Optimistic Exploration with Backward Bootstrapped Bonus for Deep Reinforcement Learning

Chenjia Bai [1] **Lingxiao Wang** [2] **Lei Han** [3] **Jianye Hao** [4] **Animesh Garg** [5] **Peng Liu** [1] **Zhaoran Wang** [2]

Abstract

Optimism in the face of uncertainty is a principled approach for provably efficient exploration for reinforcement learning in tabular and linear settings. However, such an approach is challenging in developing practical exploration algorithms for Deep Reinforcement Learning (DRL). To address this problem, we propose an Optimistic Exploration algorithm with Backward Bootstrapped Bonus (OEB3) for DRL. We construct an UCB-bonus indicating the uncertainty of Q-functions. The UCB-bonus is further utilized to estimate an optimistic Q-value, which encourages the agent to explore the scarcely visited states and actions to reduce uncertainty. In the estimation of Q-function, we adopt an episodic backward update strategy to propagate the future uncertainty to the estimated Q-function consistently. Experiments show that OEB3 outperforms several state-of-the-art exploration approaches 49 Atari games.

1. Introduction

In Reinforcement learning (RL) (Sutton & Barto, 2018), an agent aims to maximize the long-term return by interacting with an unknown environment. To find the optimal policy, the agent is required to sufficiently explore the unknown environment and exploit in depth along the optimal trajectory. Devising efficient exploration algorithms thus becomes an attractive topic in recent years of RL research. The theoretical achievements in RL offer various provably efficient exploration methods in tabular and linear Markov Decision Processes (MDPs) based on the fundamental value iteration algorithm Least-Squares Value Iteration (LSVI). Among these, *optimism in the face of uncertainty* (Auer & Ortner, 2007; Jin et al., 2018) is a principled approach for efficient exploration with well theoretical guarantees. In tabular cases, the optimism-based methods incorporate the Upper Confidence Bound (UCB) into the value function as bonus and attain the optimal worst-case regret (Azar et al., 2017; Jaksch et al., 2010; Dann & Brunskill, 2015). Randomized value function based on posterior sampling chooses actions according to the randomly sampled statistically plausible value function and is known to achieve near-optimal worst-case and Bayesian regrets (Osband & Van Roy, 2017; Russo, 2019). Recently, the theoretical analyses in tabular cases have been extended to linear MDPs where the transition and reward function are assumed to be linear. In linear cases, LSVI-UCB (Jin et al., 2020) has been demonstrated to enjoy a near-optimal worst-case regret using a provably efficient bonus. Randomized LSVI (Zanette et al., 2020) also obtains a near-optimal worst-case regret.

Although the analyses in tabular and linear cases have induced attractive approaches for efficient exploration, it is still challenging in developing a practical exploration algorithm that is essentially suitable for Deep Reinforcement Learning (DRL) (Mnih et al., 2015), which is necessary to achieve human-level performance in large-scale tasks such as Atari games and robotic tasks. A simple evidence is that, in linear case, the bonus in LSVI-UCB (Jin et al., 2020) and nontrivial noise in randomized LSVI (Zanette et al., 2020) are specifically designed for linear models (Abbasi-Yadkori et al., 2011), without generalizations to fit powerful function approximations such as neural networks.

In this paper, we propose an Optimistic Exploration algorithm with Backward Bootstrapped Bonus (OEB3) for DRL. OEB3 is an instantiation of LSVI-UCB (Jin et al., 2020) in DRL by using a general-purpose UCB-bonus to provide an optimistic Q-value and a randomized value function to perform temporally-extended exploration. This general-purpose UCB-bonus represents the disagreement of bootstrapped Q-functions (Osband et al., 2016) to measure the epistemic uncertainty of the unknown optimal value function. Importantly, this proposed UCB-bonus can also be theoretically demonstrated to be equivalent to the bonus-term

[1]Harbin Institute of Technology, Harbin, China [2]Northwestern University, Evanston, USA [3]Tencent Robotics X [4]Tianjin University [5]University of Toronto, Vector Institute. Correspondence to: Chenjia Bai <bai_chenjia@stu.hit.edu.cn>.

A long version *Principled Exploration via Optimistic Bootstrapping and Backward Induction* was accepted by ICML 2021. http://proceedings.mlr.press/v139/bai21d.html.

Workshop on Reinforcement Learning Theory of the 38^{th} International Conference on Machine Learning, 2021.

