Reinforcement Learning with Multi-Timescale Replay, April 2020. URL http://arxiv.org/abs/2004.07530. arXiv:2004.07530 [cs, stat].

Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards Continual Reinforcement Learning: A Review and Perspectives, November 2022. URL http://arxiv.org/abs/2012.13490. arXiv:2012.13490 [cs].

Johannes Kirschner, Tor Lattimore, Claire Vernade, and Csaba Szepesvari. Asymptotically Optimal Information-Directed Sampling. In *Proceedings of Thirty Fourth Conference on Learning Theory*, pp. 2777–2821. PMLR, July 2021. URL https://proceedings.mlr.press/v134/kirschner21a.html. ISSN: 2640-3498.

Roger Koenker and Kevin F. Hallock. Quantile Regression. *Journal of Economic Perspectives*, 15(4):143–156, December 2001. ISSN 0895-3309. DOI: 10.1257/jep.15.4.143. URL https://www.aeaweb.org/articles?id=10.1257/jep.15.4.143.

Salem Lahlou, Moksh Jain, Hadi Nekoei, Victor Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. DEUP: Direct Epistemic Uncertainty Prediction. Technical Report arXiv:2102.08501, arXiv, April 2022. URL http://arxiv.org/abs/2102.08501. arXiv:2102.08501 [cs, stat] type: article.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, Cambridge, 2020. ISBN 978-1-108-48682-8. DOI: 10.1017/9781108571401. URL https://www.cambridge.org/core/books/bandit-algorithms/8E39FD004E6CE036680F90DD0C6F09FC.

Honglin Li, Payam Barnaghi, Shirin Enshaeifar, and Frieder Ganz. Continual Learning Using Bayesian Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32 (9):4243–4252, September 2021. ISSN 2162-2388. DOI: 10.1109/TNNLS.2020.3017292. URL https://ieeexplore.ieee.org/document/9181489. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

Long-Ji Lin. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*, 8(3):293–321, May 1992. ISSN 1573-0565. DOI: 10.1007/BF00992699. URL https://doi.org/10.1007/BF00992699.

Qi Liu, Yanjie Li, Shiyu Chen, Ke Lin, Xiongtao Shi, and Yunjiang Lou. Distributional reinforcement learning with epistemic and aleatoric uncertainty estimation. *Information Sciences*, 644:119217, October 2023. ISSN 0020-0255. DOI: 10.1016/j.ins.2023.119217. URL https://www.sciencedirect.com/science/article/pii/S0020025523008022.

Ruishan Liu and James Zou. The Effects of Memory Replay in Reinforcement Learning, October 2017. URL http://arxiv.org/abs/1710.06574. arXiv:1710.06574 [cs, stat].

Yunzhe Liu, Raymond J. Dolan, Zeb Kurth-Nelson, and Timothy E.J. Behrens. Human Replay Spontaneously Reorganizes Experience. *Cell*, 178(3):640–652.e14, July 2019. ISSN 00928674. DOI: 10.1016/j.cell.2019.06.012. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867419306403.

Sam Lobel, Akhil Bagaria, and George Konidaris. Flipping Coins to Estimate Pseudocounts for Exploration in Reinforcement Learning, June 2023. URL http://arxiv.org/abs/2306.03186. arXiv:2306.03186 [cs].

Björn Lütjens, Michael Everett, and Jonathan P. How. Safe Reinforcement Learning with Model Uncertainty Estimates, March 2019. URL http://arxiv.org/abs/1810.08700. arXiv:1810.08700 [cs].

Vincent Mai, Kaustubh Mani, and Liam Paull. Sample Efficient Deep Reinforcement Learning via Uncertainty Estimation, May 2022. URL http://arxiv.org/abs/2201.01666. arXiv:2201.01666 [cs].

James John Martin. Bayesian decision problems and markov chains. *(No Title)*, 1967.

Marcelo G. Mattar and Nathaniel D. Daw. Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience*, 21(11):1609–1617, November 2018. ISSN 1546-1726. DOI: 10.1038/s41593-018-0232-z. URL https://www.nature.com/articles/s41593-018-0232-z. Number: 11 Publisher: Nature Publishing Group.

Augustine Mavor-Parker, Kimberly Young, Caswell Barry, and Lewis Griffin. How to Stay Curious while avoiding Noisy TVs using Aleatoric Uncertainty Estimation. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 15220–15240. PMLR, June 2022. URL https://proceedings.mlr.press/v162/mavor-parker22a.html. ISSN: 2640-3498.

Colin G. McNamara, Álvaro Tejero-Cantero, Stéphanie Trouche, Natalia Campo-Urriza, and David Dupret. Dopaminergic neurons promote hippocampal reactivation and spatial memory persistence. *Nature Neuroscience*, 17(12):1658–1660, December 2014. ISSN 1546-1726. DOI: 10.1038/nn.3843. URL https://www.nature.com/articles/nn.3843. Number: 12 Publisher: Nature Publishing Group.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015. ISSN 1476-4687. DOI: 10.1038/nature14236. URL https://www.nature.com/articles/nature14236. Number: 7540 Publisher: Nature Publishing Group.

Andrew W. Moore and Christopher G. Atkeson. Prioritized sweeping: Reinforcement learning with less data and less time. *Machine Learning*, 13(1):103–130, October 1993. ISSN 1573-0565. DOI: 10.1007/BF00993104. URL https://doi.org/10.1007/BF00993104.

Ashvin Nair, Bob McGrew, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Overcoming Exploration in Reinforcement Learning with Demonstrations, February 2018. URL http://arxiv.org/abs/1709.10089. arXiv:1709.10089 [cs].

Hiromi Narimatsu, Mayuko Ozawa, and Shiro Kumano. Collision Probability Matching Loss for Disentangling Epistemic Uncertainty from Aleatoric Uncertainty. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pp. 11355–11370. PMLR, April 2023. URL https://proceedings.mlr.press/v206/narimatsu23a.html. ISSN: 2640-3498.

Gergely Neu, Matteo Papini, and Ludovic Schwartz. Optimistic Information Directed Sampling, June 2024. URL http://arxiv.org/abs/2402.15411. arXiv:2402.15411 [cs].

Vu-Linh Nguyen, Mohammad Hossein Shaker, and Eyke Hüllermeier. How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning*, 111(1):89–122, January 2022. ISSN 1573-0565. DOI: 10.1007/s10994-021-06003-9. URL https://doi.org/10.1007/s10994-021-06003-9.

Rui Nian, Jinfeng Liu, and Biao Huang. A review On reinforcement learning: Introduction and applications in industrial process control. *Computers & Chemical Engineering*, 139:106886, August 2020. ISSN 0098-1354. DOI: 10.1016/j.compchemeng.2020.106886. URL https://www.sciencedirect.com/science/article/pii/S0098135420300557.

Nikolay Nikolov, Johannes Kirschner, Felix Berkenkamp, and Andreas Krause. Information-Directed Exploration for Deep Reinforcement Learning, March 2019. URL http://arxiv.org/abs/1812.07544. arXiv:1812.07544 [cs, stat].

Brendan O'Donoghue. Efficient Exploration via Epistemic-Risk-Seeking Policy Optimization, June 2023. URL http://arxiv.org/abs/2302.09339. arXiv:2302.09339 [cs].

OpenAI, Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, and others. Dota 2 with Large Scale Deep Reinforcement Learning, December 2019. URL http://arxiv.org/abs/1912.06680. arXiv:1912.06680 [cs, stat].

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep Exploration via Bootstrapped DQN, July 2016. URL http://arxiv.org/abs/1602.04621. arXiv:1602.04621 [cs, stat].

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. Epistemic neural networks. *arXiv preprint arXiv:2107.08924*, 2021.

Georg Ostrovski, Marc G Bellemare, Aäron Oord, and Rémi Munos. Count-based exploration with neural density models. In *International conference on machine learning*, pp. 2721–2730. PMLR, 2017a.

Georg Ostrovski, Marc G. Bellemare, Aaron van den Oord, and Remi Munos. Count-Based Exploration with Neural Density Models, June 2017b. URL http://arxiv.org/abs/1703.01310. arXiv:1703.01310 [cs].

Pierre-Yves Oudeyer and Frederic Kaplan. What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurorobotics*, 1, 2007. ISSN 1662-5218. URL https://www.frontiersin.org/articles/10.3389/neuro.12.006.2007.

Yangchen Pan, Jincheng Mei, Amir-massoud Farahmand, Martha White, Hengshuai Yao, Mohsen Rohani, and Jun Luo. Understanding and mitigating the limitations of prioritized experience replay. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, pp. 1561–1571. PMLR, August 2022. URL https://proceedings.mlr.press/v180/pan22a.html. ISSN: 2640-3498.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven Exploration by Self-supervised Prediction, May 2017. URL http://arxiv.org/abs/1705.05363. arXiv:1705.05363 [cs, stat].

L. A. Prashanth and Mohammad Ghavamzadeh. Variance-constrained actor-critic algorithms for discounted and average reward MDPs. *Machine Learning*, 105(3):367–417, December 2016. ISSN 1573-0565. DOI: 10.1007/s10994-016-5569-5. URL https://doi.org/10.1007/s10994-016-5569-5.

Daniel Russo and Benjamin Van Roy. Learning to Optimize via Information-Directed Sampling. *Operations Research*, 66(1):230–252, February 2018. ISSN 0030-364X. DOI: 10.1287/opre.2017.1663. URL https://pubsonline.informs.org/doi/10.1287/opre.2017.1663. Publisher: INFORMS.

