The Uncertainty Bellman Equation and Exploration

Brendan O'Donoghue [1] **Ian Osband** [1] **Remi Munos** [1] **Volodymyr Mnih** [1]

Abstract

We consider the exploration/exploitation problem in reinforcement learning. For exploitation, it is well known that the Bellman equation connects the value at any time-step to the expected value at subsequent time-steps. In this paper we consider a similar *uncertainty* Bellman equation (UBE), which connects the uncertainty at any time-step to the expected uncertainties at subsequent time-steps, thereby extending the potential exploratory benefit of a policy beyond individual time-steps. We prove that the unique fixed point of the UBE yields an upper bound on the variance of the posterior distribution of the Q-values induced by any policy. This bound can be much tighter than traditional count-based bonuses that compound standard deviation rather than variance. Importantly, and unlike several existing approaches to optimism, this method scales naturally to large systems with complex generalization. Substituting our UBE-exploration strategy for ϵ-greedy improves DQN performance on 51 out of 57 games in the Atari suite.

1. Introduction

We consider the reinforcement learning (RL) problem of an agent interacting with its environment to maximize cumulative rewards over time (Sutton & Barto, 1998). We model the environment as a Markov decision process (MDP), but where the agent is initially uncertain of the true dynamics and mean rewards of the MDP (Bellman, 1957; Bertsekas, 2005). At each time-step, the agent performs an action, receives a reward, and moves to the next state; from these data it can learn which actions lead to higher payoffs. This leads to the *exploration versus exploitation* trade-off: Should the agent investigate poorly understood states and actions to improve future performance or instead take actions that maximize rewards given its current knowledge?

Separating estimation and control in RL via 'greedy' algorithms can lead to premature and suboptimal exploitation. To offset this, the majority of practical implementations introduce some random noise or *dithering* into their action selection (such as ϵ-greedy). These algorithms will eventually explore every reachable state and action infinitely often, but can take exponentially long to learn the optimal policy (Kakade, 2003). By contrast, for any set of prior beliefs the optimal exploration policy can be computed directly by dynamic programming in the Bayesian belief space. However, this approach can be computationally intractable for even very small problems (Guez et al., 2012) while direct computational approximations can fail spectacularly badly (Munos, 2014).

For this reason, most provably-efficient approaches to reinforcement learning rely upon the *optimism in the face of uncertainty* (OFU) principle (Lai & Robbins, 1985; Kearns & Singh, 2002; Brafman & Tennenholtz, 2002). These algorithms give a bonus to poorly-understood states and actions and subsequently follow the policy that is optimal for this augmented optimistic MDP. This optimism incentivizes exploration but, as the agent learns more about the environment, the scale of the bonus should decrease and the agent's performance should approach optimality. At a high level these approaches to OFU-RL build up confidence sets that contain the true MDP with high probability (Strehl & Littman, 2004; Lattimore & Hutter, 2012; Jaksch et al., 2010). These techniques can provide performance guarantees that are 'near-optimal' in terms of the problem parameters. However, apart from the simple 'multi-armed bandit' setting with only one state, there is still a significant gap between the upper and lower bounds for these algorithms (Lattimore, 2016; Jaksch et al., 2010; Osband & Van Roy, 2016).

One inefficiency in these algorithms is that, although the concentration may be tight at each state and action independently, the combination of simultaneously optimistic estimates may result in an extremely over-optimistic estimate for the MDP as a whole (Osband & Van Roy, 2017). Other works have suggested that a Bayesian posterior sampling approach may not suffer from these inefficiencies and can lead to performance improvements over OFU methods

[1]DeepMind. Correspondence to: Brendan O'Donoghue <bodonoghue@google.com>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

(Strens, 2000; Osband et al., 2013; Grande et al., 2014). In this paper we explore a related approach that harnesses the simple relationship of the uncertainty Bellman equation (UBE), where we define *uncertainty* to be the variance of the Bayesian posterior of the Q-values of a policy conditioned on the data the agent has collected, in a sense similar to the parametric variance of Mannor et al. (2007). Intuitively speaking, if the agent has high uncertainty (as measured by high posterior variance) in a region of the state-space then it should explore there, in order to get a better estimate of those Q-values. We show that, just as the Bellman equation relates the value of a policy beyond a single time-step, so too does the uncertainty Bellman equation propagate uncertainty values over multiple time-steps, thereby facilitating 'deep exploration' (Osband et al., 2017; Moerland et al., 2017).

The benefit of our approach (which *learns* the solution to the UBE and uses this to guide exploration) is that we can harness the existing machinery for deep reinforcement learning with minimal change to existing network architectures. The resulting algorithm shares a connection to the existing literature of both OFU and intrinsic motivation (Singh et al., 2004; Schmidhuber, 2009; White & White, 2010). Recent work has further connected these approaches through the notion of 'pseudo-count' (Bellemare et al., 2016; Ostrovski et al., 2017), a generalization of the number of visits to a state and action. Rather than adding a pseudo-count based bonus to the rewards, our work builds upon the idea that the more fundamental quantity is the uncertainty of the value function and that naively compounding count-based bonuses may lead to inefficient confidence sets (Osband & Van Roy, 2017). The key difference is that the UBE compounds the variances at each step, rather than standard deviation.

