

Hill Climbing on Value Estimates for Search-control in Dyna

Yangchen Pan¹, Hengshuai Yao², Amir-massoud Farahmand^{3,4} and Martha White¹

¹Department of Computing Science, University of Alberta, Canada

²Huawei HiSilicon, Canada

³Vector Institute, Canada

⁴Department of Computer Science, University of Toronto, Canada

pan6@ualberta.ca, hengshuai.yao@huawei.com, farahmand@vectorinstitute.ai, whitem@ualberta.ca

Abstract

Dyna is an architecture for model-based reinforcement learning (RL), where simulated experience from a model is used to update policies or value functions. A key component of Dyna is search-control, the mechanism to generate the state and action from which the agent queries the model, which remains largely unexplored. In this work, we propose to generate such states by using the trajectory obtained from Hill Climbing (HC) the current estimate of the value function. This has the effect of propagating value from high-value regions and of preemptively updating value estimates of the regions that the agent is likely to visit next. We derive a noisy projected natural gradient algorithm for hill climbing, and highlight a connection to Langevin dynamics. We provide an empirical demonstration on four classical domains that our algorithm, HC-Dyna, can obtain significant sample efficiency improvements. We study the properties of different sampling distributions for search-control, and find that there appears to be a benefit specifically from using the samples generated by climbing on current value estimates from low-value to high-value region.

1 Introduction

Experience replay (ER) [Lin, 1992] is currently the most common way to train value functions approximated as neural networks (NNs), in an online RL setting [Adam *et al.*, 2012; Wawrzyński and Tanwani, 2013]. The buffer in ER is typically a recency buffer, storing the most recent transitions, composed of state, action, next state and reward. At each environment time step, the NN gets updated by using a mini-batch of samples from the ER buffer, that is, the agent replays those transitions. ER enables the agent to be more sample efficient, and in fact can be seen as a simple form of model-based RL [van Seijen and Sutton, 2015]. This connection is specific to the Dyna architecture [Sutton, 1990; Sutton, 1991], where the agent maintains a search-control (SC) queue of pairs of states and actions and uses a model to generate next states and rewards. These simulated transitions are used to update values. ER, then, can be seen as a variant

of Dyna with a nonparameteric model, where search-control is determined by the observed states and actions.

By moving beyond ER to Dyna with a learned model, we can potentially benefit from increased flexibility in obtaining simulated transitions. Having access to a model allows us to generate unobserved transitions, from a given state-action pair. For example, a model allows the agent to obtain on-policy or exploratory samples from a given state, which has been reported to have advantages [Gu *et al.*, 2016; Pan *et al.*, 2018; Santos *et al.*, 2012; Peng *et al.*, 2018]. More generally, models allow for a variety of choices for search-control, which is critical as it emphasizes different states during the planning phase. Prioritized sweeping [Moore and Atkeson, 1993] uses the model to obtain predecessor states, with states sampled according to the absolute value of temporal difference error. This early work, and more recent work [Sutton *et al.*, 2008; Pan *et al.*, 2018; Corneil *et al.*, 2018], showed this addition significantly outperformed Dyna with states uniformly sampled from observed states. Most of the work on search-control, however, has been limited to sampling visited or predecessor states. Predecessor states require a reverse model, which can be limiting. The range of possibilities has yet to be explored for search-control and there is room for many more ideas.

In this work, we investigate using sampled trajectories by hill climbing on our learned value function to generate states for search-control. Updating along such trajectories has the effect of propagating value from regions the agent currently believes to be high-value. This strategy enables the agent to preemptively update regions where it is likely to visit next. Further, it focuses updates in areas where approximate values are high, and so important to the agent. To obtain such states for search-control, we propose a noisy natural projected gradient algorithm. We show this has a connection to Langevin dynamics, whose distribution converges to the Gibbs distribution, where the density is proportional to the exponential of the state values. We empirically study different sampling distributions for populating the search-control queue, and verify the effectiveness of hill climbing based on estimated values. We conduct experiments showing improved performance in four benchmark domains, as compared to DQN¹, and illustrate the usage of our architecture for continuous control.

¹We use DQN to refer to the algorithm by [Mnih *et al.*, 2015] that uses ER and target network, but not the exact original architecture.

2 Background

We formalize the environment as a Markov Decision Process (MDP) $(\mathcal{S}, \mathcal{A}, \mathbb{P}, R, \gamma)$, where \mathcal{S} is the state space, \mathcal{A} is the action space, $\mathbb{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ is the transition probabilities, $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ is the reward function, and $\gamma \in [0, 1]$ is the discount factor. At each time step $t = 1, 2, \dots$, the agent observes a state $s_t \in \mathcal{S}$ and takes an action $a_t \in \mathcal{A}$, transitions to $s_{t+1} \sim \mathbb{P}(\cdot | s_t, a_t)$ and receives a scalar reward $r_{t+1} \in \mathbb{R}$ according to the reward function R .

Typically, the goal is to learn a policy to maximize the expected return starting from some fixed initial state. One popular algorithm is Q-learning, by which we can obtain approximate action-values $Q_\theta : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ for parameters θ . The policy corresponds to acting greedily according to these action-values: for each state, select action $\arg \max_{a \in \mathcal{A}} Q(s, a)$. The Q-learning update for a sampled transition $s_t, a_t, r_{t+1}, s_{t+1}$ is

$$\theta \leftarrow \theta + \alpha \delta_t \nabla Q_\theta(s_t, a_t)$$

$$\text{where } \delta_t \stackrel{\text{def}}{=} r_{t+1} + \max_{a' \in \mathcal{A}} Q_\theta(s_{t+1}, a') - Q_\theta(s_t, a_t)$$

Though frequently used, such an update may not be sound with function approximation. Target networks [Mnih *et al.*, 2015] are typically used to improve stability when training NNs, where the bootstrap target on the next step is a fixed, older estimate of the action-values.