Nikolay Savinov, Anton Raichuk, Raphaël Marinier, Damien Vincent, Marc Pollefeys, Timothy Lillicrap, and Sylvain Gelly. Episodic Curiosity through Reachability, August 2019. URL http://arxiv.org/abs/1810.02274. arXiv:1810.02274 [cs, stat].

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized Experience Replay, February 2016. URL http://arxiv.org/abs/1511.05952. arXiv:1511.05952 [cs].

Anne-Marike Schiffer. Human replay. *Nature Human Behaviour*, 3(9):895–895, September 2019. ISSN 2397-3374. DOI: 10.1038/s41562-019-0740-1. URL https://www.nature.com/articles/s41562-019-0740-1. Publisher: Nature Publishing Group.

Craig Sherstan, Dylan R Ashley, Brendan Bennett, Kenny Young, Adam White, Martha White, and Richard S Sutton. Comparing Direct and Indirect Temporal-Difference Methods for Estimating the Variance of the Return. 2018.

David Silver, Aja Huang, Chris 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(7587):484–489, January 2016. ISSN 1476-4687. DOI: 10.1038/nature16961. URL https://www.nature.com/articles/nature16961%7D. Number: 7587 Publisher: Nature Publishing Group.

Matthew J. Sobel. The Variance of Discounted Markov Decision Processes. *Journal of Applied Probability*, 19(4):794–802, 1982. ISSN 0021-9002. DOI: 10.2307/3213832. URL https://www.jstor.org/stable/3213832. Publisher: Applied Probability Trust.

Bradly C. Stadie, Sergey Levine, and Pieter Abbeel. Incentivizing Exploration In Reinforcement Learning With Deep Predictive Models, November 2015. URL http://arxiv.org/abs/1507.00814. arXiv:1507.00814 [cs, stat].

Ajay Subramanian, Sharad Chitlangia, and Veeky Baths. Reinforcement learning and its connections with neuroscience and psychology. *Neural Networks*, 145:271–287, January 2022. ISSN 0893-6080. DOI: 10.1016/j.neunet.2021.10.003. URL https://www.sciencedirect.com/science/article/pii/S0893608021003944.

Peiquan Sun, Wengang Zhou, and Houqiang Li. Attentive Experience Replay. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5900–5907, April 2020. ISSN 2374-3468. DOI: 10.1609/aaai.v34i04.6049. URL https://ojs.aaai.org/index.php/AAAI/article/view/6049. Number: 04.

Richard Sutton and Andrew G Barto. *Reinforcement learning: An Introduction*. MIT Press, 2018.

Richard S. Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, 2(4):160–163, July 1991. ISSN 0163-5719. DOI: 10.1145/122344.122377. URL https://dl.acm.org/doi/10.1145/122344.122377.

Aviv Tamar, Dotan Di Castro, and Shie Mannor. Policy gradients with variance related risk criteria. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 1651–1658, Madison, WI, USA, June 2012. Omnipress. ISBN 978-1-4503-1285-1.

Aviv Tamar, Dotan Di Castro, and Shie Mannor. Learning the Variance of the Reward-To-Go. *Journal of Machine Learning Research*, 17(13):1–36, 2016. ISSN 1533-7928. URL http://jmlr.org/papers/v17/14-335.html.

Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. # exploration: A study of count-based exploration for deep reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

Open Ended Learning Team, Adam Stooke, Anuj Mahajan, Catarina Barros, Charlie Deck, Jakob Bauer, Jakub Sygnowski, Maja Trebacz, Max Jaderberg, Michael Mathieu, Nat McAleese, Nathalie Bradley-Schmieg, Nathaniel Wong, Nicolas Porcel, Roberta Raileanu, Steph Hughes-Fitt, Valentin Dalibard, and Wojciech Marian Czarnecki. Open-Ended Learning Leads to Generally Capable Agents, July 2021. URL http://arxiv.org/abs/2107.12808. arXiv:2107.12808 [cs].

Bradley Efron Tibshirani, R. J. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, New York, May 1994. ISBN 978-0-429-24659-3. DOI: 10.1201/9780429246593.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8:279–292, 1992.

David Weinberg, Qian Wang, Thomas Ohlson Timoudas, and Carlo Fischione. A Review of Reinforcement Learning for Controlling Building Energy Systems From a Computer Science Perspective. *Sustainable Cities and Society*, 89:104351, February 2023. ISSN 2210-6707. DOI: 10.1016/j.scs.2022.104351. URL https://www.sciencedirect.com/science/article/pii/S2210670722006552.

Greg Welch, Gary Bishop, et al. An introduction to the kalman filter. 1995.

Martha White and Adam White. A Greedy Approach to Adapting the Trace Parameter for Temporal Difference Learning, October 2016. URL http://arxiv.org/abs/1607.00446. arXiv:1607.00446 [cs, stat].

Lennart Wittkuhn, Lena M. Krippner, Christoph Koch, and Nicolas W. Schuck. Replay in the human visual cortex during brief task pauses is linked to implicit learning of successor representations. *Proceedings of the National Academy of Sciences*, 122(34):e2507516122, August 2025. DOI: 10.1073/pnas.2507516122. URL https://www.pnas.org/doi/10.1073/pnas.2507516122. Publisher: Proceedings of the National Academy of Sciences.

Aolin Xu and Maxim Raginsky. Minimum excess risk in bayesian learning. *IEEE Transactions on Information Theory*, 68(12):7935–7955, 2022.

Daochen Zha, Kwei-Herng Lai, Kaixiong Zhou, and Xia Hu. Experience Replay Optimization. pp. 4243–4249, 2019. URL https://www.ijcai.org/proceedings/2019/589.

Xinlei Zhou, Han Liu, Farhad Pourpanah, Tieyong Zeng, and Xizhao Wang. A survey on epistemic (model) uncertainty in supervised learning: Recent advances and applications. *Neurocomputing*, 489:449–465, June 2022. ISSN 0925-2312. DOI: 10.1016/j.neucom.2021.10.119. URL https://www.sciencedirect.com/science/article/pii/S0925231221019068.

Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. Varibad: A very good method for bayes-adaptive deep rl via meta-learning. *arXiv preprint arXiv:1910.08348*, 2019.

Supplementary Materials

The following content was not necessarily subject to peer review.

SM 1 Further Related Work

In the main text we focus primarily on related work in uncertainty estimation for reinforcement learning that is specific to the epistemic vs. aleatoric dichotomy. Here we give an extended discussion on uncertainty estimation methods more generally.

SM 1.1 Direct Variance Estimation

Distributional RL provides a framework for computing statistics of the return beyond the mean. Efforts to compute such quantities in RL date back to Sobel (1982), who derived Bellman-like operators for higher order moments of the return in MDPs that can be used to indirectly estimate variance. This has since been extended to a greater set of problem settings and models (Prashanth & Ghavamzadeh, 2016; Tamar et al., 2016; White & White, 2016). More recently methods have also been developed to directly estimate variance (Tamar et al., 2012); arguably the simplest such scheme for TD(0) learning is the following update rule for the action-value variance $\hat{\mathcal{A}}(s,a)$ at state s,a (re-estated from Sherstan et al. (2018) for state and action):

$$\hat{\mathcal{A}}_{t+1}(s,a) \leftarrow \hat{\mathcal{A}}_t(s,a) + \bar{\alpha}\bar{\delta}_t, \tag{24}$$

where

$$\bar{\delta}_t \leftarrow \bar{r}_{t+1} + \bar{\gamma}_{t+1}\hat{\mathcal{A}}_t(s',a') - \hat{\mathcal{A}}_t(s,a), \tag{25}$$

$$\bar{r}_{t+1} \leftarrow \delta_t^2, \tag{26}$$

$$\bar{\gamma}_{t+1} \leftarrow \gamma_{t+1}^2; \tag{27}$$

δ_t is the temporal difference error of on the mean value estimate, and $\bar{\alpha}$ is the variance learning rate. $\bar{r}$ can be thought of as a 'meta' reward for the variance estimate. This update corresponds to simply regressing on the square of the mean estimate error in a standard regression problem (single state, no concept of discounting) like in the bandit experiments shown in Sec. 4. This form of estimating aleatoric uncertainty does not require quantile regression, but

SM 1.2 Bayesian methods

A more comphrehensive Bayesian approach to the reinforcement learning problem can be formulated via so-called Bayes-adaptive Markov decision processes (BAMDPs) (Martin, 1967), where an agent continuously updates a belief distribution over underlying Markov decision processes. Solutions to BAMDPs are Bayes' optimal in the sense that they optimally trade off exploration and exploitation to maximise expected return. However, in all but the smallest environments and settings, learning over this entire belief distribution is intractable (Brunskill, 2012; Asmuth & Littman, 2012).

Posterior sampling, which can be viewed as the analogue of Thompson sampling for MDPs, has been a popular method to approximate the full Bayesian posterior e.g. via ensembles (Osband et al., 2016) or dropout (Gal & Ghahramani, 2016); extensions include provision of pseudo priors (Osband et al., 2018; 2021). While these approaches have been successful in some settings, they have few guarantees. A different line of work includes using methods such as meta-learning to reason on and train the approximate posterior (Zintgraf et al., 2019; Humplik et al., 2019).

With regards to the discussions on epistemic and aleatoric uncertainty, the above methods can give the model access to a distribution over parameters that can be sampled and operated on (e.g. to calculate variance). They do not however—Bayes optimal or not—lead *per se* to a decomposition into epistemic and aleatoric uncertainty.