Algorithm 1 LSVI-UCB in linear MDP

1: **Initialize:** $\Lambda_t \leftarrow \lambda \cdot \mathbf{I}$ and $w_h \leftarrow 0$
2: **for** episode $m = 0$ **to** $M-1$ **do**
3: Receive the initial state s_0
4: **for** step $t = 0$ **to** $T-1$ **do**
5: Take action $a_t = \arg\max_a Q_t(s_t, a)$ and observe s_{t+1}
6: **end for**
7: **for** step $t = T-1$ **to** 0 **do**
8: $\Lambda_t \leftarrow \sum_{\tau=0}^{m} \phi(x_t^\tau, a_t^\tau)\phi(x_t^\tau, a_t^\tau)^\top + \lambda \cdot \mathbf{I}$
9: $w_t \leftarrow \Lambda_t^{-1} \sum_{\tau=0}^{m} \phi(x_t^\tau, a_t^\tau)[r_t(x_t^\tau, a_t^\tau) + \max_a Q_{t+1}(x_{t+1}^\tau, a)]$
10: $Q_t(\cdot,\cdot) = \min\{w_t^\top \phi(\cdot,\cdot) + \alpha[\phi(\cdot,\cdot)^\top \Lambda_t^{-1} \phi(\cdot,\cdot)]^{1/2}, T\}$
11: **end for**
12: **end for**

in LSVI-UCB (Jin et al., 2020), when moving back in linear MDPs. In our case, the Q-value plus the general-purpose UCB-bonus is shown to be an optimistic Q^+ function that is higher than the Q-value for scarcely visited state-action pairs and remains close to the Q-value for frequently visited pairs. Furthermore, we propose an extension of the Episodic Backward Update (EBU) technique (Lee et al., 2019) to propagate future uncertainties to the estimated action-value function consistently within an episode. The backward update exploits the theoretical advantage of LSVI-UCB and empirically improves the sample-efficiency significantly. Extensive evaluations show that OEB3 outperforms several strong exploration methods in 49 Atari games.

2. Background

Considering an MDP represented as $(\mathcal{S}, \mathcal{A}, T, \mathbb{P}, r)$, where $T \in \mathbb{Z}_+$ is the episode length, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, r is the reward function, and $\mathbb{P}$ is the unknown dynamics. In each timestep, the agent observes the current state s_t and takes an action a_t, and then it receives a reward r_t and the next state s_{t+1}. The action-value function $Q^\pi(s_t, a_t) := \mathbb{E}_\pi\big[\sum_{i=t}^{T-1} \gamma^{i-t} r_i\big]$ represents the expected cumulative reward starting from state s_t by taking action a_t and following policy $\pi(a_t|s_t)$ until the end of the episode. $\gamma \in [0, 1)$ is the discount factor. The optimal value function $Q^* = \max_\pi Q^\pi$, and the optimal action $a^* = \text{argmax}_{a \in \mathcal{A}} Q^*(s, a)$.

LSVI-UCB (Jin et al., 2020) uses an optimistic Q-value with LSVI in linear MDP. We denote the feature map of the state-action pair as $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. Furthermore, the transition kernel and reward function are assumed to be linear in ϕ. The LSVI-UCB algorithm is shown in Algorithm 1. For lines 3-6, the agent executes the policy to collect data in an episode. For lines 7-11, the parameter w_t of Q-function is updated in closed-form by following the regularized least-squares problem as $w_t \leftarrow \text{argmin}_{w \in \mathbb{R}^d} \sum_{\tau=0}^{m} \big[r_t(s_t^\tau, a_t^\tau) + \max_{a \in \mathcal{A}} Q_{t+1}(s_{t+1}^\tau, a) - w^\top \phi(s_t^\tau, a_t^\tau)\big]^2 + \lambda \|w\|^2$, where m is the total number of episodes, and τ is the episodic index. The least-squares problem has the explicit solution $w_t = \Lambda_t^{-1} \sum_{\tau=0}^{m} \phi(x_t^\tau, a_t^\tau)\big[r_t(x_t^\tau, a_t^\tau) + \max_a Q_{t+1}(x_{t+1}^\tau, a)\big]$ (line 9), where Λ_t is the Gram matrix. The value function is estimated by $Q_t(s, a) \approx w_t^\top \phi(s, a)$. LSVI-UCB uses an UCB-bonus (Abbasi-Yadkori et al., 2011) in line 10