ER and Dyna can both be used to improve sample efficiency of DQN. Dyna is a model-based method that simulates (or replays) transitions, to reduce the number of required interactions with the environment. A model is sometimes available a priori (e.g., from physical equations of the dynamics) or is learned using data collected through interacting with the environment. The generic Dyna architecture, with explicit pseudo-code given by [Sutton and Barto, 2018, Chapter 8], can be summarized as follows. When the agent interacts with the real world, it updates both the action-value function and the learned model using the real transition. The agent then performs n planning steps. In each planning step, the agent samples (\tilde{s}, \tilde{a}) from the search-control queue, generates next state \tilde{s}' and reward \tilde{r} from (\tilde{s}, \tilde{a}) using the model, and updates the action-values using Q-learning with the tuple $(\tilde{s}, \tilde{a}, \tilde{r}, \tilde{s}')$.

3 A Motivating Example

In this section we provide an example of how the value function surface changes during learning on a simple continuous-state GridWorld domain. This provides intuition on why it is useful to populate the search-control queue with states obtained by hill climbing on the estimated value function, as proposed in the next section.

Consider the GridWorld in Figure 1(a), which is a variant of the one introduced by [Peng and Williams, 1993]. In each episode, the agent starts from a uniformly sampled point from the area $[0, 0.05]^2$ and terminates when it reaches the goal area $[0.95, 1.0]^2$. There are four actions {UP, DOWN, LEFT, RIGHT}; each leads to a 0.05 unit move towards the corresponding direction. As a cost-to-goal problem, the reward is -1 per step.

In Figure 1, we plot the value function surface after 0, 14k and 20k mini-batch updates to DQN. We visualize the

gradient ascent trajectories with 100 gradient steps starting from five states $(0.1, 0.1)$, $(0.9, 0.9)$, $(0.1, 0.9)$, $(0.9, 0.1)$, and $(0.3, 0.4)$. The gradient of the value function used in the gradient ascent is

$$\nabla_s V(s) = \nabla_s \max_a Q_\theta(s, a), \quad (1)$$

At the beginning, with a randomly initialized NN, the gradient with respect to state is almost zero, as seen in Figure 1(b). As the DQN agent updates its parameters, the gradient ascent generates trajectories directed towards the goal, though after only 14k steps, these are not yet contiguous, as seen Figure 1(c). After 20k steps, as in Figure 1(d), even though the value function is still inaccurate, the gradient ascent trajectories take all initial states to the goal area. This suggests that as long as the estimated value function roughly reflects the shape of the optimal value function, the trajectories provide a demonstration of how to reach the goal—or high-value regions—and speed up learning by focusing updates on these relevant regions.

More generally, by focusing planning on regions the agent *thinks* are high-value, it can quickly correct value function estimates before visiting those regions, and so avoid unnecessary interaction. We demonstrate this in Figure 1(e), where the agent obtains gains in performance by updating from high-value states, even when its value estimates have the wrong shape. After 20k learning steps, the values are flipped by negating the sign of the parameters in the output layer of the NN. HC-Dyna, introduced in Section 5, quickly recovers compared to DQN and OnPolicy updates from the ER buffer. Planning steps help pushing down these erroneously high-values, and the agent can recover much more quickly.

4 Effective Hill Climbing

To generate states for search control, we need an algorithm that can climb on the estimated value function surface. For general value function approximators, such as NNs, this can be difficult. The value function surface can be very flat or very rugged, causing the gradient ascent to get stuck in local optima and hence interrupt the gradient traveling process. Further, the state variables may have very different numerical scales. When using a regular gradient ascent method, it is likely for the state variables with a smaller numerical scale to immediately go out of the state space. Lastly, gradient ascent is unconstrained, potentially generating unrealizable states.

In this section, we propose solutions for all these issues. We provide a noisy invariant projected gradient ascent strategy to generate meaningful trajectories of states for search-control. We then discuss connections to Langevin dynamics, a model for heat diffusion, which provides insight into the sampling distribution of our search-control queue.

4.1 Noisy Natural Projected Gradient Ascent

To address the first issue, of flat or rugged function surfaces, we propose to add Gaussian noise on each gradient ascent step. Intuitively, this provides robustness to flat regions and avoids getting stuck in local maxima on the function surface, by diffusing across the surface to high-value regions.

To address the second issue of vastly different numerical scales among state variables, we use a standard strategy to be

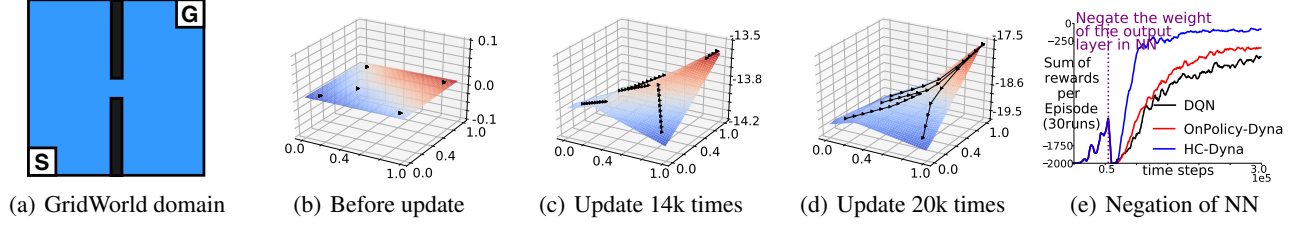


Figure 1: (b-d) The value function on the GridWorld domain with gradient ascent trajectories. (e) shows learning curves (sum of rewards per episode v.s. time steps) where each algorithm needs to recover from a bad NN initialization (i.e. the value function looks like the reverse of (d)).

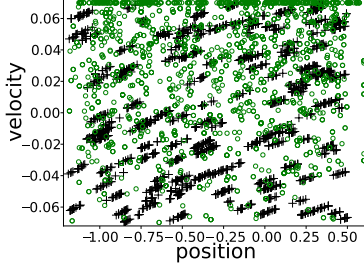


Figure 2: The search-control queue filled by using (+) or not using (o) natural gradient on MountainCar-v0.

invariant to scale: natural gradient ascent. A popular choice of natural gradient is derived by defining the metric tensor as the Fisher information matrix [Amari and Douglas, 1998; Amari, 1998; Thomas *et al.*, 2016]. We introduce a simple and computationally efficient metric tensor: the inverse of covariance matrix of the states Σ_s^{-1} . This choice is simple, because the covariance matrix can easily be estimated online. We can define the following inner product:

$$\langle s, s' \rangle = s^\top \Sigma_s^{-1} s', \forall s, s' \in S,$$

which induces a vector space—the Riemannian manifold—where we can compute the distance of two points s and $s + \Delta$ that are close to each other by $d(s, s + \Delta) \stackrel{\text{def}}{=} \Delta^\top \Sigma_s^{-1} \Delta$. The steepest ascent updating rule based on this distance metric becomes $s \leftarrow s + \alpha \Sigma_s g$, where g is the gradient vector.