The observation that the higher moments of a value function also satisfy a form of Bellman equation is not new and has been observed by some of the early papers on the subject (Sobel, 1982). Unlike most prior work, we focus upon the *epistemic* uncertainty over the value function, as captured by the Bayesian posterior, *i.e.*, the uncertainty due to estimating a parameter using a finite amount of data, rather than the higher moments of the reward-to-go (Lattimore & Hutter, 2012; Azar et al., 2012; Mannor & Tsitsiklis, 2011; Bellemare et al., 2017). For application to rich environments with complex generalization we will use a deep learning architecture to *learn* a solution to the UBE, in the style of (Tamar et al., 2016).

2. Problem formulation

We consider a finite horizon, finite state and action space MDP, with horizon length $H \in \mathbb{N}$, state space $\mathcal{S}$, action space $\mathcal{A}$ and rewards at time period h denoted by $r^h \in \mathbb{R}$. A policy $\pi = (\pi^1, \ldots, \pi^H)$ is a sequence of functions where each $\pi^h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$ is a mapping from state-action pair to the probability of taking that action at that state, *i.e.*, π^h_{sa} is the probability of taking action a at state s at time-step h and $\sum_a \pi^h_{sa} = 1$ for all $s \in \mathcal{S}$. At each time-step h the agent receives a state s^h and a reward r^h and selects an action a^h from the policy π^h, and the agent moves to the next state s^{h+1}, which is sampled with probability $P^h_{s^{h+1}s^h a^h}$, where $P^h_{s'sa}$ is the probability of transitioning from state s to state s' after taking action a at time-step h. The goal of the agent is to maximize the expected total discounted return J under its policy π, where $J(\pi) = \mathbf{E}\left[\sum_{h=1}^H r^h \mid \pi\right]$. Here the expectation is with respect to the initial state distribution, the state-transition probabilities, the rewards, and the policy π.

The action-value, or Q-value, at time step l of a particular state under policy π is the expected total return from taking that action at that state and following π thereafter, *i.e.*, $Q^l_{sa} = \mathbf{E}\left[\sum_{h=l}^H r^h \mid s^l = s, a^l = a, \pi\right]$ (we suppress the dependence on π in this notation). The value of state s under policy π at time-step h, $V^h(s) = \mathbf{E}_{a\sim\pi^h_s} Q^h_{sa}$, is the expected total discounted return of policy π from state s.

The Bellman operator $\mathcal{T}^h$ for policy π at each time-step h relates the value at each time-step to the value at subsequent time-steps via dynamic programming (Bellman, 1957),

$$\mathcal{T}^h Q^{h+1}_{sa} = \mu^h_{sa} + \sum_{s',a'} \pi^h_{s'a'} P^h_{s'sa} Q^{h+1}_{s'a'} \qquad (1)$$

for all (s, a), where $\mu = \mathbf{E}\, r$ is the mean reward. The Q-values are the unique fixed point of equation (1), *i.e.*, the solution to $\mathcal{T}^h Q^{h+1} = Q^h$ for $h = 1, \ldots, H$, where Q^{H+1} is defined to be zero. Several reinforcement learning algorithms have been designed around minimizing the residual of equation (1) to propagate knowledge of immediate rewards to long term value (Sutton, 1988; Watkins, 1989). In the next section we examine a similar relationship for propagating the *uncertainties* of the Q-values, we call this relationship the uncertainty Bellman equation.

3. The uncertainty Bellman equation

In this section we derive a Bellman-style relationship that propagates the uncertainty (variance) of the Bayesian posterior distribution over Q-values across multiple time-steps. Propagating the potential value of exploration over many time-steps, or *deep exploration*, is important for statistically efficient RL (Kearns & Singh, 2002; Osband et al., 2017). Our main result, which we state in Theorem 1, is based upon nothing more than the dynamic programming recursion in equation (1) and some crude upper bounds of several intermediate terms. We will show that even in very simple settings this approach can result in well-calibrated

uncertainty estimates where common count-based bonuses are inefficient (Osband & Van Roy, 2017).

3.1. Posterior variance estimation

We consider the Bayesian case, where we have priors over the mean reward μ and the transition probability matrix P which we denote by ϕ_μ and ϕ_P respectively. We collect some data generated by the policy π and use it to derive posterior distributions over μ and P, given the data. We denote by $\mathcal{F}_t$ the sigma-algebra generated by all the history up to episode t (*e.g.*, all the rewards, actions, and state transitions for all episodes), and let the posteriors over the mean reward and transition probabilities be denoted by $\phi_{\mu|\mathcal{F}_t}$ and $\phi_{P|\mathcal{F}_t}$ respectively. If we sample $\hat{\mu} \sim \phi_{\mu|\mathcal{F}_t}$ and $\hat{P} \sim \phi_{P|\mathcal{F}_t}$, then the resulting Q-values that satisfy

$$\hat{Q}^h_{sa} = \hat{\mu}^h_{sa} + \sum_{s',a'} \pi^h_{s'a'} \hat{P}^h_{s'sa} \hat{Q}^{h+1}_{s'a'}, \quad h = 1, \dots, H,$$

where $\hat{Q}^{H+1} = 0$, are a sample from the implicit posterior over Q-values, conditioned on the history $\mathcal{F}_t$ (Strens, 2000). In this section we compute a bound on the variance (uncertainty) of the random variable $\hat{Q}$. For the analysis we will require some additional assumptions.

Assumption 1. *The MDP is a directed acyclic graph.*

This assumption means that the agent cannot revisit a state within the same episode, and is a common assumption in the literature (Osband et al., 2014). Note that any finite horizon MDP that doesn't satisfy this assumption can be converted into one that does by 'unrolling' the MDP so that each state s is replaced by H copies of the state, one for each step in the episode.