SM 1.3 Counts

Another category of methods that are frequently used in reinforcement learning and related paradigms like bandits is based around notions of counts e.g. of state visitation. Such counts can be used to construct intervals/bounds on confidence of learned quantities. This is the foundation of well established exploration methods in tabular settings called upper confidence bounds (Auer, 2002b;a). In function approximation settings, much of the focus has been on constructing accurate *pseudo* counts that incorporate state similarities (Bellemare et al., 2016a; Ostrovski et al., 2017a; Tang et al., 2017). Despite the well demarcated distinction between count-based methods and those that address the Bayesian posterior above, with access to any mean-zero unit-variance distribution, an ensemble of mean-predictors of that distribution can be used to estimate pseudo-counts (Lobel et al., 2023). As a result, it is generally possible to convert a Bayesian posterior into pseudo-counts.

SM 1.4 Model-Based

A set of methods that is further removed from those used in our work, but are often motivated by similar questions consists of learning a model of the environment. Downstream quantities like the prediction error of the environment model can be used as proxies for uncertainty or novelty e.g. for exploration bonuses. Much of this work falls under the domain of intrinsic motivation (Barto, 2013). Some of the methods in this area e.g. curiosity (Pathak et al., 2017) attempt implicitly to make the distinction between epistemic uncertainty and aleatoric uncertainty to avoid the noisy TV problem.

SM 1.5 Beyond the prioritisation variable

Altering the prioritized experience replay is not confined to changing the prioritization variable. In Zha et al. (2019), the replay policy is adapted through gradient optimization. Balaji et al. (2020) introduces a regularization technique, enhancing continual learning by storing a compressed network activity version for replay. Additional methods encompass the utilization of sub-buffers storing transitions at multiple time scales (Kaplanis et al., 2020), replay for sparse rewards (Andrychowicz et al., 2017; Nair et al., 2018), and employing diverse sampling strategies (Pan et al., 2022). Further endeavors are aiming to understand the effects of PER in RL (Liu & Zou, 2017; Fedus et al., 2020).

SM 1.6 Information Directed Sampling

A closely related method is information-directed sampling (IDS, Russo & Van Roy 2018; Kirschner et al. 2021; Neu et al. 2024), which relies on a similar notion of information gain to the one defined in Sec. 3.2. In IDS, the action policy minimizes a quantity called the *information ratio* $\Psi_t(\pi)$, and the policy induced by IDS, π_t^{IDS}, is defined as

$$\pi_t^{\text{IDS}} = \arg\min_{\pi} \left\{ \Psi_t(\pi) = \frac{\Delta_t(\pi)}{g_t(\pi)} \right\}, \tag{28}$$

where t denotes the time index, $\Delta_t(\pi)$ is the expected regret averaged across a given policy π, and $g_t(\pi)$ is the expected *information gain* across all actions.

The information ratio minimized by IDS measures the *cost per unit of information acquired by following a certain policy*, effectively inducing a policy that selects actions with high *information gain* (i.e., informative actions) while avoiding those with high regret. This ratio explicitly formulates the exploration–exploitation trade-off: information gain versus regret/reward.

Importantly, the information gain for a given action is defined as

$$g_t(a) = \mathbb{E}\left[H(\alpha_t) - H(\alpha_{t+1}) \,|\, \mathcal{T}_t, A_t = a\right], \quad \text{and} \quad g_t(\pi) = \sum_{a \in A} \pi(a) g_t(a), \tag{29}$$

with α_t denoting the best possible policy before observing the outcome of action a at time t in trajectory $\mathcal{T}_t$, and α_{t+1} denoting the best possible policy after observing the outcome of that action. In

other words, this information gain is analogous to ours in that it quantifies how much information a new observation provides. In our case (UPER), observations are the transitions used to improve a value estimate, while in IDS the focus is on reducing uncertainty about the optimal policy after observing the outcome of an action. Another important difference is that IDS is used to sample actions for exploration, whereas UPER samples transitions from the buffer. Nevertheless, both methods could in principle be used for exploration and prioritization. Future work could explore their relationship and how both approaches could be leveraged for prioritization and intrinsic motivation to promote exploration.

SM 2 DEUP decomposition

Consider the total uncertainty as defined in Lahlou et al. (2022) (but adapted for RL), which can be decomposed into epistemic uncertainty (distance between the mean estimation and true mean) and aleatoric uncertain (target variance) as:

$$\begin{aligned}
\mathcal{U}(Q_\psi, s, a) &= \int \left(\Theta(s', r) - Q_\psi(s, a)\right)^2 P(s', r|s, a) ds' dr && (30)\\
&= \mathbb{E}_{s',r}\left[\left(\Theta(s', r) - Q_\psi(s, a)\right)^2\right] && (31)\\
&= \mathbb{E}_{s',r}\left[\Theta(s', r)^2\right] - 2Q_\psi(s, a)\mathbb{E}_{s',r}\left[\Theta(s', r)\right] + Q_\psi(s, a)^2 && (32)\\
&= \mathbb{V}_{s',r}\left[\Theta(s', r)\right] + \mathbb{E}_{s',r}\left[\Theta(s', r)\right]^2 - 2Q_\psi(s, a)\mathbb{E}_{s',r}\left[\Theta(s', r)\right] + Q_\psi(s, a)^2 && (33)\\
&= \underbrace{\mathbb{V}_{s',r}\left[\Theta(s', r)\right]}_{\text{aleatoric } \mathcal{A}(s,a)} + \underbrace{\left(Q_\psi(s, a) - \mathbb{E}_{s',r}\left[\Theta(s', r)\right]\right)^2}_{\text{epistemic } \mathcal{E}(Q_\psi,s,a)} && (34)
\end{aligned}$$

SM 2.1 Uncertainty decomposition in quantile regression

Here we provide some extra intuition on the difference between MSE curves when prioritising by total uncertainty $\mathcal{U}$, td-error $|\delta|$, estimated epistemic uncertainty $\hat{\mathcal{E}}_\delta$ and true epistemic uncertainty $\mathcal{E}^*$. Let's start by considering a single agent trained using quantile regression as explained in Sec. 2.3.2. Consider the expected squared error of all quantiles indexed by τ and the target distribution Z, also defined in Sec. 3.1 as $\mathcal{U}$:

$$\begin{aligned}
\mathcal{U}^2 = \mathbb{E}_{\tau,r\sim Z}\left[(r - \theta_\tau)^2\right] &= \mathbb{E}_r\left[r^2\right] - 2\mathbb{E}_r[r]\mathbb{E}_\tau\left[\theta_\tau\right] + \mathbb{E}_\tau\left[\theta_\tau^2\right], && (35)\\
&= \mathbb{V}_r\left[r\right] + \bar{r}^2 - 2\bar{r}Q(a) + Q(a)^2 + \mathbb{V}_\tau\left[\theta_\tau\right], && (36)\\
&= \underbrace{(\bar{r} - Q(a))^2}_{(\mathcal{E}^*)^2} + \underbrace{\mathbb{V}_r\left[r\right]}_{\text{Target variance}} + \underbrace{\mathbb{V}_\tau\left[\theta_\tau\right]}_{\text{Estimation variance}}. && (37)
\end{aligned}$$

The first term is the true epistemic uncertainty $\mathcal{E}^*$, second term and third term are the variance from the target, and the estimation variance. When using the total uncertainty as priority variable $p_i = \mathcal{U}$, the target and estimation uncertainty will be considered in the priority, therefore oversampling the noisiest arm as shown in the sampling probabilities depicted in Figures 7 and 8. When using the TD-error $p_i = |\delta_i|$, consider the expected squared TD-error

$$\begin{aligned}
\mathbb{E}_r\left[\delta^2\right] &= \left[(r - \mathbb{E}_\tau\left[\theta_\tau\right])^2\right], && (38)\\
&= \underbrace{(\bar{r} - Q(a))^2}_{(\mathcal{E}^*)^2} + \underbrace{\mathbb{V}_r\left[r\right]}_{\text{Target variance}}. && (39)
\end{aligned}$$

Therefore, the TD-error does not prioritise by estimation variance, but it includes the target variance. Eventually, the target variance will be equal to the estimation variance, but from the start of the training, this is not true. Hence, the TD-error will also oversample the noisiest arm, but less compared to prioritising by total uncertainty $\mathcal{U}$. In practice, we do not have direct to $\mathbb{V}_r\left[r\right]$, in fact this is a quantity we are trying to estimate by using quantile regression. We have implicit access to the true

distance $\mathcal{E}^*$ (epistemic uncertainty) through the decomposition $\mathcal{U} = \mathcal{E} + \mathcal{A}$ as explain in Sec. 2.3.3, which is used to estimate epistemic uncertainty as in Sec. 3. Prioritising using information gain achieve similar results compare to the direct use of $\mathcal{E}^*$ to prioritise replay. For further discussion about epistemic uncertainty ratios, refer to SM 3.3.