$$r^{\text{ucb}} = [\phi(s, a)^\top \Lambda_t^{-1} \phi(s, a)]^{1/2} \tag{1}$$

to measure the uncertainty of state-action pairs. The term $u := (\phi^\top \Lambda_t^{-1} \phi)^{-1}$ can be intuitively considered as a pseudo count of the state-action pair in the representation space of ϕ. Thus, the bonus $r^{\text{ucb}} = 1/\sqrt{u}$ represents the uncertainty along the direction of ϕ. By adding the bonus to the Q-value, we obtain an optimistic value function Q^+, which serves as an upper bound of Q to encourage exploration. The bonus in each step is propagated from the end of the episode by the backward update of the Q-value (lines 7-11), which follows the principle of dynamic programming. Theoretical analysis shows that LSVI-UCB achieves a near-optimal worst-case regret of $\tilde{\mathcal{O}}(\sqrt{d^3 T^3 L^3})$ with proper selection of α and λ, where L is the total number of steps.

3. Proposed Method

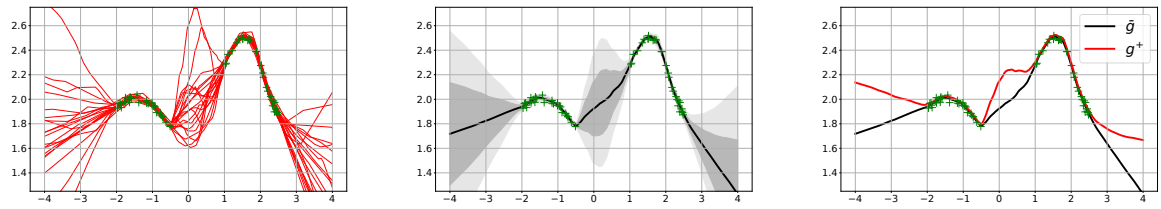
We utilize bootstrapped DQN to construct a general-purpose UCB-bonus, which is theoretically consistent with LSVI-UCB for linear MDPs. We also integrate bootstrapped Q-functions and UCB-bonus into the backward update, which follows the principle of dynamic programming.

3.1. General-Purpose UCB-Bonus

Optimistic exploration uses an optimistic action-value function Q^+ to encourage exploration by adding a bonus term to the standard Q-value. Thus Q^+ serves as an upper bound of the standard Q. The bonus term represents the epistemic uncertainty that results from lacking experiences of the corresponding states and actions. For DRL with deep Q network, it is impractical to derive a closed-form optimistic bonus like (1). Instead, we propose a general-purpose UCB-bonus $\mathcal{B}(s_t, a_t)$ by measuring the disagreement of multiple bootstrapped Q-values $\{Q^k(s_t, a_t)\}_{k=1}^K$ of the state-action pair (s_t, a_t) in a bootstrapped DQN. That is,

$$\mathcal{B}(s_t, a_t) := \sqrt{\frac{1}{K} \sum_{k=1}^{K} \Big(Q^k(s_t, a_t) - \bar{Q}(s_t, a_t)\Big)^2}, \tag{2}$$

where $\bar{Q}(s_t, a_t)$ is the mean of the bootstrapped Q-values. A similar uncertainty measurement was used in Chen et al. (2017). We surprisingly find that this simple form in (2) is also provably efficient for linear MDPs. Indeed, the following theorem establishes the connection between the general-purpose UCB-bonus defined in (2) and the bonus in LSVI-UCB defined in (1).



(a) (b) (c)

Figure 1. Illustration of the general-purpose UCB-bonus in a simple regression task. Green markers indicate there are 60 data points. (a) Regression curves of 20 neural networks. (b) Mean estimation (black curve) and uncertainty measurement (shadow region). (c) The optimistic value (red) and mean value (black).

Theorem 1. *In linear MDPs, the UCB-bonus $\mathcal{B}(s_t, a_t)$ in OEB3 is equivalent to the bonus-term $[\phi_t^\top \Lambda_t^{-1} \phi_t]^{1/2}$ in LSVI-UCB, where $\Lambda_t \leftarrow \sum_{\tau=0}^{m} \phi(x_t^\tau, a_t^\tau)\phi(x_t^\tau, a_t^\tau)^\top + \lambda \cdot \mathbf{I}$, and m is the current episode.*

In Theorem 1, we cast the variance that defines the UCB-bonus of OEB3 as the posterior variance of value functions under the Bayesian learning regime. We remark that the bootstrapped distribution of value functions coincides with the posterior under a Bayesian setting where the prior is uninformative (Friedman et al., 2001). We refer to Appendix A for the details and complete statement. Theorem 1 shows that the general-purpose UCB-bonus in (2) is provably efficient and equivalent to bonus-term in LSVI-UCB for linear cases. Importantly, (2) is a general form for arbitrary Q functions such as deep neural networks.