We demonstrate the utility of using the natural gradient scaling. Figure 2 shows the states from the search-control queue filled by hill climbing in early stages of learning (after 8000 steps) on MountainCar. The domain has two state variables with very different numerical scale: position $\in [-1.2, 0.6]$ and velocity $\in [-0.07, 0.07]$. Using a regular gradient update, the queue shows a state distribution with many states concentrated near the top since it is very easy for the velocity variable to go out of boundary. In contrast, the one with natural gradient, shows clear trajectories with an obvious tendency to the right top area (position ≥ 0.5), which is the goal area.

We use projected gradient updates to address the third issue regarding unrealizable states. We explain the issue and solution using the Acrobot domain. The first two state variables are $\cos \theta, \sin \theta$, where θ is the angle between the first robot arm’s link and the vector pointing downwards. This induces the restriction that $\cos^2 \theta + \sin^2 \theta = 1$. The hill climbing process could generate many states that do not satisfy this restriction.

This could potentially degrade performance, since the NN needs to generalize to these states unnecessarily. We can use a projection operator Π to enforce such restrictions, whenever known, after each gradient ascent step. In Acrobot, Π is a simple normalization. In many settings, the constraints are simple box constraints, with projection just inside the boundary.

Now we are ready to introduce our final hill climbing rule:

$$s \leftarrow \Pi(s + \alpha \Sigma_s g + \mathcal{N}), \quad (2)$$

where \mathcal{N} is Gaussian noise and α a stepsize. For simplicity, we set the stepsize to $\alpha = 0.1 / \|\Sigma_s g\|$ across all results in this work, though of course there could be better choices.

4.2 Connection to Langevin Dynamics

The proposed hill climbing procedure is similar to Langevin dynamics, which is frequently used as a tool to analyze optimization algorithms or to acquire an estimate of the expected parameter values w.r.t. some posterior distribution in Bayesian learning [Welling and Teh, 2011]. The overdamped Langevin dynamics can be described by a stochastic differential equation (SDE) $dW(t) = \nabla U(W_t)dt + \sqrt{2}dB_t$, where $B_t \in \mathbb{R}^d$ is a d -dimensional Brownian motion and U is a continuous differentiable function. Under some conditions, it turns out that the Langevin diffusion $(W_t)_{t \geq 0}$ converges to a unique invariant distribution $p(x) \propto \exp(U(x))$ [Chiang *et al.*, 1987].

By apply the Euler-Maruyama discretization scheme to the SDE, we acquire the discretized version $Y_{k+1} = Y_k + \alpha_{k+1} \nabla U(Y_k) + \sqrt{2\alpha_{k+1}} Z_{k+1}$ where $(Z_k)_{k \geq 1}$ is an i.i.d. sequence of standard d -dimensional Gaussian random vectors and $(\alpha_k)_{k \geq 1}$ is a sequence of step sizes. This discretization scheme was used to acquire samples from the original invariant distribution $p(x) \propto \exp(U(x))$ through the Markov chain $(Y_k)_{k \geq 1}$ when it converges to the chain’s stationary distribution [Roberts, 1996]. The distance between the limiting distribution of $(Y_k)_{k \geq 1}$ and the invariant distribution of the underlying SDE has been characterized through various bounds [Durmus and Moulines, 2017].

When we perform hill climbing, the parameter θ is constant at each time step t . By choosing the function U in the SDE above to be equal to V_θ , we see that the state distribution $p(s)$ in our search-control queue is approximately²

$$p(s) \propto \exp(V_\theta(s)).$$

²Different assumptions on $(\alpha_k)_{k \geq 1}$ and properties of U can give convergence claims with different strengths. Also refer to [Welling and Teh, 2011] for the discussion on the use of a preconditioner.

Algorithm 1 HC-Dyna

Input: budget k for the number of gradient ascent steps (e.g., $k = 100$), stochasticity η for gradient ascent (e.g., $\eta = 0.1$), ρ percentage of updates from SC queue (e.g., $\rho = 0.5$), d the number of state variables, i.e. $\mathcal{S} \subset \mathbb{R}^d$. Initialize empty SC queue B_{sc} and ER buffer B_{er} .
 $\hat{\Sigma}_s \leftarrow \mathbf{I}$ (empirical covariance matrix)
 $\mu_{ss} \leftarrow \mathbf{0} \in \mathbb{R}^{d \times d}, \mu_s \leftarrow \mathbf{0} \in \mathbb{R}^d$ (auxiliary variables for computing empirical covariance matrix, sample average will be maintained for μ_{ss}, μ_s)
 $\epsilon_a \leftarrow 0$ (threshold for accepting a state)
for $t = 1, 2, \dots$ **do**
 Observe (s, a, s', r) and add it to B_{er}
 $\mu_{ss} \leftarrow \frac{\mu_{ss}(t-1) + ss^\top}{t}, \mu_s \leftarrow \frac{\mu_s(t-1) + s}{t}$
 $\hat{\Sigma}_s \leftarrow \mu_{ss} - \mu_s \mu_s^\top$
 $\epsilon_a \leftarrow (1 - \beta)\epsilon_a + \beta \text{distance}(s', s)$ for $\beta = 0.001$
 Sample s_0 from $B_{er}, \tilde{s} \leftarrow \infty$
 for $i = 0, \dots, k$ **do**
 $g_{s_i} \leftarrow \nabla_s V(s_i) = \nabla_s \max_a Q_\theta(s_i, a)$
 $s_{i+1} \leftarrow \Pi(s_i + \frac{0.1}{\|\hat{\Sigma}_s g_{s_i}\|} \hat{\Sigma}_s g_{s_i} + X_i), X_i \sim \mathcal{N}(0, \eta \hat{\Sigma}_s)$

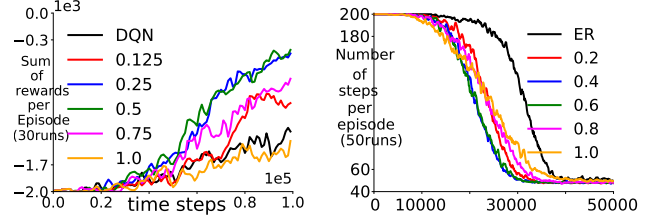
 if $\text{distance}(\tilde{s}, s_{i+1}) \geq \epsilon_a$ **then**
 Add s_{i+1} into $B_{sc}, \tilde{s} \leftarrow s_{i+1}$
 for n times **do**
 Sample a mixed mini-batch b , with proportion ρ from B_{sc} and $1 - \rho$ from B_{er}
 Update parameters θ (i.e. DQN update) with b

An important difference between the theoretical limiting distribution and the actual distribution acquired by our hill climbing method is that our trajectories would also include the states during the *burn-in* or transient period, which refers to the period before the stationary behavior is reached. We would want to point out that those states play an essential role in improving learning efficiency as we will demonstrate in section 6.2.