SM 3 Prioritisation Quantities based on Uncertainty

SM 3.1 Information gain derivation

Given the setup in Sec. 3.2, consider a hypothetical dataset of points $x_i \sim \mathcal{N}(\mu_x, \sigma_x^2)$. Our objective is to estimate the posterior distribution of the mean after observing one sample $P(\nu|x_i) \propto P(x_i|\nu)P(\nu)$ with a prior distribution of the mean $\nu \sim \mathcal{N}(\mu, \sigma^2)$. Following the observation of a single sample x_i, the posterior distribution is Gaussian with variance $\sigma_\nu^2 = \frac{\sigma^2\sigma_x^2}{\sigma_x^2+\sigma^2}$. Knowing that the entropy of a Gaussian random variable is $\mathcal{H}(P(\nu)) = 1/2\log(2\pi e\sigma^2)$, we proceed to compute the information gain (or entropy reduction) of the posterior distribution as

$$\Delta\mathcal{H} = \mathcal{H}(P(\nu)) - \mathcal{H}(P(\nu|x_i)) \tag{40}$$

$$= \frac{1}{2}\log\left(2\pi e\sigma^2\right) - \frac{1}{2}\log\left(2\pi e\left(\frac{\sigma^2\sigma_x^2}{\sigma_x^2+\sigma^2}\right)\right) \tag{41}$$

$$= \frac{1}{2}\log\left(1 + \frac{\sigma^2}{\sigma_x^2}\right). \tag{42}$$

We consider $\sigma^2 = \hat{\mathcal{E}}_\delta$ as a form of epistemic uncertainty that can be reduced by sampling more points, and $\sigma_x^2 = \hat{\mathcal{A}}$ as aleatoric uncertainty, which is the underlying irreducible noise of the data, giving a prioritisation variable

$$p_i = \Delta\mathcal{H}_\delta = \frac{1}{2}\log\left(1 + \frac{\hat{\mathcal{E}}_\delta(s,a)}{\hat{\mathcal{A}}(s,a)}\right). \tag{43}$$

As discussed in the main text, other form of priority variables p_i can be effective in some settings. We extend the discussion about uncertainty ratios in the following sections, and show empirical results in the arm bandit task in SM 4. This form of information gain is related to the Information Directed Sampling (IDS) described in SM 1.6.

SM 3.2 Variance as Uncertainty Estimation

To justify our choice of $\sigma^2 = \hat{\mathcal{E}}$ and $\sigma_x^2 = \hat{\mathcal{E}}$ in the information gain described in Eq. 11, we train an ensemble of distribution regressors to learn the mean from Gaussian samples ($\mu_x = 2, \sigma_x = 1$). This ensemble is compared to the Bayesian posterior distribution of the mean (Gaussian prior, likelihood, and posterior) as detailed in Sec. 3.2. The ensemble, composed of 50 distribution quantile regressors, is initialized with the same prior as the Bayesian model – a unit variance Gaussian centered at 0 – by sampling 50 values from this prior and setting the initial mean of each quantile regressor accordingly. Both the ensemble and Bayesian models are trained using samples from the data distribution. The ensemble training process follows the method described in the paper, and where each regressor is updated with a probability of 0.5 to introduce ensemble variability. The updates are performed using quantile regression as outlined in Sec. 2.3.2. At each time step, the ensemble's estimated posterior is computed by averaging the means of all regressors and calculating the variance of these means.

Fig. 3 (a) and (b) illustrate the posterior evolution of both models from the same starting prior, given more samples. Both posteriors exhibit similar trends (the Bayesian model converges faster to the mean, due to the use of TD-updates with a smaller learning rate in the ensemble). In the Bayesian model, posterior sharpness is quantified by its variance, σ_ν^2, whereas for the ensemble, it corresponds to the epistemic uncertainty $\hat{\mathcal{E}}$ from Eq. 8. Both measures converge to zero, but at different

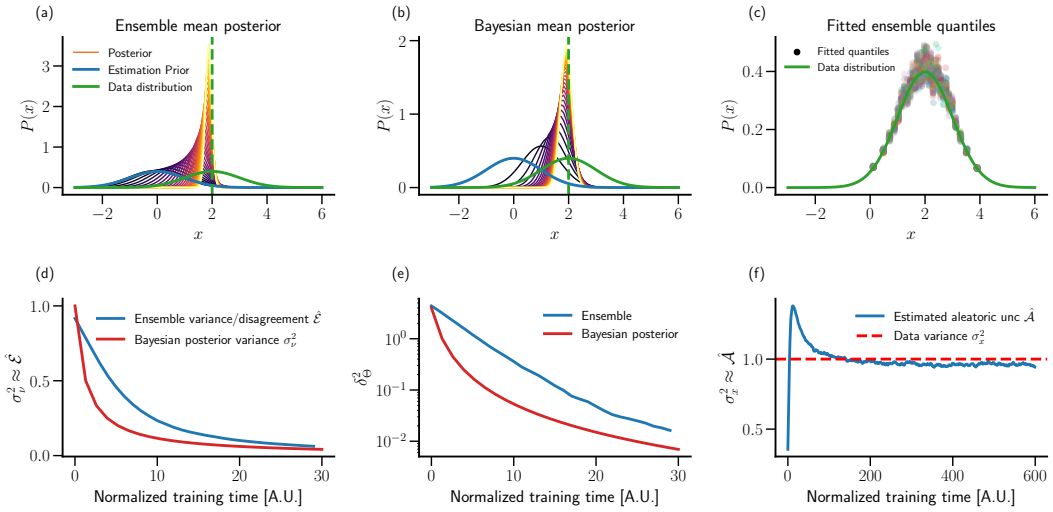

Figure 3: **Variances in the information gain can be approximated by epistemic and aleatoric uncertainty in the information gain**: **(a) and (b)** Evolution during training of the posterior of the mean using an ensemble (gaussian fitted to members of the ensemble at each step) and an ideal Gaussian respectively, as described in Sec. 3.2. Training progresses from purple to yellow. **(c)**: Fitted ensemble quantiles to true data distribution. **(d)**: Ensemble disagreement (equivalent to variance of the posterior estimated with ensemble as $\hat{\mathcal{E}}$ in Eq. 8) and true posterior variance σ_ν^2 from ideal Gaussian. **(e)**: Distance to the target true value δ_Θ. **(f)**: Data variance σ_x^2 approximated with $\mathcal{A}$ in Eq. 8. Training time was scaled to show a match between Gaussian posterior and uncertainty measures.

rates Fig. 3d. The aleatoric uncertainty of the data, by definition the variance σ_x^2, is well approximated by $\hat{\mathcal{A}}$ from Eq. 8, and shown in Fig. 3f. The slight underestimation of the variance is a known issue in quantile regression, as quantiles often fail to capture lower probability regions (Fig. 3c), leading to an underestimation of the distribution's variance. Our contribution to prioritization involves incorporating the distance to the target δ_Θ from Eq. 23 (Fig. 3e). This approach prioritizes transitions not only based on the reduction in posterior variance but also on the regressor's proximity to the target.

SM 3.3 Uncertainty Ratios

Having arrived at various methods for estimating epistemic and aleatoric uncertainty using distributional reinforcement learning, we now consider how to construct prioritisation variables from these estimates. Naively, one might consider prioritising directly using the epistemic uncertainty estimate; but neglecting the inherent noise or aleatoric uncertainty entirely ignores the 'learnability' of the data. Many methods in related learning domains can be interpreted as incorporating both uncertainties, including Kalman learning Welch et al. (1995); Gershman (2017), active learning Cohn et al. (1996), weighted least-squares regression Greene (2000), and corresponding extensions in deep learning and reinforcement learning Mai et al. (2022). To gain an intuition on how the choice of functional form might impact our particular use-case of prioritisation for various magnitudes of epistemic and aleatoric uncertainty.

$\mathcal{E}/\mathcal{A}$ has desirable properties. For instance under Bayesian learning of Gaussian distributions, $\log(1 + \mathcal{E}/A)$ maximises information gain (see Sec. 3.2), but discontinuities around very low noise must be dealt with—for instance by adding small constants to the denominator. Normalising instead with the total uncertainty is another way of handling the discontinuities. $\mathcal{E}^2/\mathcal{U}$ in particular corresponds to maximising reduction in variance under Bayesian learning in the same Gaussian setting. Both of these forms have the advantage over e.g. $\mathcal{E}/\mathcal{A}$ of preferring low epistemic uncertainty for

equal ratios of epistemic and aleatoric uncertainties, i.e. they are not constant along the diagonal of the phase diagram. More generally, it is difficult to say *a priori* which functional form is optimal. Many factors, including the data distributions, model and learning rule will play a role. Further discussion on these considerations can be found in Sec. SM 3.4. These trade-offs are also borne out empirically in the experimental Sec. 4 & Sec. 5 below.

SM 3.4 Bias as Temperature

Lahlou et al. (2022) and others make an equivalence between excess risk and epistemic uncertainty. Concretely, if $f^*(x)$ is the Bayes optimal predictor, the excess risk is defined as:

$$\text{ER}(f, x) = R(f, x) - R(f^*, x), \tag{44}$$

where R is the risk and $R(f^*, x)$ can be thought of as the aleatoric uncertainty.

One possible issue arises in overstating the connection between excess risk and epistemic uncertainty. Consider the case where there is model mis-specification, and f^* is not in the model class; then assuming the model class is fixed (as is standard), then the lower bound of $\text{ER}(f, x)$ is non-zero. Stated differently, it is *not* fully reducible, which is often viewed as a central property of epistemic uncertainty. For some applications this distinction may not be important; there is some non-zero lower bound to the epistemic uncertainty but the ordering and correlations are intact under this equivalence. But it could also play a significant role. For us in particular, adopting this equivalence has two related consequences:

1. The model mis-specification acts as a temperature for our prioritisation distribution;
2. The ratio, or more generally the functional form of our prioritisation variable, can offset this temperature.

To make the above equation fully reducible, we would need to further subtract a term capturing the difference between the Bayes predictor, and the best predictor in the model class i.e. the model bias or mis-specification term. Let us denote this term by C, and assume it constant over the domain. And let us denote the fully reducible uncertainty by η. In the case where we use the excess risk, the prioritisation of sample i is given by

$$\text{[Vanilla]} \qquad p_i = \frac{\eta_i + C}{\sum_i (\eta_i + C)} = \frac{\eta_i + C}{NC + \sum_i \eta_i}. \tag{45}$$

It is easy to see how C acts as a temperature. In the limit of large C we get a uniform distribution over samples. Similarly if $C = 0$ we recover the 'true' distribution for reducible uncertainty.