The optimistic Q^+ is obtained by summing up $\mathcal{B}(s_t, a_t)$ and the estimated Q-function, which takes the form as

$$Q^+(s_t, a_t) := Q(s_t, a_t) + \alpha \mathcal{B}(s_t, a_t), \tag{3}$$

where α is a tuning parameter. We use a simple regression task with neural networks to illustrate the proposed UCB-bonus, as shown in Figure 1. We use 20 neural networks with the same network architecture to solve the same regression problem.

According to Osband et al. (2016), the differences among the outcomes of fitting the 20 neural networks is a result of random initializations. For a given input x, the networks yield different estimations $\{g_i(x)\}_{i=1}^{20}$. It follows from Figure 1(a) that the estimations $\{g_i(x)\}_{i=1}^{20}$ behave similar in the region with large amount of observations, resulting in small disagreement of the estimations. However, for regions with less observations, the disagreement of the estimations inflates a lot. In Figure 1(b), we illustrate the confidence bound of the regression results $\bar{g}(x) \pm \tilde{\sigma}(g_i(x))$ and $\bar{g}(x) \pm 2\tilde{\sigma}(g_i(x))$, where $\bar{g}(x)$ and $\tilde{\sigma}(g_i(x))$ are the mean and standard deviation of the estimations. The standard deviation $\tilde{\sigma}(g_i(x))$ captures the epistemic uncertainty of regression results. Figure 1(c) shows the optimistic estimation $g^+(x) = \bar{g}(x) + \tilde{\sigma}(g_i(x))$ plus the standard deviation. Clearly, the optimistic estimation g^+ is close to $\bar{g}$ in the region with dense observations, and it is larger than $\bar{g}$ in the region with fewer observations.

In DRL, the bootstrapped Q-functions $\{Q^k(s_t, a_t)\}_{k=1}^{K}$, estimated by fitting the target Q-function, perform similarly as $\{g_i(x)\}_{i=1}^{20}$ in the above regression task. A higher UCB-bonus $\mathcal{B}(s_t, a_t) := \tilde{\sigma}(Q^k(s_t, a_t))$ indicates a higher epistemic uncertainty of the action-value function with (s_t, a_t). Therefore, Q^+ produces optimistic estimation for novel state-action pairs and behaves similar to the Q-function in areas that are well explored by the agent. Hence, the optimistic estimation Q^+ encourages the agent to explore the potentially informative state-action pairs efficiently.

3.2. Backward Update of Uncertainty

OEB3 adopts BEBU for backward update when updating the action-value function. BEBU collects a complete trajectory from the replay buffer for each update. Such an approach allows OEB3 to infer the long-term effect in an episode for decision making. In contrast, DQN and Bootstrapped DQN sample one-step transitions, which loses the information containing long-term effects.

It has to be mentioned that BEBU is required to propagate future uncertainty to the estimated action-value function consistently via UCB-bonus. For instance, let $t_2 > t_1$ be indices of two steps in an episode. If Q_{t_2} updates after that of Q_{t_1}, then the uncertainty propagated to Q_{t_1} is inconsistent with that propagated to Q_{t_2}.

To integrate the general-purpose UCB-bonus into bootstrapped Q-learning, we propose a novel Q-target by adding the bonus in both the immediate reward and the next-Q value. The proposed Q-target needs to be suitable for BEBU in training. Formally, the Q-target for updating Q^k is defined as

$$\begin{aligned} y_t^k := &\left[r(s_t, a_t) + \alpha_1 \mathcal{B}(s_t, a_t; \theta)\right] + \gamma \\ &\left[Q^k(s_{t+1}, a'; \theta^{k-}) + \alpha_2 \mathbb{1}_{a' \neq a_{t+1}} \tilde{\mathcal{B}}^k(s_{t+1}, a'; \theta^-)\right], \end{aligned} \tag{4}$$

where $a' = \operatorname{argmax}_a Q^k(s_{t+1}, a; \theta^{k-})$. The choice of a' is

Table 1. Summary of human-normalized scores in 49 games. BEBU, BEBU-UCB, BEBU-IDS and OEB3 are trained for 20M frames.