Assumption 2. *The mean rewards are bounded in a known interval, i.e., $\mu^h_{sa} \in [-R_{\max}, R_{\max}]$ for all (s, a).*

This assumption means we can bound the absolute value of the Q-values as $|Q^h_{sa}| \leq Q_{\max} = HR_{\max}$. We will use this quantity in the bound we derive below. This brings us to our first lemma.

Lemma 1. *For any random variable x let*

$$\mathbf{var}_t x = \mathbf{E}((x - \mathbf{E}(x|\mathcal{F}_t))^2|\mathcal{F}_t)$$

denote the variance of x conditioned on $\mathcal{F}_t$. Under the assumptions listed above, the variance of the Q-values under the posterior satisfies the Bellman inequality

$$\mathbf{var}_t \hat{Q}^h_{sa} \leq \nu^h_{sa} + \sum_{s',a'} \pi^h_{s'a'} \, \mathbf{E}(P^h_{s'sa}|\mathcal{F}_t) \mathbf{var}_t \hat{Q}^{h+1}_{s'a'}$$

for all (s, a) and $h = 1, \dots, H$, where $\mathbf{var}_t \hat{Q}^{H+1} = 0$ and where we call ν^h_{sa} the local uncertainty *at (s, a), and it is given by*

$$\nu^h_{sa} = \mathbf{var}_t \hat{\mu}^h_{sa} + Q^2_{\max} \textstyle\sum_{s'} \mathbf{var}_t \hat{P}^h_{s'sa}.$$

Proof. See the appendix. □

We refer to ν in the above lemma as the *local* uncertainty since it depends only on locally available quantities, and so can be calculated (in principle) at each state-action independently. With this lemma we are ready to prove our main theorem.

Theorem 1 (Solution of the uncertainty Bellman equation)**.** *Under assumptions 1 and 2, for any policy π there exists a unique u that satisfies the uncertainty Bellman equation*

$$u^h_{sa} = \nu^h_{sa} + \sum_{s',a'} \pi^h_{s'a'} \, \mathbf{E}(P^h_{s'sa}|\mathcal{F}_t) u^{h+1}_{s'a'} \tag{2}$$

for all (s, a) and $h = 1, \dots, H$, where $u^{H+1} = 0$, and furthermore $u \geq \mathbf{var}_t \hat{Q}$ pointwise.

Proof. Let $\mathcal{U}^h$ be the Bellman operator that defines the uncertainty Bellman equation, *i.e.*, rewrite equation (2) as

$$u^h = \mathcal{U}^h u^{h+1},$$

then to prove the result we use two essential properties of the Bellman operator for a fixed policy. Firstly, the solution to the Bellman equation exists and is unique, and secondly the Bellman operator is monotonically non-decreasing in its argument, *i.e.*, if $x \geq y$ pointwise then $\mathcal{U}^h x \geq \mathcal{U}^h y$ pointwise (Bertsekas, 2005). The proof proceeds by induction; assume that for some h we have $\mathbf{var}_t \hat{Q}^{h+1} \leq u^{h+1}$, then we have

$$\mathbf{var}_t \hat{Q}^h \leq \mathcal{U}^h \mathbf{var}_t \hat{Q}^{h+1} \leq \mathcal{U}^h u^{h+1} = u^h,$$

where we have used the fact that the variance satisfies the Bellman *inequality* from lemma 1, and the base case holds because $\mathbf{var}_t \hat{Q}^{H+1} = u^{H+1} = 0$. □

We conclude with a brief discussion on why the variance of the posterior is useful for exploration. If we had access to the true posterior distribution over the Q-values then we could take actions that lead to states with higher uncertainty by, for example, using Thompson sampling (Thompson, 1933; Strens, 2000), or constructing Q-values that are high probability upper bounds on the true Q-values and using the OFU principle (Kaufmann et al., 2012). However, calculating the true posterior is intractable for all but very small problems. Due to this difficulty prior work has sought to approximate the posterior distribution (Osband et al., 2017), and use that to drive exploration. In that spirit we develop another approximation of the posterior, in this case it is motivated by the Bayesian central limit theorem which states that, under some mild conditions, the posterior distribution converges to a Gaussian as the amount of data increases (Berger, 2013). With that in

mind, rather than computing the full posterior we approximate it as $\mathcal{N}(\bar{Q}, \mathbf{diag}(u))$ where u is the solution to the uncertainty Bellman equation (2), and consequently is a guaranteed upper bound on the true variance of the posterior, and $\bar{Q}$ denotes the mean Q-values under the posterior at episode t, *i.e.*, the unique solution to

$$\bar{Q}^h_{sa} = \mathbf{E}(\hat{\mu}^h_{sa}|\mathcal{F}_t) + \sum_{s',a'} \pi^h_{s'a'} \mathbf{E}(\hat{P}^h_{s'sa}|\mathcal{F}_t)\bar{Q}^{h+1}_{s'a'},$$

for $h = 1, \ldots, H$, and $\bar{Q}^{H+1} = 0$. With this approximate posterior we can perform Thompson sampling as an exploration heuristic. Specifically, at state s and time-step h we select the action using

$$a = \operatorname*{argmax}_b (\bar{Q}^h_{sb} + \zeta_b (u^h_{sb})^{1/2}) \tag{3}$$

where ζ_b is sampled from $\mathcal{N}(0, 1)$. Our goal is for the agent to explore states and actions where it has higher uncertainty. This is in contrast to the commonly used ϵ-greedy (Mnih et al., 2013) and Boltzmann exploration strategies (Mnih et al., 2016; O'Donoghue et al., 2017; Haarnoja et al., 2017) which simply inject noise into the agents actions. We shall see in the experiments that our strategy can dramatically outperform these naive heuristics.