It is of course hard to measure this model mis-specification term. In large networks we can assume the capacity is unlikely to be restrictive, but perhaps other parts of the training regime could play a part. Importantly, the above holds true not just for model mis-specification, but also if there is any systematic error in the epistemic uncertainty estimate (i.e. think of C as an error on the epistemic uncertainty estimate).

SM 3.5 Prioritisation Distribution Entropy

Assuming the above effect is significant, might a different functional form (as discussed in Sec. SM 3.3) for prioritisation alleviate the impact? Consider the following additional options:

$$[\mathcal{E}/\mathcal{U}] \qquad p_i = \frac{\frac{\eta_i + C}{\eta_i + C + \beta_i}}{\sum_i \frac{\eta_i + C}{\eta_i + C + \beta_i}}; \tag{46}$$

$$[\mathcal{E}^2/\mathcal{U}] \qquad p_i = \frac{\frac{(\eta_i + C)^2}{\eta_i + C + \beta_i}}{\sum_i \frac{(\eta_i + C)^2}{\eta_i + C + \beta_i}}; \tag{47}$$

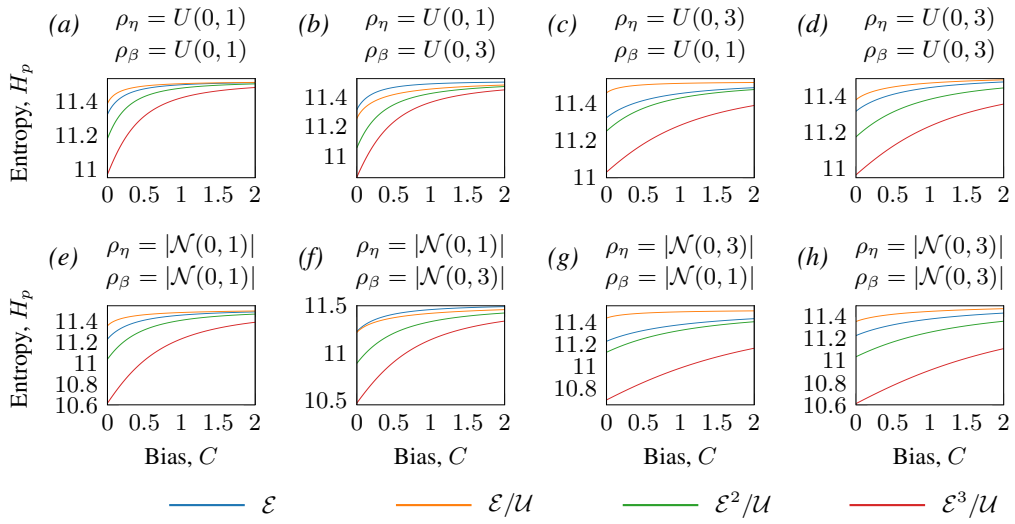

Figure 4: **Ratios can reduce entropy of distribution under bias.**

and more generally,

$$[\mathcal{E}^m/\mathcal{U}] \qquad p_i = \frac{\frac{(\eta_i+C)^m}{\eta_i+C+\beta_i}}{\sum_i \frac{(\eta_i+C)^m}{\eta_i+C+\beta_i}}. \tag{48}$$

In the limit of large C all of these forms tend to a uniform distribution. However, at what rate? And is there anything else interesting we can say?

Consider the following toy problem:

- Populate "replay" buffer with N samples;
- Each sample's reducible uncertainty is sampled from ρ_η;
- Each sample's reducible uncertainty is sampled from ρ_β;
- C is constant over the samples.

We can plot as a function of C the entropy of the prioritisation distribution for the functional forms above. Such a plot is shown for various choices of ρ_η, ρ_β in Fig. 4. Clearly, as C increases the entropy in the distribution increases and saturates at some maximum entropy. There is some variation in the entropy ordering depending on the exact ρ_η, ρ_β distributions; in some instances the vanilla form is lower entropy than $\mathcal{E}/\mathcal{U}$, but in general the entropy remains lower for longer (as a function of C) when the exponent in the nominator is higher. This is not a particularly surprising result, but lends support to the idea that a higher order function of $\mathcal{E}$ in a ratio form is desirable for prioritisation.

SM 3.6 Relation to $\mathcal{E}$ under 0 Bias

Now let us consider a more interesting measure. Ordinarily, or naively—in the sense that this is the first order approach—we want our prioritisation variable to be the vanilla prescription; and ideally we would want C to be 0. We can measure the difference, which we denote δ_i to this ideal for each functional form as a function of C. This plot is show for various choices of ρ_η, ρ_β in Fig. 5.

In general, the standard $\mathcal{E}/\mathcal{U}$ ratio is poor, it has systematically higher mean and variance of error. Beyond that, a clear trade-off emerges: as you increase the exponent m, then for high C there is lower deviation from the 'correct' distribution for priority. This is related to maintaining lower entropy and tending to a uniform distribution more slowly. However, for lower C you are likely to be more wrong, catastrophically so. This trade-off for $m = 3$ is effectively crossed when the red

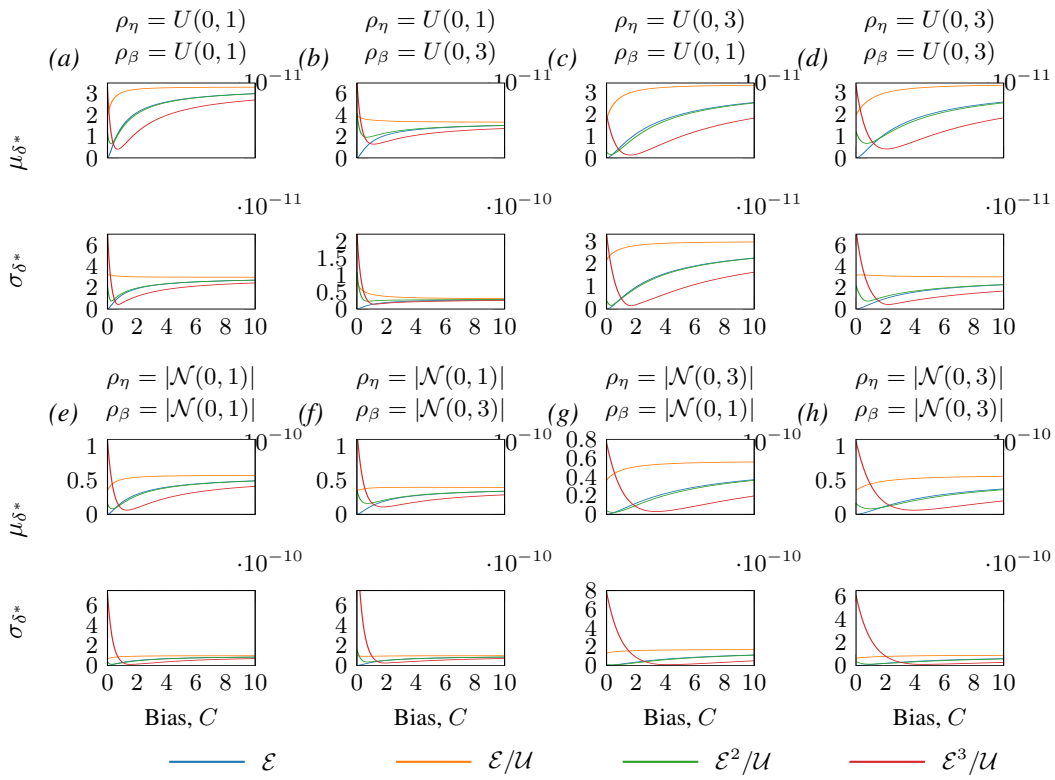

Figure 5: $\mathcal{E}^2/\mathcal{U}$ **closely approximates** E **for non-trivial bias.**

line intersects with the blue in these plots. The point at which this intersection happens will be a function of various things, primarily the underlying distributions—in this case ρ_η, ρ_β.

Interestingly however, for $m = 2$ there is very fast convergence of $\mathcal{E}^2/\mathcal{U}$ and $\mathcal{E}$ as a function of C. So while $m = 3$ has a very stark trade-off, $m = 2$ is less extreme: For low C it may make you more wrong but generally you will have very similar average error by this metric to the vanilla case; all the while the entropy of the distribution will be much lower and more informative (as shown in Fig. 4). This toy model is clearly very simplistic, not least the lack of variation in C over the samples; but future work could be dedicated to understanding these trade-offs more formally in the context of prioritized replay.

SM 3.7 Off-setting Bias with TD Term

Leaving aside the ratio forms, the consequences of the temperature effect may differ depending on the choice of epistemic uncertainty estimate we use. The methods we discuss in Sec. 2 & Sec. 3 all effectively use the equivalence of excess risk and epistemic uncertainty, and so do not explicitly consider the possibility of model bias. The possible exception is the method resulting from the expansion of the average error over the quantiles and ensemble in Sec. 3.1. The main difference between this decomposition and that of Clements et al. (2020) is a term that encodes the distance from the target:

$$\delta_\Theta^2 = \left(\Theta - \mathbb{E}_{\psi,i}\left[\theta_i(\psi)\right]\right)^2 . \tag{49}$$

This term *could* guard against two possible shortcomings of the decomposition in Clements et al. (2020):

1. Consider the pathological case in which each ensemble is initialised identically, then each quantile will have zero variance and the epistemic uncertainty measure from Clements et al. (2020) will be zero. Even if there is independence at initialisation, there may be characteristic learn-

ing trajectories or other systematic biases that push the ensemble together and lead to an underestimate in epistemic uncertainty. Here, the term above—if treated as part of the epistemic uncertainty—can continue to drive learning in ways we want.