Frames	200M					20M				
	DQN	UBE	BootDQN	NoisyNet	**BootDQN-IDS**	Bayesian-DQN	BEBU	BEBU-UCB	**BEBU-IDS**	**OEB3**
Mean	241%	440%	553%	651%	**757%**	224%	553%	610%	**622%**	**765%**
Median	93%	126%	139%	172%	**187%**	27%	36%	38%	**44%**	**50%**

determined by the target Q-value without considering the bonus. The immediate reward is added by $\mathcal{B}(s_t, a_t; \theta)$ with a factor α_1, where the bonus $\mathcal{B}$ is computed by bootstrapped Q-network with parameter θ. The next-Q value is added by $\mathbb{1}_{a' \neq a_{t+1}} \tilde{\mathcal{B}}^k(s_{t+1}, a'; \theta^-)$ with factor α_2, where the bonus $\tilde{\mathcal{B}}^k$ is computed by the target network with parameter θ^-. We assign different bonus $\tilde{\mathcal{B}}^k$ of next-Q value to different heads, since the choices of a' are different among the heads. Meanwhile, we assign the same bonus $\mathcal{B}$ of immediate reward for all the heads. We introduce an indicator function $\mathbb{1}_{a' \neq a_{t+1}}$ to control backward update of Q-values. More specifically, in the t-th step, the action-value function Q^k is updated optimistically at the state-action pair (s_{t+1}, a_{t+1}) due to the backward update. Thus, we ignore the bonus of next-Q value in the update of Q^k when a' is equal to a_{t+1}.

3.3. Comparison with LSVI-UCB

We remark that both LSVI-UCB and OEB3 constructs the confidence interval of value functions based on the frequentist approaches. Specifically, LSVI-UCB constructs the confidence intervals explicitly based on the linear model, whereas OEB3 constructs the confidence interval based on the non-parametric bootstrapped approach. In OEB3, we adopt Bootstrapped Q-values to calculate the standard deviation of Q-functions with neural network parameterization, which coincides with the bonus in LSVI-UCB on linear MDPs. When the sample size increases, the distribution of bootstrapped Q-values converges asymptotically to the posterior under a Bayesian setting where the prior is uninformative (Friedman et al., 2001). Hence, in Theorem 1, we use the Bayesian setting as a simplification to motivate our algorithm while this is not necessary. A recent approach also uses a similar way to motivate the worst-case regret of randomized value functions (Russo, 2019).

4. Experimental Results

We evaluate the algorithms in 49 Atari games. Directly comparing OEB3 with baselines using Bootstrapped DQN is not fair, since OEB3 uses backward update for training. To achieve fair comparison, we reimplement all Bootstrapped DQN-based baselines with BEBU. We compare the following methods. (1) **OEB3**: the proposed principled exploration method. (2) **BEBU**: a reimplementation of Bootstrapped DQN (Osband et al., 2016) with BEBU. (3) **BEBU-UCB**: BEBU with optimistic actions selected by the upper bound of Q (Chen et al., 2017; Lee et al., 2020). (4) **BEBU-IDS**: integrating homoscedastic IDS (Nikolov et al., 2019) into BEBU without distributional RL.

We additionally compare the performance of DQN (Mnih et al., 2015), NoisyNet (Fortunato et al., 2018), Bootstrapped DQN (BootDQN) (Osband et al., 2016), BootDQN-IDS (Nikolov et al., 2019), UBE (O'Donoghue et al., 2018) in 200M training frames, and Bayesian DQN (Azizzadenesheli et al., 2018) in 20M training frames. We choose NoisyNet as a baseline since it has been evaluated on the entire Atari suite (instead of several hard exploration games) such that it performs substantially better than existing bonus-based methods (Taiga et al., 2020), including CTS-counts (Bellemare et al., 2016), PixelCNN-counts (Ostrovski et al., 2017), RND (Burda et al., 2019), and ICM (Pathak et al., 2017). An ensemble policy by a majority vote of Q-heads is used for 30 no-op evaluation.