5 Hill Climbing Dyna

In this section, we provide the full algorithm, called Hill Climbing Dyna, summarized in Algorithm 1. The key component is to use the Hill Climbing procedure developed in the previous section, to generate states for search-control (SC). To ensure some separation between states in the search-control queue, we use a threshold ϵ_a to decide whether or not to add a state into the queue. We use a simple heuristic to set this threshold on each step, as the following sample average: $\epsilon_a \approx \epsilon_a^{(T)} = \sum_{t=1}^T \frac{\|s_{t+1} - s_t\|_2 / \sqrt{d}}{T}$. The start state for the gradient ascent is randomly sampled from the ER buffer.

In addition to using this new method for search control, we also found it beneficial to include updates on the experience generated in the real world. The mini-batch sampled for training has ρ proportion of transitions generated by states from the SC queue, and $1 - \rho$ from the ER buffer. For example, for $\rho = 0.75$ with a mini-batch size of 32, the updates consists of 24 ($= 32 \times 0.75$) transitions generated from states in the SC queue and 8 transitions from the ER buffer. Previous work using Dyna for learning NN value functions also used such



(a) (Continuous state) GridWorld

(b) TabularGridWorld

Figure 3: The effect of mixing rate on learning performance. The numerical label means HC-Dyna with a certain mixing rate.

mixed mini-batches [Holland *et al.*, 2018].

One potential reason this addition is beneficial is that it alleviates issues with heavily skewing the sampling distribution to be off-policy. Tabular Q-learning is an off-policy learning algorithm, which has strong convergence guarantees under mild assumptions [Tsitsiklis, 1994]. When moving to function approximation, however, convergence of Q-learning is much less well understood. The change in sampling distribution for the states could significantly impact convergence rates, and potentially even cause divergence. Empirically, previous prioritized ER work pointed out that skewing the sampling distribution from the ER buffer can lead to a biased solution [Schaul *et al.*, 2016]. Though the ER buffer is not on-policy, because the policy is continually changing, the distribution of states is closer to the states that would be sampled by the current policy than those in SC. Using mixed states from the ER buffer, and those generated by Hill Climbing, could alleviate some of the issues with this skewness.

Another possible reason that such mixed sampling could be necessary is due to model error. The use of real experience could mitigate issues with such error. We found, however, that this mixing has an effect even when using the true model. This suggests that this phenomenon indeed is related to the distribution over states.

We provide a small experiment in the GridWorld, depicted in Figure 1, using both a continuous-state and a discrete-state version. We include a discrete state version, so we can demonstrate that the effect persists even in a tabular setting when Q-learning is known to be stable. The continuous-state setting uses NNs—as described more fully in Section 6—with a mini-batch size of 32. For the tabular setting, the mini-batch size is 1; updates are randomly selected to be from the SC queue or ER buffer proportional to ρ . Figure 3 shows the performance of HC-Dyna as the mixing proportion increases from 0 (ER only) to 1.0 (SC only). In both cases, a mixing rate around $\rho = 0.5$ provides the best results. Generally, using too few search-control samples do not improve performance; focusing too many updates on search-control samples seems to slightly speed up early learning, but then later learning suffers. In all further experiments in this paper, we set $\rho = 0.5$.

6 Experiments

In this section, we demonstrate the utility of (DQN-)HC-Dyna in several benchmark domains, and then analyze the learning effect of different sampling distributions to generate states for the search-control queue.

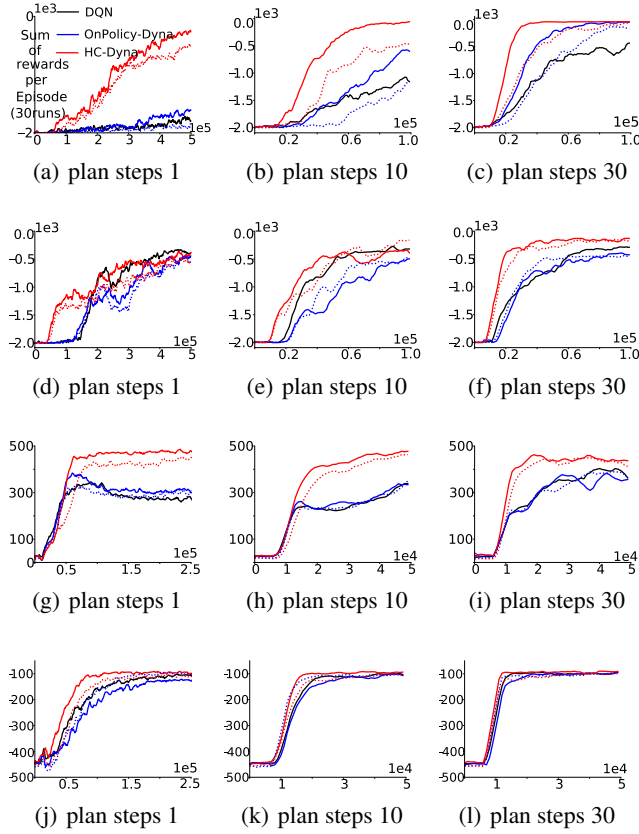


Figure 4: Evaluation curves (sum of episodic reward v.s. environment time steps) of (DQN-)HC-Dyna, (DQN-)OnPolicy-Dyna, DQN on GridWorld (a-c), MountainCar-v0 (d-f), CartPole-v1 (g-i) and Acrobot-v1 (j-l). The curves plotted by *dotted* lines are using online learned models. Results are averaged over 30 random seeds.