3.2. Comparison to traditional exploration bonus

Consider a simple decision problem with known deterministic transitions, unknown rewards, and two actions at a root node, as depicted in Figure 1. The first action leads to a single reward r_1 sampled from $\mathcal{N}(\mu_1, \sigma^2)$ at which point the episode terminates, and the second action leads to an chain of length H consisting of states each having random reward r_2 independently sampled from $\mathcal{N}(\mu_2/H, \sigma^2/H)$.

Take the case where each action at the root has been taken n times and where the uncertainty over the rewards at each state concentrates like $1/n$ (*e.g.*, when the prior is an improper Gaussian). In this case the *true* uncertainty about the value of each action is identical and given by σ^2/n. This is also the answer we get from the uncertainty Bellman equation, since for action 1 we obtain $u_1 = \sigma^2/n$ (since $\mathbf{var}_t\, P = 0$) and for action 2 the uncertainty about the reward at each state along the chain is given by σ^2/Hn and so we have $u_2 = \sum_{h=1}^{H} \sigma^2/Hn = \sigma^2/n$.

Rather than considering the variance of the value as a whole, the majority of existing approaches to OFU provide exploration bonuses at each state and action independently and then combine these estimates via union bound. In this context, even a state of the art algorithm such as UCRL2 (Jaksch et al., 2010) would augment the rewards at each state with a bonus proportional to the standard deviation of the reward estimate at each state (Bellemare et al., 2016). For the first action this would be $\mathrm{ExpBonus}_1 = \sigma/\sqrt{n}$, but for the second action this would be accumulated along the chain to be

$$\mathrm{ExpBonus}_2 = \sum_{h=1}^{H} \frac{\sigma}{\sqrt{Hn}} = \sigma\sqrt{\frac{H}{n}}$$

In other words, the bonus afforded to the second action is a factor of $\sqrt{H}$ larger than the true uncertainty. The agent would have to take the second action a factor of H more times than the first action in order to have the same effective bonus given to each one. If the first action was actually superior in terms of expected reward, it would take the agent far longer to discover that than an agent using the correct uncertainties to select actions. The essential issue is that, unlike the variance, the standard deviations do not obey a Bellman-style relationship.

In Figure 2 we show the results of an experiment showing this phenomenon. Action 1 had expected reward $\mu_1 = 1$, and action 2 had expected reward $\mu_2 = 0$. We set $\sigma = 1$ and $H = 10$, and the results are averaged over 500 seeds. We compare two agents, one using the uncertainty Bellman equation to drive exploration and the other agent using a count-based reward bonus. Both agents take actions and use the results to update their beliefs about the value of each action. The agent using the UBE takes the action yielded by Thompson sampling as in equation (3). The exploration-bonus agent takes the action i that maximizes $\hat{Q}_i + \beta \log(t) \mathrm{ExpBonus}_i$ (the $\log(t)$ term is required to achieve a regret bound (Jaksch et al., 2010), but doesn't materially affect the previous argument) where $\beta > 0$ is a hyper-parameter chosen by a sweep and where $\hat{Q}_i$ is the estimate of the value of action i. Figure 2 shows the *regret* of each agent vs number of episodes. Regret measures how sub-optimal the rewards the agent has received so far are, relative to the (unknown) optimal policy, and lower regret is better (Cesa-Bianchi & Lugosi, 2006).

The agent using the uncertainty Bellman equation has well calibrated uncertainty estimates and consequently quickly figures out that the first action is better. By contrast, the exploration bonus agent takes significantly longer to determine that the first action is better due to the fact that the bonus afforded to the second action is too large, and consequently it suffers significantly higher regret.

4. Estimating the local uncertainty

Section 3 outlined how the uncertainty Bellman equation can be used to propagate local estimates of the variance of $\hat{Q}$ to global estimates for the uncertainty. In this section we present some pragmatic approaches to estimating the local uncertainty ν that we can then use for practical learning algorithms inspired by Theorem 1. We do not claim that these approaches are the only approaches to estimating the local uncertainty, or even that these simple approximations

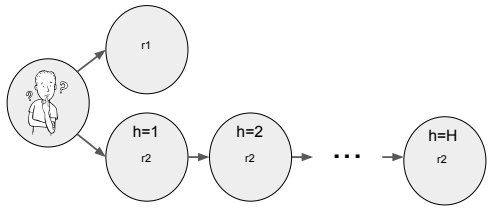

Figure 1: Simple tabular MDP.

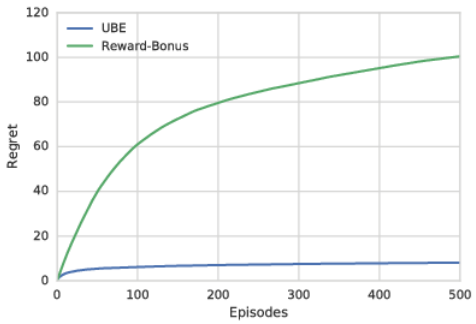

Figure 2: Regret over time for the simple tabular MDP.

are in any sense the 'best'. Investigating these choices is an important area of future research, but outside the scope of this short paper. We present a simple progression from tabular representations, to linear function approximation and then to non-linear neural network architectures.