Table 1 reports the overall performance of all the methods on 49 Atari games. According to Table 1, BootDQN-IDS performs better than UBE, BootDQN, and NoisyNet. Thus, BootDQN-IDS outperforms popular bonus-based exploration methods that perform worse than NoisyNet (Taiga et al., 2020). We then reimplement BootDQN-IDS with BEBU, and we refer this version to as BEBU-IDS. We observe that OEB3 outperforms BEBU-IDS in both mean and medium scores, as well as outperforming all other bonus-based methods in the backward update setting. We report the detailed raw scores in Appendix B. OEB3 outperforms BEBU, BEBU-UCB, and BEBU-IDS in 36, 34, and 35 games out of all 49 games, respectively.

5. Conclusion

In this work, we have proposed a principled exploration method, i.e., OEB3, that shares nice theoretical properties as LSVI-UCB. By integrating with backward update, the sample efficiency is further enhanced. As far as we see, our work seems to establish the first empirical attempt of uncertainty propagation in deep RL, which exploits the core benefit of theoretical analysis. Moreover, we observe that the connection between theoretical analysis and practical algorithm provides strong empirical performance, which hopefully raises insights on combining theory and practice to the community.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *NeurIPS*, pp. 2312–2320, 2011.

Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *ICML*, pp. 322–332. PMLR, 2019.

Auer, P. and Ortner, R. Logarithmic online regret bounds for undiscounted reinforcement learning. In *NeurIPS*, pp. 49–56, 2007.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *ICML*, pp. 263–272, 2017.

Azizzadenesheli, K., Brunskill, E., and Anandkumar, A. Efficient exploration through bayesian deep q-networks. In *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–9. IEEE, 2018.

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

Burda, Y., Edwards, H., Storkey, A., and Klimov, O. Exploration by random network distillation. In *ICLR*, 2019.

Chen, R. Y., Sidor, S., Abbeel, P., and Schulman, J. Ucb exploration via q-ensembles. *arXiv preprint arXiv:1706.01502*, 2017.

Dann, C. and Brunskill, E. Sample complexity of episodic fixed-horizon reinforcement learning. In *NeurIPS*, pp. 2818–2826, 2015.

Fortunato, M., Azar, M. G., Piot, B., Menick, J., Osband, I., Graves, A., Mnih, V., Munos, R., Hassabis, D., Pietquin, O., Blundell, C., and Legg, S. Noisy networks for exploration. In *ICLR*, 2018.

Friedman, J., Hastie, T., and Tibshirani, R. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

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

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? In *NeurIPS*, pp. 4863–4873, 2018.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *CoLT*, pp. 2137–2143, 2020.

Lee, K., Laskin, M., Srinivas, A., and Abbeel, P. Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning. *arXiv preprint arXiv:2007.04938*, 2020.

Lee, S. Y., Sungik, C., and Chung, S.-Y. Sample-efficient deep reinforcement learning via episodic backward update. In *NeurIPS*, pp. 2110–2119, 2019.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M. A., Fidjeland, A., 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*, 518:529–533, 2015.

Nikolov, N., Kirschner, J., Berkenkamp, F., and Krause, A. Information-directed exploration for deep reinforcement learning. In *ICLR*, 2019.

Osband, I. and Van Roy, B. Why is posterior sampling better than optimism for reinforcement learning? In *ICML*, pp. 2701–2710, 2017.

Osband, I., Blundell, C., Pritzel, A., and Van Roy, B. Deep exploration via bootstrapped dqn. In *NeurIPS*, pp. 4026–4034, 2016.

Ostrovski, G., Bellemare, M. G., van den Oord, A., and Munos, R. Count-based exploration with neural density models. In *ICML*, pp. 2721–2730, 2017.

O'Donoghue, B., Osband, I., Munos, R., and Mnih, V. The uncertainty bellman equation and exploration. In *ICML*, pp. 3836–3845, 2018.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction. In *ICML*, pp. 2778–2787, 2017.

Russo, D. Worst-case regret bounds for exploration via randomized value functions. In *NeurIPS*, pp. 14410–14420, 2019.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Taiga, A. A., Fedus, W., Machado, M. C., Courville, A., and Bellemare, M. G. On bonus based exploration methods in the arcade learning environment. In *ICLR*, 2020.

West, M. Outlier models and prior distributions in bayesian linear regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(3):431–439, 1984.

Zanette, A., Brandfonbrener, D., Brunskill, E., Pirotta, M., and Lazaric, A. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pp. 1954–1964, 2020.