6.1 Results in Benchmark Domains

In this section, we present empirical results on four classic domains: the GridWorld (Figure 1(a)), MountainCar, CartPole and Acrobot. We present both discrete and continuous action results in the GridWorld, and compare to DQN for the discrete control and to Deep Deterministic Policy Gradient (DDPG) for the continuous control [Lillicrap *et al.*, 2016]. The agents all use a two-layer NN, with ReLU activations and 32 nodes in each layer. We include results using both the true model and the learned model, on the same plots. We further include multiple planning steps n , where for each real environment step, the agent does n updates with a mini-batch of size 32.

In addition to ER, we add an on-policy baseline called **OnPolicy-Dyna**. This algorithm samples a mini-batch of states (not the full transition) from the ER buffer, but then generates the next state and reward using an on-policy action. This baseline distinguishes when the gain of HC-Dyna algorithm is due to on-policy sampled actions, rather than because of the states in our search-control queue.

Discrete Action

The results in Figure 4 show that (a) HC-Dyna never harms performance over ER and OnPolicy-Dyna, and in some cases

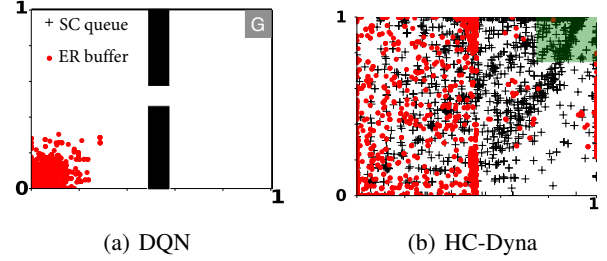


Figure 5: Figure (a)(b) show buffer(red \cdot)/queue(black $+$) distribution on GridWorld ($s \in [0, 1]^2$) by uniformly sampling 2k states. (a) is showing ER buffer when running DQN, hence there is no “+” in it. (b) shows 0.2% of the ER samples fall in the green shadow (i.e. high value region), while 27.8% samples from the SC queue are there.

significantly improves performance, (b) these gains persist even under learned models and (c) there are clear gains from HC-Dyna even with a small number of planning steps. Interestingly, using multiple mini-batch updates per time step can significantly improve the performance of all the algorithms. DQN, however, has very limited gain when moving from 10 to 30 planning steps on all domains except GridWorld, whereas HC-Dyna seems to more noticeably improve from more planning steps. This implies a possible limit of the usefulness of only using samples in the ER buffer.

We observe that the on-policy actions does not always help. The GridWorld domain is in fact the only one where on-policy actions (OnPolicy-Dyna) shows an advantage as the number of planning steps increase. This result provides evidence that the gain of our algorithm is due to the states in our search-control queue, rather than on-policy sampled actions. We also see that even though both model-based methods perform worse when the model has to be learned compared to when the true model is available, HC-Dyna is consistently better than OnPolicy-Dyna across all domains/settings.

To gain intuition for why our algorithm achieves superior performance, we visualize the states in the search-control queue for HC-Dyna in the GridWorld domain (Figure 5). We also show the states in the ER buffer at the same time step, for both HC-Dyna and DQN to contrast. There are two interesting outcomes from this visualization. First, the modification to search-control significantly changes where the agent explores, as evidenced by the ER buffer distribution. Second, HC-Dyna has many states in the SC queue that are near the goal region even when its ER buffer samples concentrate on the left part on the square. The agent can still update around the goal region even when it is physically in the left part of the domain.

Continuous Control

Our architecture can easily be used with continuous actions, as long as the algorithm estimates values. We use DDPG [Lillicrap *et al.*, 2016] as an example for use inside HC-Dyna. DDPG is an actor-critic algorithm that uses the deterministic policy gradient theorem [Silver *et al.*, 2014]. Let $\pi_\psi(\cdot) : \mathcal{S} \rightarrow \mathcal{A}$ be the actor network parameterized by ψ , and $Q_\theta(\cdot, \cdot) : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ be the critic. Given an initial state value s , the gradient ascent direction can be computed by $\nabla_s Q_\psi(s, \pi_\theta(s))$. In fact, because the gradient step causes small changes, we can further approximate this gra-

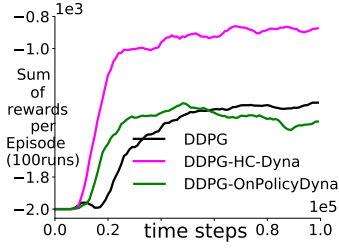


Figure 6: HC-Dyna for continuous control with DDPG. We used 5 planning steps in this experiment.

dient more efficiently using $\nabla_s Q_\psi(s, a^*)$, $a^* \stackrel{\text{def}}{=} \pi_\theta(s)$, without backpropagating the gradient through the actor network. We modified the GridWorld in Figure 1(a) to have action space $\mathcal{A} = [-1, 1]^2$ and an action $a_t \in \mathcal{A}$ is executed as $s_{t+1} \leftarrow s_t + 0.05a_t$. Figure 6 shows the learning curve of DDPG, and DDPG with OnPolicy-Dyna and with HC-Dyna. As before, HC-Dyna shows significant early learning benefits and also reaches a better solution. This highlights that improved search-control could be particularly effective for algorithms that are known to be prone to local minima, like Actor-Critic algorithms.

6.2 Investigating Sampling Distributions for Search-control

We next investigate the importance of two choices in HC-Dyna: (a) using trajectories to high-value regions and (b) using the agent’s value estimates to identify these important regions. To test this, we include following sampling methods for comparison: (a) HC-Dyna: hill climbing by using \hat{V}_θ (our algorithm); (b) Gibbs: sampling $\propto \exp(\hat{V}_\theta)$; (c) HC-Dyna-Vstar: hill climbing by using V^* and (d) Gibbs-Vstar: sampling $\propto \exp(V^*)$, where V^* is a pre-learned optimal value function. We also include the baselines OnPolicyDyna, ER and Uniform-Dyna, which uniformly samples states from the whole state space. All strategies mix with ER, using $\rho = 0.5$, to better give insight into performance differences.