Tabular value estimate. Consider the case where the posterior over the mean rewards concentrates at least as fast the reciprocal of the visit count, *i.e.*,

$$\mathbf{var}_t \hat{\mu}^h_{sa} \leq \sigma_r^2 / n^h_{sa}$$

where σ_r is the variance of the reward process and n^h_{sa} is the *visit count* of the agent to state s and action a at time-step h, up to episode t. This is the case when, for example, the rewards and the prior over the mean reward are both Gaussian. Furthermore, if we assume that the prior over the transition function is Dirichlet then it is straightforward to show that

$$\sum_{s'} \mathbf{var}_t \hat{P}^h_{s'sa} \leq 1/n^h_{sa}$$

since the likelihood of the transition function is a categorical distribution, which is conjugate to the Dirichlet distribution and the variance of a Dirichlet concentrates like the reciprocal of the sum of the counts of each category. Under these assumptions we can bound the local uncertainty as

$$\nu^h_{sa} \leq (\sigma_r^2 + Q^2_{\max})/n^h_{sa}.$$

In other words, the local uncertainty can be modeled under these assumptions as a constant divided by the visit count.

Linear value estimate. In the non-tabular case we need some way to estimate the inverse counts in order to approximate the local uncertainty. Consider a linear value function estimator $\hat{Q}^h_{sa} = \phi(s)^T w_a$ for each state and action with fixed basis functions $\phi(s) : \mathcal{S} \to \mathbb{R}^D$ and learned weights $w_a \in \mathbb{R}^D$, one for each action. This setting allows for some generalization between states and actions through the basis functions. For any fixed dataset we can find the least squares solution for each action a (Boyan, 1999),

$$\text{minimize}_{w_a} \sum_{i=1}^N (\phi(s_i)^T w_a - y_i)_2^2,$$

where each $y_i \in \mathbb{R}$ is a regression target (*e.g.*, a Monte Carlo return from that state-action). The solution to this problem is $w^\star_a = (\Phi_a^T \Phi_a)^{-1} \Phi_a^T y$, where Φ_a is the matrix consisting of the $\phi(s_i)$ vectors stacked row-wise (we use the subscript a to denote the fact that action a was taken at these states). We can compute the variance of this estimator, which will provide a proxy for the inverse counts (Bellemare et al., 2016). If we model the targets y_i as IID with unit variance, then $\mathbf{var}_t\, w^\star_a = (\Phi_a^T \Phi_a)^{-1}$. Given a new state vector ϕ_s, the variance of the Q-value estimate at (s, a) is then $\mathbf{var}_t \phi_s^T w^\star_a = \phi_s^T (\Phi_a^T \Phi_a)^{-1} \phi_s$, which we can take to be our estimate of the inverse counts, *i.e.*, set $(\hat{n}^h_{sa})^{-1} = \phi_s^T (\Phi_a^T \Phi_a)^{-1} \phi_s$. Now we can estimate the local uncertainty as

$$\hat{\nu}^h_{sa} = \beta^2 (\hat{n}^h_{sa})^{-1} = \beta^2 \phi_s^T (\Phi_a^T \Phi_a)^{-1} \phi_s \tag{4}$$

for some β, which in the tabular case (*i.e.*, where $\phi(s) = e_s$ and $D = |\mathcal{S}|$) is equal to β^2 / n^h_{sa}, as expected.

An agent using this notion of uncertainty must maintain and update the matrix $\Sigma_a = (\Phi_a^T \Phi_a)^{-1}$ as it receives new data. Given new sample ϕ, the updated matrix Σ_a^+ is given by

$$\begin{aligned} \Sigma_a^+ &= \left(\begin{bmatrix} \Phi_a \\ \phi^T \end{bmatrix}^T \begin{bmatrix} \Phi_a \\ \phi^T \end{bmatrix} \right)^{-1} = (\Phi_a^T \Phi_a + \phi \phi^T)^{-1} \\ &= \Sigma_a - (\Sigma_a \phi \phi^T \Sigma_a)/(1 + \phi^T \Sigma_a \phi) \end{aligned} \tag{5}$$

by the Sherman-Morrison-Woodbury formula (Golub & Van Loan, 2012), the cost of this update is one matrix multiply and one matrix-matrix subtraction per step.

Neural networks value estimate. If we are approximating our Q-value function using a neural network then the above analysis does not hold. However if the last layer of the network is linear, then the Q-values are approximated as $Q^h_{sa} = \phi(s)^T w_a$, where w_a are the weights of the last layer associated with action a and $\phi(s)$ is the output of the network up to the last layer for state s. In other words we can

think of a neural network as learning a useful set of basis functions such that a linear combination of them approximates the Q-values. Then, if we ignore the uncertainty in the ϕ mapping, we can reuse the analysis for the purely linear case to derive an *approximate* measure of local uncertainty that might be useful in practice.

This scheme has some advantages. As the agent progresses it is learning a state representation that helps it achieve the goal of maximizing the return. The agent will learn to pay attention to small but important details (*e.g.*, the ball in Atari 'breakout') and learn to ignore large but irrelevant changes (*e.g.*, if the background suddenly changes). This is a desirable property from the point of view of using these features to drive exploration, because the states that differ only in irrelevant ways will be aliased to (roughly) the same state representation, and states that differ is small but important ways will be mapped to quite different state vectors, permitting a more task-relevant measure of uncertainty.