2. However, it could be that the ensemble behaves nicely and the metric over the ensemble from Clements et al. (2020) is principally a good one, *but* that there is significant model bias. This could also be captured by the term above but would need to be *subtracted* from the total error in order to get a fully reducible measure for epistemic uncertainty (as per the argument discussed above).

Which of the two problems is more pronounced is difficult to know *a priori*, and could be an avenue for future work. Empirically, the performance of the UPER agent in Sec. 5 suggests that the former is the greater effect—at least on the atari benchmark with the model architecture and learning setting used.

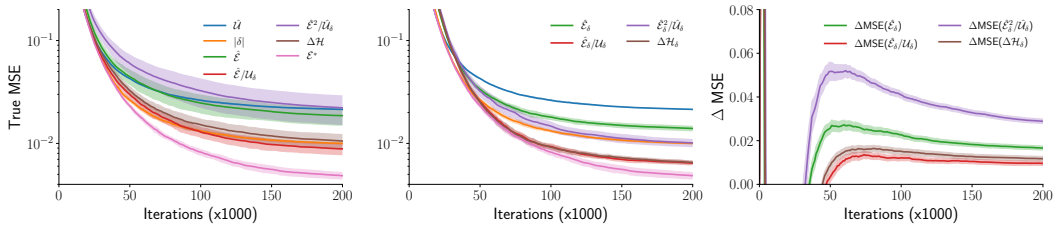

Figure 6: Comparison of MSE for different prioritisation scheemes. Left panel, shows ratios and information gain based on epistemic uncertainty $\hat{\mathcal{E}}$ proposed by Clements et al. (2020). Middle panel, shows ratios and information gain based on our proposed *target epistemic uncertianty* $\hat{\mathcal{E}}_\delta$. Right panel, different in MSE between curves in the left panel and right panel for the shifted arm task. For instance, $\Delta\text{MSE}(\hat{\mathcal{E}}_\delta) = \text{MSE}(\hat{\mathcal{E}}) - \text{MSE}(\hat{\mathcal{E}}_\delta)$, showing that our proposed $\hat{\mathcal{E}}_\delta$ is in general better for prioritisation in the arm-bandit task. Averaged across 10 seeds.

SM 4 Arm-Bandit Task

The hyperparameters used in the Arm-Bandit Task shown in Sec. 4 are shown below:

- Number of train steps: 20^5
- Learning rate annealing: $0.005 \cdot 2^{-\text{iters}/40000}$.
- Init variance estimation: uniformly sampled from to 0.1
- Number of agents in the ensemble: 30
- $\alpha = 0.7$, β is annealing from 0.5 to 1 in 0.4 to 1 in proportional prioritisation as in the original work by Schaul et al. (2016).
- n arms: $n_a = 5$, $\bar{r} = 2$, $\sigma_{\max} = 2$ and $\sigma_{\min} = 0.1$.
- Number of quantiles: 30.
- Quantiles initialized as uniform distribution between -1 and 1. For the main results in **??**, θ_τ are initialized randomly between -1 and 1, then sorted to describe a cumulative distribution.
- Each agent in the ensemble is updated with probability 1/2 on each step.
- For the shifted arm experiment, the mean reward per arm $\bar{r}(a) = 3, 2.75, 2.5, 2.25, 2$ for arms $1, 2, 3, 4$ and 5.

Fig. 6 show the mean squared error from the estimated $Q(a) = \mathbb{E}_{j,\psi}\left[\theta_j(\psi)\right]$ to the true mean, where ψ denotes agents in the ensemble case. Fig. 8 and Fig. 8 show the probability of sampling each arm from the memory buffer throughout the training, and the mean square error from the estimated arm value $Q(a)$ to the true arm value $\bar{r}$ (the same for every arm). In addition, we depict the evolution of uncertainty quantities for all prioritisation variables for the arm bandit task in Fig. 9.

SM 5 Gridworld Experiments

The hyperparameters used in Fig. 1 are listed below:

- Learning rate: 0.1
- Discount factor, γ: 0.9
- Exploration co-efficient, ϵ: 0.95
- Buffer capacity: 10,000
- Episode timeout: 1000 steps
- Random reward distribution: $\mathcal{N}(0, 2)$

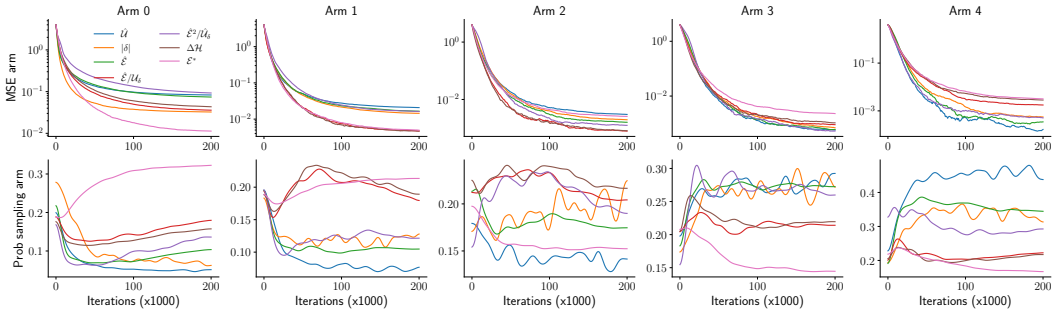

Figure 7: Comparison of MSE for different prioritisation scheemes using $\hat{\mathcal{E}}$ based prioritisation. Total uncertainty $\mathcal{U}$ and TD-error prioritisation tend to oversample high variance arms compared to epistemic uncertainty prioritisation.

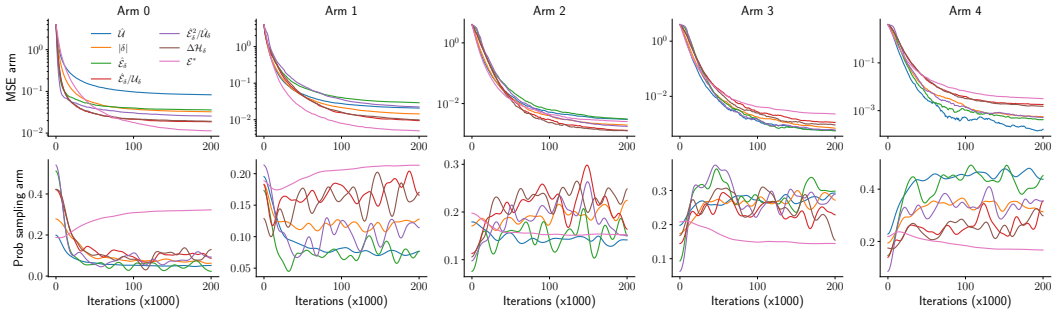

Figure 8: Comparison of MSE for different prioritisation scheemes using $\hat{\mathcal{E}}_\delta$ based prioritisation. Total uncertainty $\mathcal{U}$ and TD-error prioritisation tend to oversample high variance arms compared to epistemic uncertainty prioritisation.

For every 10 steps of 'direct' interaction and learning from the environment, the agent makes 5 updates with 'indirect' learning from the buffer replay. The data shown in the plots consists of 100 repeats and is smoothed over a window of 10.

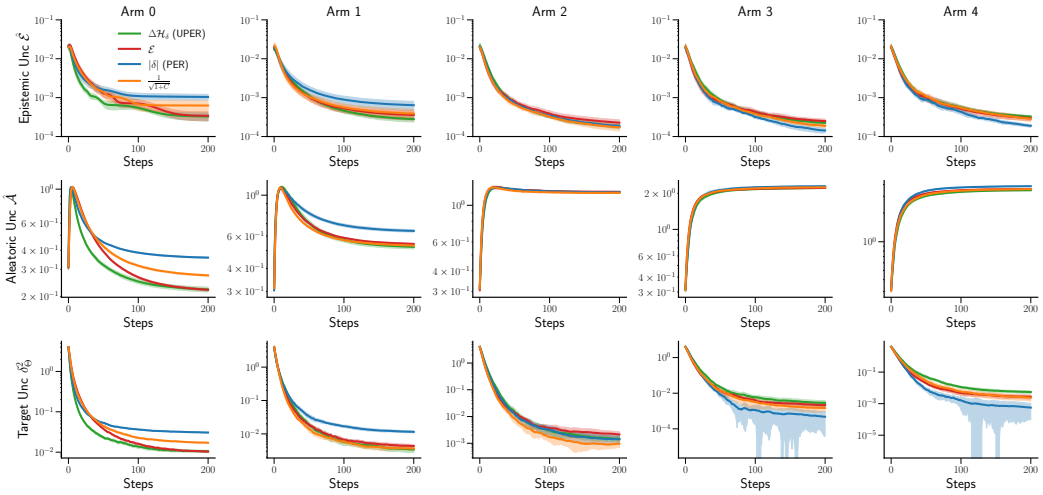

Figure 9: Epistemic uncertainty $\hat{\mathcal{E}}$ and target uncertainty δ^2_Θ decrease more rapidly for lower noise arm (first column), for UPER compared to other methods. The inclusion of aleatoric uncertainty in the prioritization variable, as utilized in the information gain formula, aims to sample transitions with high epistemic uncertainty for its reduction, while also avoiding transitions with high aleatoric uncertainty with less learnable content. This rationale is reflected in the ratio presented in the derived $\Delta\mathcal{H}_\delta$, and shown its effect in the sampling probabilities plotted in Fig. 8. The TD-error tends to oversample noisier transitions, resulting in less frequent updates for the least noisy arm, consequently leading to higher levels of epistemic and target uncertainty for that arm.