To facilitate sampling from the Gibbs distribution and computing the optimal value function, we test on a simplified TabularGridWorld domain of size 20×20 , without any obstacles. Each state is represented by an integer $i \in \{1, \dots, 400\}$, assigned from bottom to top, left to right on the square with 20×20 grids. HC-Dyna and HC-Dyna-Vstar assume that the state space is continuous on the square $[0, 1]^2$ and each grid can be represented by its center’s (x, y) coordinates. We use the finite difference method for hill climbing.

Comparing to the Gibbs Distribution

As we pointed out the connection to the Langevin dynamics in Section 4.2, the limiting behavior of our hill climbing strategy is approximately a Gibbs distribution. Figure 7(a) shows that HC-Dyna performs the best among all sampling distributions, including Gibbs and other baselines. This result suggests that the states during the burn-in period matter. Figure 7(b) shows the state count by randomly sampling *the same number of states* from the HC-Dyna’s search-control queue and from that filled by Gibbs distribution. We can see that the Gibbs one

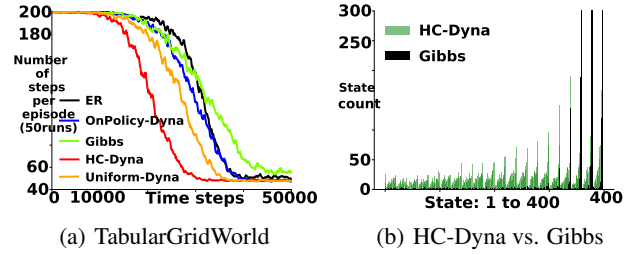


Figure 7: (a) Evaluation curve; (b) Search-control distribution

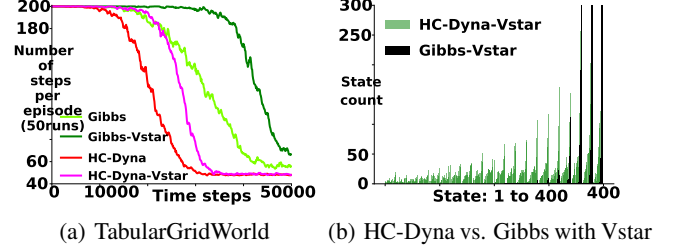


Figure 8: (a) Evaluation curve; (b) Search-control distribution

concentrates its distribution only on very high value states.

Comparing to True Values

One hypothesis is that the value estimates guide the agent to the goal. A natural comparison, then, is to use the optimal values, which should point the agent directly to the goal. Figure 8(a) indicates that using the estimates, rather than true values, is more beneficial for planning. This result highlights that there does seem to be some additional value to focusing updates based on the agent’s current value estimates. Comparing state distribution of Gibbs-Vstar and HC-Dyna-Vstar in Figure 8(b) to Gibbs and HC-Dyna in Figure 7(b), one can see that both distributions are even more concentrated, which seems to negatively impact performance.

7 Conclusion

We presented a new Dyna algorithm, called HC-Dyna, which generates states for search-control by using hill climbing on value estimates. We proposed a noisy natural projected gradient ascent strategy for the hill climbing process. We demonstrate that using states from hill climbing can significantly improve sample efficiency in several benchmark domains. We empirically investigated, and validated, several choices in our algorithm, including the use of natural gradients, the utility of mixing with ER samples, the benefits of using estimated values for search control. A natural next step is to further investigate other criteria for assigning importance to states. Our HC strategy is generic for any smooth function; not only for value estimates. A possible alternative is to investigate importance based on error in a region, or based more explicitly on optimism or uncertainty, to encourage systematic exploration.

Acknowledgments

We would like to acknowledge funding from the Canada CIFAR AI Chairs Program, Amii and NSERC.

A Appendix

The appendix includes all algorithmic and experimental details.

A.1 Algorithmic details

We include the classic Dyna architecture [Sutton, 1991; Sutton and Barto, 2018] in Algorithm 2 and our algorithm with additional details in Algorithm 3.

Algorithm 2 Generic Dyna Architecture: Tabular Setting

```

Initialize  $Q(s, a)$  and model  $\mathcal{M}(s, a)$ ,  $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$ 
At each learning step:
Receive  $s, a, s', r$ 
observe  $s$ , take action  $a$  by  $\epsilon$ -greedy w.r.t.  $Q(s, \cdot)$ 
execute  $a$ , observe reward  $R$  and next state  $s'$ 
Q-learning update for  $Q(s, a)$ 
update model  $\mathcal{M}(s, a)$  (i.e. by counting)
put  $(s, a)$  into search-control queue
for  $i=1:n$  do
    sample  $(\tilde{s}, \tilde{a})$  from search-control queue
     $(\tilde{s}', \tilde{R}) \leftarrow \mathcal{M}(\tilde{s}, \tilde{a})$  // simulated transition
    Q-learning update for  $Q(\tilde{s}, \tilde{a})$  // planning update

```

A.2 Experimental details

Implementation details of common settings. The Grid-World domain is written by ourselves, all other discrete action domains are from OpenAI Gym [Brockman *et al.*, 2016] with version 0.8.2. The exact environment names we used are: MountainCar-v0, CartPole-v1, Acrobot-v1. Deep learning implementation is based on tensorflow with version 1.1.0 [Abadi *et al.*, 2015]. On all domains, we use Adam optimizer, Xavier initializer, set mini-batch size $b = 32$, buffer size 100k. All activation functions are ReLU except the output layer of the Q -value is linear, and the output layer of the actor network is tanh. The output layer parameters were initialized from a uniform distribution $[-0.0003, 0.0003]$, all other parameters are initialized using Xavier initialization [Glorot and Bengio, 2010].

As for model learning, we learn a difference model to alleviate the effect of outliers, that is, we learn a neural network model with input s_t and output $s_{t+1} - s_t$. The neural network has two 64 units hidden ReLU-layers. The model is learned in an online manner and by using samples from ER buffer with a fixed learning rate as 0.0001 and mini-batch size 128 across all experiments.

Termination condition on OpenAI environments. On OpenAI, each environment has a time limit and the termination flag will be true if either the time limit reached or the actual termination condition satisfied. However, theoretically we should truncate the return if and only if the actual termination condition satisfied. All of our experiments are conducted by setting discount rate $\gamma = 0.0$ if and only if the actual termination condition satisfied. For example, on mountain car, $done = true$ if and only if the position ≥ 0.5 .