5. Deep Reinforcement Learning

Previously we proved that under certain conditions we can bound the variance of the posterior distribution of the Q-values, and we used the resulting uncertainty values to derive an exploration strategy. Here we discuss the application of that strategy to deep-RL. In this case several of the assumptions we have made to derive theorem 1 are violated. This puts us firmly in the territory of heuristic. Specifically, the MDPs we apply this to will not be directed acyclic graphs, the policy that we are estimating the uncertainty over will not be fixed, we cannot exactly compute the local uncertainty, and we won't be solving the UBE exactly. However, empirically, we demonstrate that this heuristic can perform well in practice, despite the underlying assumptions being violated.

Our strategy involves *learning* the uncertainty estimates, and then using them to sample Q-values from the approximate posterior, as in equation (3). The technique is described in pseudo-code in Algorithm 1. We refer to the technique as 'one-step' since the uncertainty values are updated using a one-step SARSA Bellman backup, but it is easily extendable to the n-step case. The algorithm takes as input a neural network which has two output 'heads', one which is attempting to learn the optimal Q-values as normal, the other is attempting to learn the uncertainty values of the current policy (which is constantly changing). We do not allow the gradients from the uncertainty output head to flow into the trunk of the network; this ensures the Q-value estimates are not perturbed by the changing uncertainty signal. For the local uncertainty measure we use the linear basis approximation described in section 4. Algorithm 1 incorporates a discount factor $\gamma \in (0, 1)$, since deep RL often uses a discount even in the purely episodic case. In this case the Q-learning update uses a γ discount and the Uncertainty Bellman equation (2) is augmented with a γ^2 discount factor.

Algorithm 1 One-step UBE exploration with linear uncertainty estimates.

Require: Neural network outputting Q and u estimates, Q-value learning subroutine `qlearn`, Thompson sampling hyper-parameter $\beta > 0$
Initialize $\Sigma_a = \mu I$ for each a, where $\mu > 0$
Get initial state s, take initial action a
for episode $t = 1, \ldots,$ **do**
  for time-step $h = 2, \ldots, H + 1$ **do**
    Retrieve $\phi(s)$ from input to last network layer
    Receive new state s' and reward r
    Calculate $\hat{Q}^h_{s'b}$ and $u^h_{s'b}$ for each action b
    Sample $\zeta_b \sim \mathcal{N}(0, 1)$ for each b and calculate

$$a' = \operatorname*{argmax}_b (\hat{Q}^h_{s'b} + \beta \zeta_b (u^h_{s'b})^{1/2})$$

    Calculate

$$y = \begin{cases} \phi(s)^T \Sigma_a \phi(s), \text{ if } h = H + 1 \\ \phi(s)^T \Sigma_a \phi(s) + \gamma^2 u^h_{s'a'}, \text{ o.w.} \end{cases}$$

    Take gradient step with respect to error

$$(y - u^{h-1}_{sa})^2$$

    Update Q-values using $\texttt{qlearn}(s, a, r, s', a')$
    Update Σ_a according to eq. (5)
    Take action a'
    Set $s \leftarrow s'$, $a \leftarrow a'$
  end for
end for

5.1. Experimental results

Here we present results of Algorithm (1) on the Atari suite of games (Bellemare et al., 2012), where the network is attempting to learn the Q-values as in DQN (Mnih et al., 2013; 2015) and the uncertainties simultaneously. The only change to vanilla DQN we made was to replace the ϵ-greedy policy with Thompson sampling over the learned uncertainty values, where the β constant in (3) was chosen to be 0.01 for all games, by a parameter sweep. We used the exact same network architecture, learning rate, optimizer, pre-processing and replay scheme as described in Mnih et al. (2015). For the uncertainty sub-network we used a single fully connected hidden layer with 512 hidden units followed by the output layer. We trained the uncertainty head using a separate RMSProp optimizer (Tieleman & Hinton, 2012) with learning rate 10^{-3}. The addition of the uncertainty head and the computation associated with

it, only reduced the frame-rate compared to vanilla DQN by about 10% on a GPU, so the additional computational cost of the approach is negligible.

We compare two versions of our approach: a 1-step method and an n-step method where we set n to 150. The n-step method accumulates the uncertainty signal over n time-steps before performing an update which should lead to the uncertainty signal propagating to earlier encountered states faster, at the expense of increased variance of the signal. Note that in all cases the Q-learning update is always 1-step; our n-step implementation only affects the uncertainty update.

We compare our approaches to vanilla DQN, and also to an exploration bonus intrinsic motivation approach, where the agent receives an augmented reward consisting of the extrinsic reward and the square root of the linear uncertainty in equation (4), which was scaled by a hyper-parameter chosen to be 0.1 by a sweep. In this case a stochastic policy was still required for good performance and so we used ϵ-greedy with the DQN annealing schedule.

We trained all strategies for 200M frames (about 8 days on a GPU). Each game and strategy was tested three times per method with the same hyper-parameters but with different random seeds, and all plots and scores correspond to an average over the seeds. All scores were normalized by subtracting the average score achieved by an agent that takes actions uniformly at random. Every 1M frames the agents were saved and evaluated (without learning) on 0.5M frames, where each episode is started from the random start condition described in (Mnih et al., 2015). The final scores presented correspond to first averaging the evalution score in each period across seeds, then taking the max average episodic score observed during any evalution period. Of the tested strategies the n-step UBE approach was the highest performer in 32 out of 57 games, the 1-step UBE approach in 14 games, DQN in 1 game, the exploration bonus strategy in 7 games, and there were 3 ties. In Table 1 we give the mean and median normalized scores as percentage of an expert human normalized score across all games, and the number of games where the agent is 'super-human', for each tested algorithm. Note that the mean scores are significantly affected by a single outlier with very high score ('Atlantis'), and therefore the median score is a better indicator of agent performance. In Figure 3 we plot the number of games at super-human performance against frames for each method, and in Figure 4 we plot the median performance across all games versus frames, where a score of 1.0 denotes human performance. The results across all 57 games, as well as the learning curves for all 57 games, are given in the appendix.