SM 6 Atari Experiments

Cumulated training improvement of UPER over PER, QRDQN, QR-PER and QR-ENS-PER are shown in Fig. 11 to Fig. 14. The accumulate percent improvement $C_{\text{UPER/PER}}$, (same for $C_{\text{UPER/QRDQN}}$ and the rest), is computed as

$$C_{\text{UPER/PER}} = \frac{\sum_t [\text{UPER}_{\text{human}}(t) - \text{PER}_{\text{human}}(t)]}{\sum_t \text{PER}_{\text{human}}(t)} \cdot 100 \tag{50}$$

where t indexes training time, and $\text{UPER}_{\text{human}}$ (same for $\text{PER}_{\text{human}}$ and $\text{QRDQN}_{\text{human}}$) denotes human normalized performance.

For the baseline experiments we use the same implementations as those of the original papers, including hyperparemeter specifications. For our UPER method, we performed a limited hyperparameter sweep over 3 key hyperparameters: learning rate and ϵ for the optimizer, and the priority exponent. The sweep ranged 3×10^{-5} to 5×10^{-5} for the learning rate, 6.1×10^{-7} to 3.125×10^{-4} for ϵ and 0.6 to 1 for the priority exponent. We chose values for our final experiments based on average performance over 2 seeds across a sub-selection of 5 Atari games (chopper command, asterix, gopher, space invaders, and battlezone).

SM 6.1 QR Models Ablation

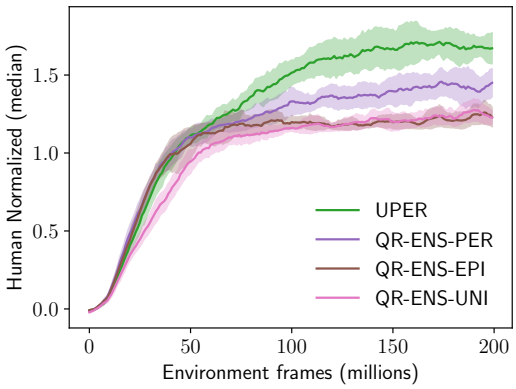

Figure 10: Comparison of ablated prioritization variables. Median Human Normalized Score for QR-DQN ensembles, where only the prioritization variable is changed. UPER, PER, EPI, and UNI use the information gain in Eq. 11, the TD-error, target epistemic uncertainty in Eq. 13, and uniform sampling, respectively.

To demonstrate the effectiveness of the information gain prioritization, and to confirm that the performance improvement stems from our proposed prioritization variable, we compared UPER to identical QR-DQN ensemble agents, maintaining the same architecture but altering only the prioritization variable. The results are presented in Fig. 10. UPER outperforms alternative approaches such as QR-DQN-PER, which uses the TD-error to prioritize (as previously shown in Fig. 2), QR-ENS-EPI, which directly prioritizes using epistemic uncertainty as defined in Eq. 13, and QR-ENS-UNI, which uses uniform sampling. These findings highlight the significance of both epistemic uncertainty and aleatoric uncertainty in prioritizing replay, as included in the information gain term. Additionally, these results confirm that the performance improvement can be solely attributed to the prioritization variable, as the QR-DQN ensemble architecture employed in each agent remains constant.

SM 6.2 Computational cost

For the main Atari-57 benchmark results, average clock time training for PER, QR-DQN, and UPER (standard DQN, distributed RL agent, and ensemble of distributed RL agents) are $\approx$ 150 hours, $\approx$ 149 hours, and $\approx$ 162 hours respectively, all implemented in JAX running in Tesla V100 NVIDIA Tensor Cores.

To generate Table 1, we conducted experiments on a laptop equipped with an i5-10500H CPU (2.50GHz) and a 6GB NVIDIA GeForce RTX 3060 Mobile/Max-Q (not the same architecture as

the main results in the paper, which uses Tesla V100 NVIDIA Tensor Cores). We ran 40 iterations of Pong for each model, using the last 20 iterations to avoid initialization and buffer filling times. The experiments were conducted on both CPU and GPU using different network architectures. In each iteration, the agent processed 1000 frames and performed one batch update of 64 transitions, with 4 frames per iteration. For all these runs, we used the publicly available implementation of DQN Zoo by DeepMind. Table 1 shows the time it takes for each iteration (1000 frames and a batch update) in seconds, along with standard deviations. There are two main conclusions from this experiment. First, most of the time consumed during each iteration is spent running the game engine (the 1000 frames per iteration), which is typically run on the CPU. This is evident from the small difference in time between QR-DQN and DQN in both the CPU and GPU cases. This difference could be larger in favor of the GPU if the batch size is increased and the frames per iteration are reduced. Second, we are significantly leveraging the parallelization capabilities of GPUs, as shown by the reduced times for the QR-DQN-ENS model (the architecture needed for UPER) when comparing GPU to CPU performance. The 2-second gap per iteration when comparing QR-DQN-ENS with QR-DQN and DQN is further reduced by utilizing V100 GPUs, as demonstrated by the training times reported in the main Atari-57 experiment.

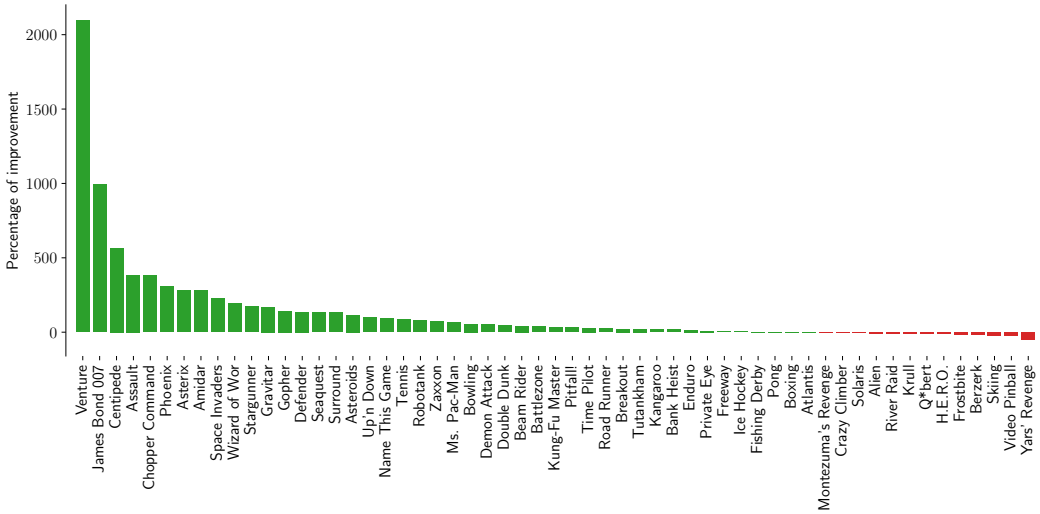

Figure 11: Cumulated training improvement of UPER over PER defined as $C_{\text{UPER}/\text{PER}}$.

SM 7 C51

To help assess whether the UPER methodology would also work when used in conjunction with other deep learning algorithms beyond QR-DQN, we performed a smaller scale set of experiments using the C51 algorithm Bellemare et al. (2017). We selected 5 Atari games in which ablations from Hessel et al. (2017) suggested vanilla PER was ineffective or even detrimental. Results on these 5 games comparing an ensemble C51 agent with PER vs an ensemble C51 agent with UPER are shown in Fig. 16. Our method is significantly better on 4 games and similar in the fifth.

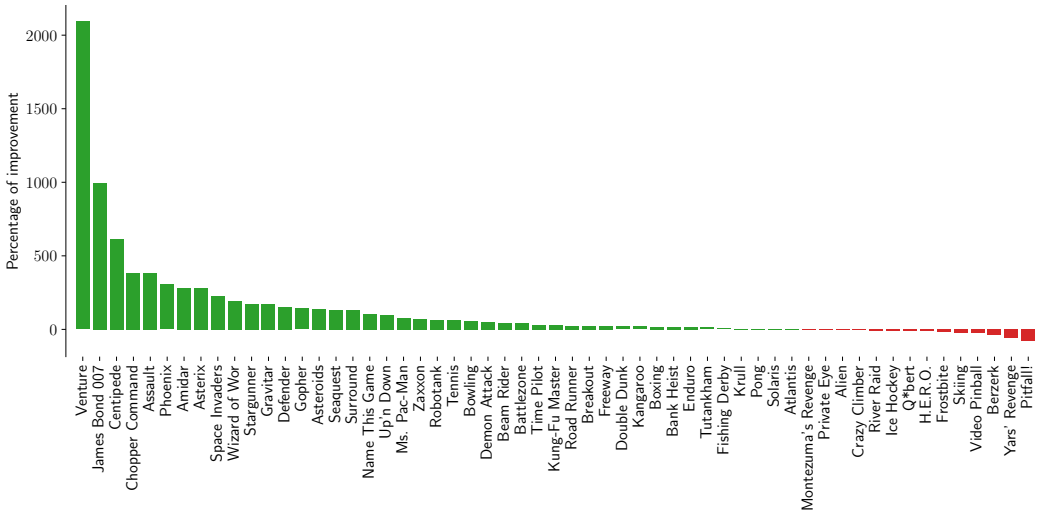

Figure 12: Cumulated training improvement of UPER over QR-DQN defined as $C_{\text{UPER/QR-DQN}}$.