Algorithm 3 (DQN-)HC-Dyna with additional details

```

 $B_s$ : search-control queue,  $B$ : the experience replay buffer
 $\mathcal{M} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S} \times \mathbb{R}$ , the environment model
 $k$ : number of HC steps,  $\alpha$ : gradient ascent step size
 $\rho \in [0, 1]$ : mixing factor in a mini-batch, i.e.  $\rho b$  samples in a mini-batch are simulated from model
 $\epsilon_a$ : threshold for whether one should add a state to  $B_s$ 
 $n$ : number of planning steps
 $Q, Q'$ : current and target Q networks, respectively
 $b$ : the mini-batch size
 $\tau$ : update target network  $Q'$  every  $\tau$  updates to  $Q$ 
 $t \leftarrow 0$  is the time step
 $n_\tau \leftarrow 0$  is the number of parameter updates
 $d$ : number of state variables, i.e.  $\mathcal{S} \subset \mathbb{R}^d$ 
 $\hat{\Sigma}_s$  is the empirical covariance matrix of states
 $\mu_{ss} \leftarrow \mathbf{0} \in \mathbb{R}^{d \times d}, \mu_s \leftarrow \mathbf{0} \in \mathbb{R}^d$  (assistant variables for computing empirical covariance matrix, sample average will be maintained for  $\mu_{ss}, \mu_s$ )
 $\eta$  is the scalar to scale the empirical covariant matrix
 $\mathcal{N}$  is the multivariate gaussian distribution used as injected noise
To reproduce experiment:  $\rho = 0.5, b = 32, k = 100, \eta = 0.1, \tau = 1000, \alpha = 0.1$ , DQN learning rate 0.0001
while true do
    Observe  $s_t$ , take action  $a_t$  (i.e.  $\epsilon$ -greedy w.r.t.  $Q$ )
    Observe  $s_{t+1}, r_{t+1}$ , add  $(s_t, a_t, s_{t+1}, r_{t+1})$  to  $B$ 
     $\mu_{ss} \leftarrow \frac{\mu_{ss}(t-1) + s_t s_t^\top}{t}, \mu_s \leftarrow \frac{\mu_s(t-1) + s_t}{t}$ 
     $\hat{\Sigma}_s \leftarrow \mu_{ss} - \mu_s \mu_s^\top$ 
     $\epsilon_a \leftarrow (1 - \beta)\epsilon_a + \beta \frac{\|s_{t+1} - s_t\|_2}{\sqrt{d}}$  for  $\beta = 0.001$ 
    // Gradient ascent hill climbing
    sample  $s_0$  from  $B, \tilde{s} \leftarrow \infty$ 
    for  $i = 0, \dots, k$  do
         $g_{s_i} \leftarrow \nabla_s V(s_i) = \nabla_s \max_a Q(s_i, a)$ 
         $s_{i+1} \leftarrow \Pi(s_i + \frac{\alpha}{\|\hat{\Sigma}_s g_{s_i}\|} \hat{\Sigma}_s g_{s_i} + X_i), X_i \sim \mathcal{N}(0, \eta \hat{\Sigma}_s)$ 
    if distance( $\tilde{s}, s_{i+1}$ )  $\geq \epsilon_a$  then
        Add  $s_{i+1}$  into  $B_{sc}, \tilde{s} \leftarrow s_{i+1}$ 
    //  $n$  planning updates: sample  $n$  mini-batches
    for  $n$  times do
        Sample  $\rho b$  states from  $B_s$ , for each state  $\hat{s}$ , take action  $\hat{a}$  according to  $\epsilon$ -greedy w.r.t  $Q$ 
        For each state action pairs  $(\hat{s}, \hat{a})$ , query the model:  $\hat{s}', \hat{r} \leftarrow \mathcal{M}(\hat{s}, \hat{a})$ , call this Mini-batch-1
        // then sample from ER buffer
        Sample Mini-batch-2 with size  $b - \rho b$  from  $B$ , stack Mini-batch-1, Mini-batch-2 into Mini-batch, hence Mini-batch has size  $b$ 
        (DQN) parameter update with mini-batch  $b_m$ 
         $n_\tau \leftarrow n_\tau + 1$ 
        if  $mod(n_\tau, \tau) == 0$  then
             $Q' \leftarrow Q$ 
        // update the neural network of the model  $\mathcal{M}$  if applicable
     $t \leftarrow t + 1$ 

```

Experimental details of TabularGridWorld domain. The purpose of using the tabular domain is to study the learning performances by using different sampling distribution to fill the search-control queue. Our TabularGridWorld is similar to the continuous state domain introduced in 1(a) except that we do not have a wall and we introduce stochasticity to make it more representative. Four actions are available and can take the agent to the next $\{up, down, left, right\}$ grid respectively. An action can be executed successfully with probability 0.8 otherwise a random action is taken. The TabularGridWorld size is 20×20 and each episode start from left-bottom grid and would terminate if reached the right-top grid or 1k time steps. The return will not be truncated unless the right-top grid is reached. The discount rate is $\gamma = 1.0$. For all algorithms, we fixed the exploration noise as $\epsilon = 0.2$ and sweep over learning rate $\{2^0, 2^{-0.25}, 2^{-0.5}, 2^{-0.75}, 2^{-1}, 2^{-1.5}, 2^{-2.0}, 2^{-2.5}\} = \{1.0, 0.8409, 0.70711, 0.59460, 0.5, 0.35356, 0.25, 0.17678\}$. We fix using exploration noise $\epsilon = 0.2$ and mixing rate $\rho = 0.5$. We use 10 planning steps for all algorithms. We evaluate each algorithm every 100 environment time steps. Parameter is optimized by using the last 20% evaluation episodes to ensure convergence.

For our algorithm HC-Dyna, we do not sweep any additional parameters. We fix doing 80 gradient ascent steps per environment time step and the injected noise is gaussian $\mathcal{N}(0, 0.05)$. When adding the noise or using finite difference method for computing gradient, we logically regard the domain as $[0, 1]^2$ and hence each grid is a square with length $1/20 = 0.05$. Specifically, when in a grid, we find its corresponding center's x, y coordinates as its location to add noise. As for gradient ascent with finite difference approximation, given a state s , we compute the value increasing rate from each of its 8 neighbors and pick up the one with largest increasing rate as the next state. That is, $s \leftarrow \arg \max_{s'} \frac{\hat{V}(s') - \hat{V}(s)}{\|s - s'\|}$. Both the search-control queue size and ER buffer size are setted as $1e5$.