Of particular interest is the game 'Montezuma's Revenge', a notoriously difficult exploration game where no one-step algorithm has managed to learn anything useful. Our 1-step strategy learns in 200M frames a policy that is able to consistently get about 500 points, which is the score the agent gets for picking up the first key and moving into the second room. In Figure 5 we show the learning progress of the agents for 500M frames where we set the Thompson sampling parameter slightly higher; 0.016 instead of 0.01 (since this game is a challenging exploration task it stands to reason that a higher exploration parameter is required). By the end of 500M frames the n-step agent is consistently getting around 3000 points, which is several rooms of progress. These scores are close to state-of-the-art, and are state-of-the-art for one-step methods (like DQN) to the best of our knowledge.

In the recent work by Bellemare et al. (2016), and the follow-up work by Ostrovski et al. (2017), the authors add an intrinsic motivation signal to a DQN-style agent that has been modified to use the full Monte Carlo return of the episode when learning the Q-values. Using Monte Carlo returns dramatically improves the performance of DQN in a way unrelated to exploration, and due to that change we cannot compare the numerical results directly. In order to have a point of comparison we implemented our own intrinisic motivation exploration signal, as discussed above. Similarly, we cannot compare directly to the numerical results obtained by Bootstrap DQN (Osband et al., 2016) since that agent is using Double-DQN, a variant of DQN that achieves a higher performance in a way unrelated to exploration. However, we note that our approach achieves a higher evaluation score in 27 out of the 48 games tested in the Bootstrap DQN paper despite using an inferior base DQN implementation, and it runs at a significantly lower computational and memory cost.

	mean	**median**	**> human**
DQN	688.60	79.41	21
DQN Intrinsic Motivation	472.93	76.73	24
DQN UBE 1-step	776.40	94.54	26
DQN UBE n-step	439.88	126.41	35

Table 1: Scores for the Atari suite, as a percentage of human score.

6. Conclusion

In this paper we derived a Bellman equation for the uncertainty over the Q-values of a policy. This allows an agent to propagate uncertainty across many time-steps in the same way that value propagates through time in the standard dynamic programming recursion. This uncertainty can be used by the agent to make decisions about which states and actions to explore, in order to gather more data about the environment and learn a better policy. Since the uncertainty satisfies a Bellman recursion, the agent can learn it using

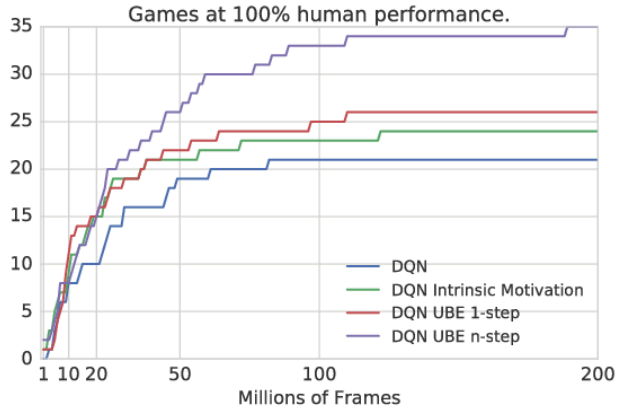

Figure 3: Number of games at super-human performance.

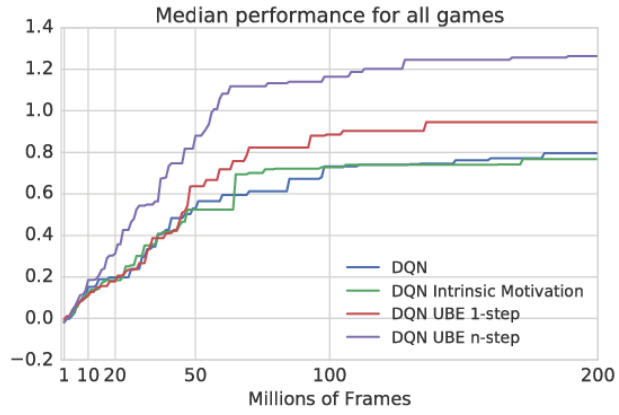

Figure 4: Normalized median performance across all games, a score of 1.0 is human-level performance.

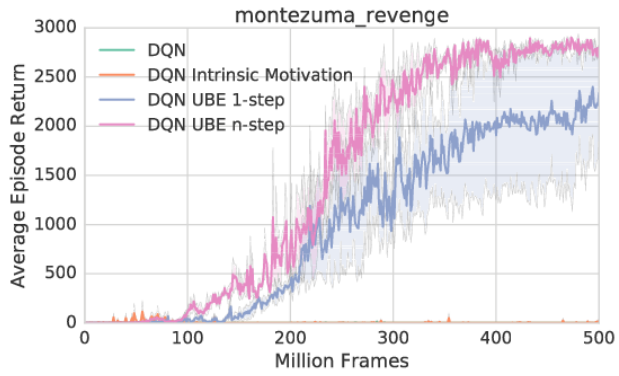

Figure 5: Montezuma's Revenge performance.

the same reinforcement learning machinery that has been developed for value functions. We showed that a heuristic algorithm based on this learned uncertainty can boost the performance of standard deep-RL techniques. Our technique was able to significantly improve the performance of DQN across the Atari suite of games, when compared against naive strategies like ϵ-greedy.