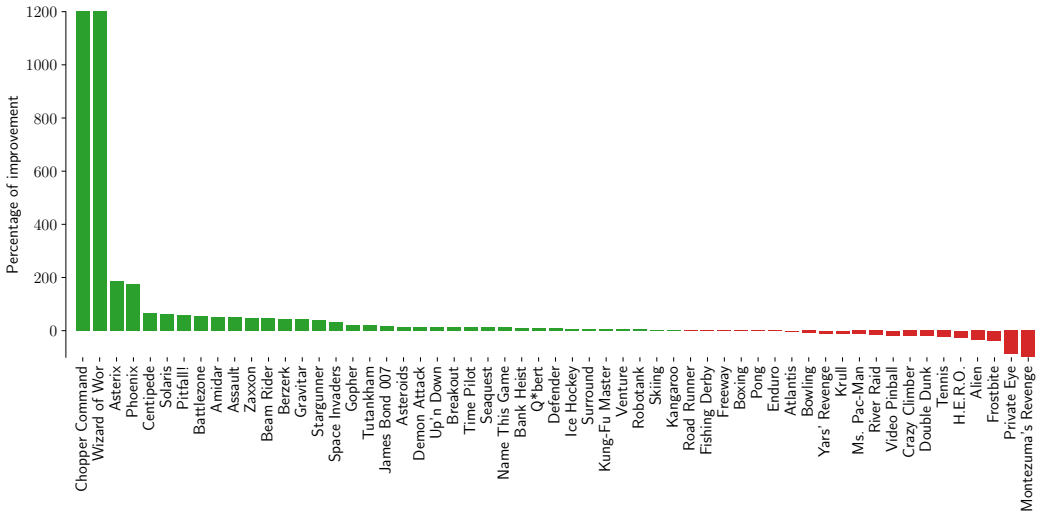

Figure 13: Cumulated training improvement of UPER over QR-PER defined as $C_{\text{UPER/QR-PER}}$.

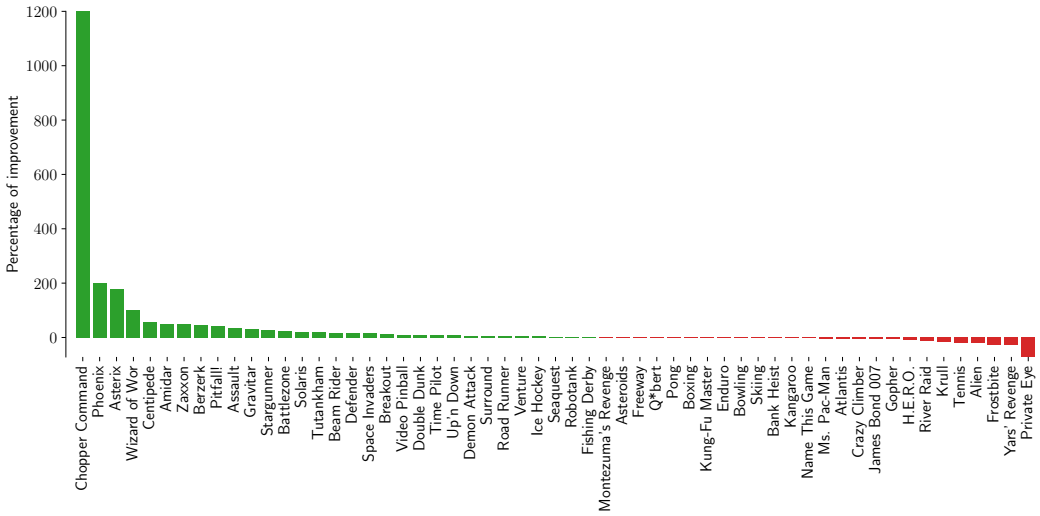

Figure 14: Cumulated training improvement of UPER over QR-ENS-PER defined as $C_{\text{UPER}/\text{QR-ENS-PER}}$.

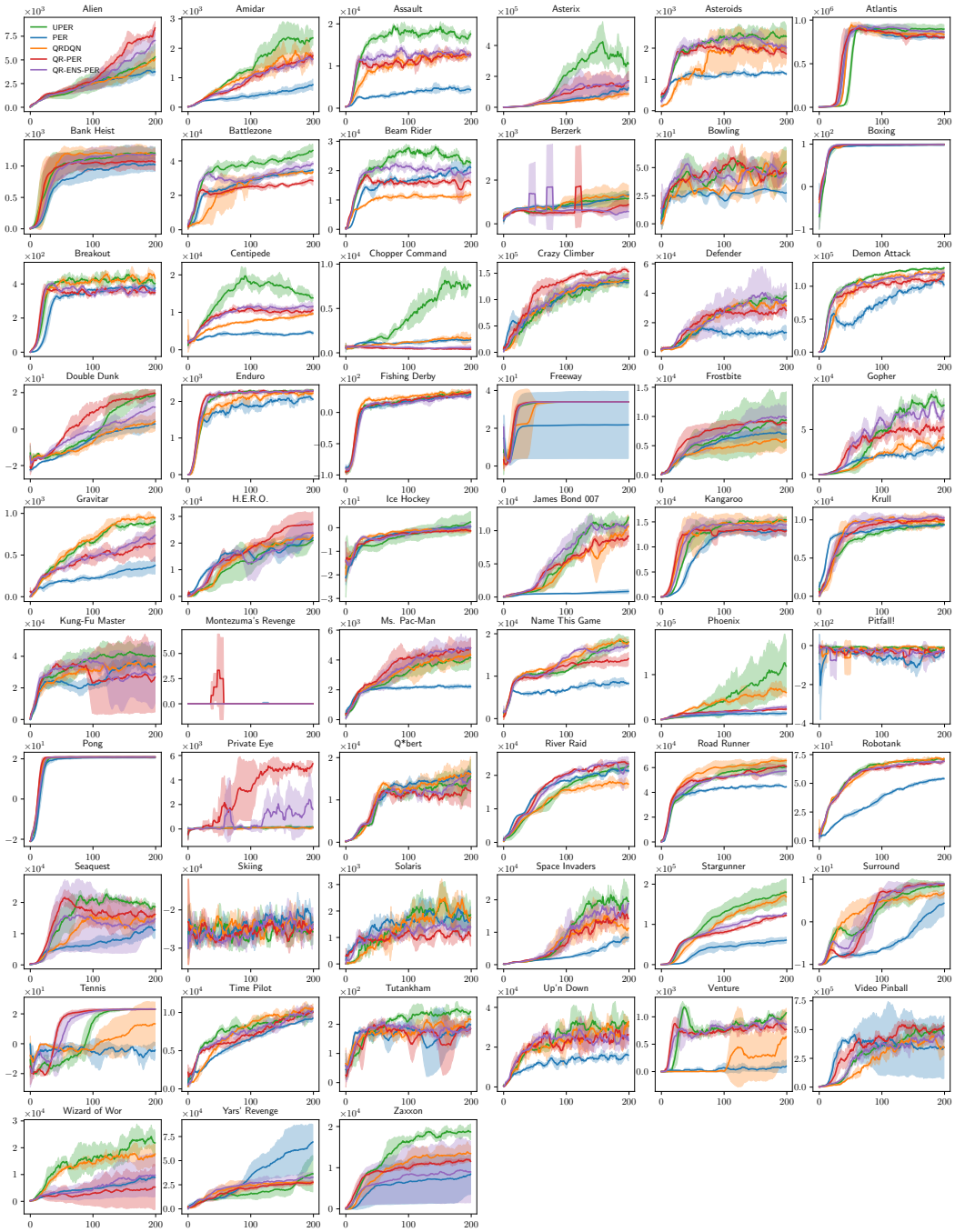

Figure 15: Average performance and corresponding standard deviation for all games across 3 seeds.

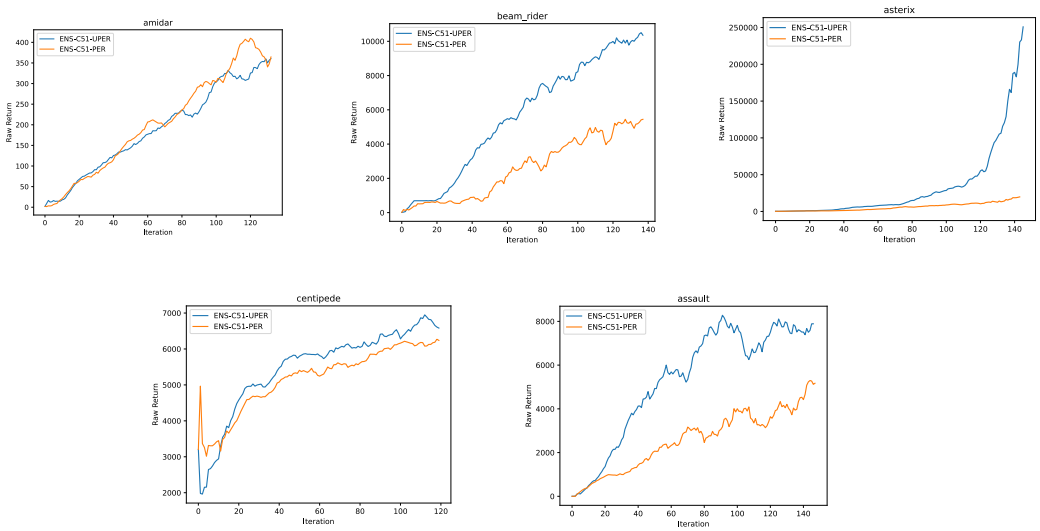

Figure 16: Performance of an ensemble C51 agent with PER vs ensemble C51 agent with PER for 5 Atari games. Average across 2 seeds.