The optimal value function used for HC-Dyna-Vstar and Gibbs-Vstar on this domain is acquired by taking the value function at the end of training ER for $1e6$ steps and averaged over 50 random seeds.

Experimental details of continuous state domains. All continuous state domain, we set discount rate $\gamma = 0.99$. We set the episode length limit as 2000 for both GridWorld and MountainCar, while keep other domains as the default setting. We use warmup steps 5000 for all algorithms.

For all Q networks, we consistently use a neural network with two 32 units hidden ReLU-layers. We use target network moving frequency $\tau = 1000$ and sweep learning rate $\{0.001, 0.0001, 0.00001\}$ for vanilla DQN with ER with planning step 5, then we directly use the same best learning rate (0.0001) for all other experiments. For our particular parameters, we fixed the same setting across all domains: mixing rate 0.5 and ϵ_a is sample average, number of gradient steps $k = 100$ with gradient ascent step size 0.1 and queue size $1e6$. We incrementally update the empirical covariance matrix. When evaluating each algorithm, we keep a small noise $\epsilon = 0.05$ when taking action and evaluate one episode every 1000 environment time steps for each run.

References

- [Abadi *et al.*, 2015] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, and et al. TensorFlow: Large-scale machine learning on heterogeneous systems. 2015. Software available from tensorflow.org.
- [Adam *et al.*, 2012] Sander Adam, Lucian Busoniu, and Robert Babuska. Experience replay for real-time reinforcement learning control. *IEEE Transactions on Systems, Man, and Cybernetics*, pages 201–212, 2012.
- [Amari and Douglas, 1998] Shun-Ichi Amari and Scott C. Douglas. Why natural gradient? *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1213–1216, 1998.
- [Amari, 1998] Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.
- [Brockman *et al.*, 2016] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym. arXiv:1606.01540, 2016.
- [Chiang *et al.*, 1987] Tzuu-Shuh Chiang, Chii-Ruey Hwang, and Shuenn Jyi Sheu. Diffusion for global optimization in \mathbb{R}^n . *SIAM Journal on Control and Optimization*, pages 737–753, 1987.
- [Corneil *et al.*, 2018] Dane S. Corneil, Wulfram Gerstner, and Johanni Brea. Efficient model-based deep reinforcement learning with variational state tabulation. *ICML*, pages 1049–1058, 2018.
- [Durmus and Moulines, 2017] Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, pages 1551–1587, 2017.
- [Glorot and Bengio, 2010] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2010.
- [Gu *et al.*, 2016] Shixiang Gu, Timothy P. Lillicrap, Ilya Sutskever, and Sergey Levine. Continuous Deep Q-Learning with Model-based Acceleration. In *ICML*, pages 2829–2838, 2016.
- [Holland *et al.*, 2018] G. Zacharias Holland, Erik Talvitie, and Michael Bowling. The effect of planning shape on dyna-style planning in high-dimensional state spaces. *CoRR*, abs/1806.01825, 2018.
- [Lillicrap *et al.*, 2016] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *ICLR*, 2016.
- [Lin, 1992] Long-Ji Lin. Self-Improving Reactive Agents Based On Reinforcement Learning, Planning and Teaching. *Machine Learning*, 1992.
- [Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, and et al. Human-level control through deep reinforcement learning. *Nature*, 2015.
- [Moore and Atkeson, 1993] Andrew W. Moore and Christopher G. Atkeson. Prioritized sweeping: Reinforcement learning with less data and less time. *Machine learning*, pages 103–130, 1993.
- [Pan *et al.*, 2018] Yangchen Pan, Muhammad Zaheer, Adam White, Andrew Patterson, and Martha White. Organizing experience: a deeper look at replay mechanisms for sample-based planning in continuous state domains. In *IJCAI*, pages 4794–4800, 2018.
- [Peng and Williams, 1993] Jing Peng and Ronald J Williams. Efficient learning and planning within the Dyna framework. *Adaptive behavior*, 1993.
- [Peng *et al.*, 2018] Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, Kam-Fai Wong, and Shang-Yu Su. Deep Dyna-Q: Integrating planning for task-completion dialogue policy learning. In *Annual Meeting of the Association for Computational Linguistics*, pages 2182–2192, 2018.
- [Roberts, 1996] Richard L. Roberts, Gareth O. and Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.
- [Santos *et al.*, 2012] Matilde Santos, José Antonio Martín H., Victoria López, and Guillermo Botella. Dyna-H: A heuristic planning reinforcement learning algorithm applied to role-playing game strategy decision systems. *Knowledge-Based Systems*, 32:28–36, 2012.
- [Schaul *et al.*, 2016] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized Experience Replay. In *ICLR*, 2016.
- [Silver *et al.*, 2014] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *ICML*, pages I–387–I–395, 2014.
- [Sutton and Barto, 2018] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.
- [Sutton *et al.*, 2008] Richard S. Sutton, Csaba Szepesvári, Alborz Geramifard, and Michael Bowling. Dyna-style planning with linear function approximation and prioritized sweeping. In *UAI*, pages 528–536, 2008.
- [Sutton, 1990] Richard S. Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *ML*, 1990.
- [Sutton, 1991] Richard S. Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *SIGART Bulletin*, 2(4):160–163, 1991.
- [Thomas *et al.*, 2016] Philip Thomas, Bruno Castro Silva, Christoph Dann, and Emma Brunskill. Energetic natural gradient descent. In *ICML*, pages 2887–2895, 2016.
- [Tsitsiklis, 1994] John N. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, pages 185–202, 1994.
- [van Seijen and Sutton, 2015] Harm van Seijen and Richard S. Sutton. A deeper look at planning as learning from replay. In *ICML*, pages 2314–2322, 2015.

[Wawrzyński and Tanwani, 2013] Paweł Wawrzyński and Ajay Kumar Tanwani. Autonomous reinforcement learning with experience replay. *Neural Networks*, pages 156–167, 2013.

[Welling and Teh, 2011] Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *ICML*, pages 681–688, 2011.