7. Acknowledgments

We thank Marc Bellemare, David Silver, Koray Kavukcuoglu, and Mohammad Gheshlaghi Azar for useful discussion and suggestions on the paper.

References

Azar, M. G., Munos, R., and Kappen, B. On the sample complexity of reinforcement learning with a generative model. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.

Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., and Munos, R. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, pp. 1471–1479, 2016.

Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 2012.

Bellemare, M. G., Dabney, W., and Munos, R. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pp. 449–458, 2017.

Bellman, R. *Dynamic programming*. Princeton University Press, 1957.

Berger, J. O. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

Bertsekas, D. P. *Dynamic programming and optimal control*, volume 1. Athena Scientific, 2005.

Boyan, J. A. Least-squares temporal difference learning. In *ICML*, pp. 49–56, 1999.

Brafman, R. I. and Tennenholtz, M. R-max: A general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Golub, G. H. and Van Loan, C. F. *Matrix computations*, volume 3. JHU Press, 2012.

Grande, R., Walsh, T., and How, J. Sample efficient reinforcement learning with Gaussian processes. In *International Conference on Machine Learning*, pp. 1332–1340, 2014.

Guez, A., Silver, D., and Dayan, P. Efficient Bayes-adaptive reinforcement learning using sample-based search. In *Advances in Neural Information Processing Systems*, pp. 1025–1033, 2012.

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

Kakade, S. M. *On the sample complexity of reinforcement learning*. PhD thesis, University of London London, England, 2003.

Kaufmann, E., Cappé, O., and Garivier, A. On Bayesian upper confidence bounds for bandit problems. In *Artificial Intelligence and Statistics*, pp. 592–600, 2012.

Kearns, M. and Singh, S. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2-3):209–232, 2002.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Lattimore, T. Regret analysis of the anytime optimally confident ucb algorithm. *arXiv preprint arXiv:1603.08661*, 2016.

Lattimore, T. and Hutter, M. Pac bounds for discounted mdps. In *International Conference on Algorithmic Learning Theory*, pp. 320–334. Springer, 2012.

Mannor, S. and Tsitsiklis, J. Mean-variance optimization in Markov decision processes. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 177–184, 2011.

Mannor, S., Simester, D., Sun, P., and Tsitsiklis, J. N. Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322, 2007.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. In *NIPS Deep Learning Workshop*. 2013.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. Human-level control through deep reinforcement learning. *Nature*, 518 (7540):529–533, 02 2015. URL `http://dx.doi.org/10.1038/nature14236`.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 1928–1937, 2016.

Moerland, T. M., Broekens, J., and Jonker, C. M. Efficient exploration with double uncertain value networks. In *31st Conference on Neural Information Processing Systems (NIPS)*, 2017.

Munos, R. From bandits to monte-carlo tree search: The optimistic principle applied to optimization and planning. *Foundations and Trends® in Machine Learning*, 7(1):1–129, 2014.

O'Donoghue, B., Munos, R., Kavukcuoglu, K., and Mnih, V. Combining policy gradient and Q-learning. In *International Conference on Learning Representations (ICLR)*, 2017.

Osband, I. and Van Roy, B. On lower bounds for regret in reinforcement learning. *stat*, 1050:9, 2016.

Osband, I. and Van Roy, B. Why is posterior sampling better than optimism for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Osband, I., Russo, D., and Van Roy, B. (More) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pp. 3003–3011, 2013.

Osband, I., Van Roy, B., and Wen, Z. Generalization and exploration via randomized value functions. *arXiv preprint arXiv:1402.0635*, 2014.

Osband, I., Blundell, C., Pritzel, A., and Van Roy, B. Deep exploration via bootstrapped DQN. In *Advances In Neural Information Processing Systems*, pp. 4026–4034, 2016.

Osband, I., Russo, D., Wen, Z., and Van Roy, B. Deep exploration via randomized value functions. *arXiv preprint arXiv:1703.07608*, 2017.

Ostrovski, G., Bellemare, M. G., Oord, A. v. d., and Munos, R. Count-based exploration with neural density models. *arXiv preprint arXiv:1703.01310*, 2017.

Schmidhuber, J. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In *Anticipatory Behavior in Adaptive Learning Systems*, pp. 48–76. Springer, 2009.

Singh, S. P., Barto, A. G., and Chentanez, N. Intrinsically motivated reinforcement learning. In *NIPS*, volume 17, pp. 1281–1288, 2004.

Sobel, M. J. The variance of discounted Markov decision processes. *Journal of Applied Probability*, 19(04):794–802, 1982.

Strehl, A. and Littman, M. Exploration via modelbased interval estimation, 2004.

Strens, M. A Bayesian framework for reinforcement learning. In *ICML*, pp. 943–950, 2000.

Sutton, R. and Barto, A. *Reinforcement Learning: an Introduction*. MIT Press, 1998.

Sutton, R. S. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

Tamar, A., Di Castro, D., and Mannor, S. Learning the variance of the reward-to-go. *Journal of Machine Learning Research*, 17(13):1–36, 2016.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Tieleman, T. and Hinton, G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4, 2012.

Watkins, C. J. C. H. *Learning from delayed rewards*. PhD thesis, University of Cambridge England, 1989.

White, M. and White, A. Interval estimation for reinforcement-learning algorithms in continuous-state domains. In *Advances in Neural Information Processing Systems*, pp. 2433–2441, 2010.