

---

# Search on the Replay Buffer: Bridging Planning and Reinforcement Learning

---

Benjamin Eysenbach <sup>$\theta\phi$</sup> , Ruslan Salakhutdinov <sup>$\theta$</sup> , Sergey Levine <sup>$\phi\psi$</sup>   
 <sup>$\theta$</sup> CMU,  <sup>$\phi$</sup> Google Brain,  <sup>$\psi$</sup> UC Berkeley  
 beysenba@cs.cmu.edu

## Abstract

The history of learning for control has been an exciting back and forth between two broad classes of algorithms: planning and reinforcement learning. Planning algorithms effectively reason over long horizons, but assume access to a local policy and distance metric over collision-free paths. Reinforcement learning excels at learning policies and the relative values of states, but fails to plan over long horizons. Despite the successes of each method in various domains, tasks that require reasoning over long horizons with limited feedback and high-dimensional observations remain exceedingly challenging for both planning and reinforcement learning algorithms. Frustratingly, these sorts of tasks are potentially the most useful, as they are simple to design (a human only need to provide an example goal state) and avoid reward shaping, which can bias the agent towards finding a sub-optimal solution. We introduce a general-purpose control algorithm that combines the strengths of planning and reinforcement learning to effectively solve these tasks. Our aim is to decompose the task of reaching a distant goal state into a sequence of easier tasks, each of which corresponds to reaching a particular subgoal. Planning algorithms can automatically find these waypoints, but only if provided with suitable abstractions of the environment – namely, a graph consisting of nodes and edges. Our main insight is that this graph can be constructed via reinforcement learning, where a goal-conditioned value function provides edge weights, and nodes are taken to be previously seen observations in a replay buffer. Using graph search over our replay buffer, we can automatically generate this sequence of subgoals, even in image-based environments. Our algorithm, search on the replay buffer (SoRB), enables agents to solve sparse reward tasks over one hundred steps, and generalizes substantially better than standard RL algorithms.<sup>1</sup>

## 1 Introduction

How can agents learn to solve complex, temporally extended tasks? Classically, planning algorithms give us one tool for learning such tasks. While planning algorithms work well for tasks where it is easy to determine distances between states and easy to design a local policy to reach nearby states, both of these requirements become roadblocks when applying planning to high-dimensional (e.g., image-based) tasks. Learning algorithms excel at handling high-dimensional observations, but reinforcement learning (RL) – learning for control – fails to reason over long horizons to solve temporally extended tasks. In this paper, we propose a method that combines the strengths of planning and RL, resulting in an algorithm that can plan over long horizons in tasks with high-dimensional observations.

Recent work has introduced goal-conditioned RL algorithms (Pong et al., 2018; Schaul et al., 2015) that acquire a single policy for reaching many goals. In practice, goal-conditioned RL succeeds at

---

<sup>1</sup>Run our algorithm in your browser: [http://bit.ly/rl\\_search](http://bit.ly/rl_search)

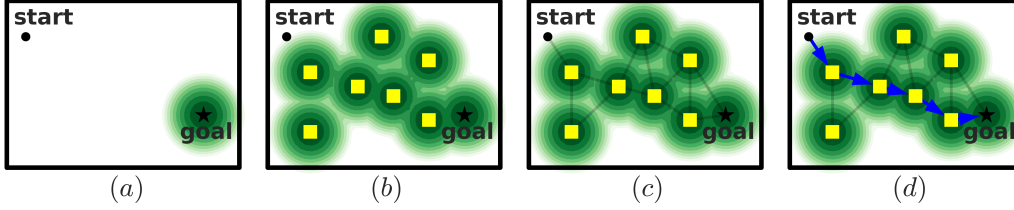


Figure 1: **Search on the Replay Buffer:** (a) Goal-conditioned RL often fails to reach distant goals, but can successfully reach the goal if starting nearby (inside the green region). (b) Our goal is to use observations in our replay buffer (yellow squares) as waypoints leading to the goal. (c) We automatically find these waypoints by using the agent’s value function to predict when two states are nearby, and building the corresponding graph. (d) We run graph search to find the sequence of waypoints (blue arrows), and then use our goal-conditioned policy to reach each waypoint.

reaching nearby goals but fails to reach distant goals; performance degrades quickly as the number of steps to the goal increases (Levy et al., 2019; Nachum et al., 2018). Moreover, goal-conditioned RL often requires large amounts of reward shaping (Chiang et al., 2019) or human demonstrations (Lynch et al., 2019; Nair et al., 2018), both of which can limit the asymptotic performance of the policy by discouraging the policy from seeking novel solutions.

We propose to solve long-horizon, sparse reward tasks by decomposing the task into a series of easier goal-reaching tasks. We learn a goal-conditioned policy for solving each of the goal-reaching tasks. Our main idea is to reduce the problem of finding these subgoals to solving a shortest path problem over states that we have previously visited, using a distance metric extracted from our goal-conditioned policy. We call this algorithm Search on Replay Buffer (SoRB), and provide a simple illustration of the algorithm in Figure 1.

Our primary contribution is an algorithm that bridges planning and deep RL for solving long-horizon, sparse reward tasks. We develop a practical instantiation of this algorithm using ensembles of distributional value functions, which allows us to *robustly* learn distances and use them for *risk-aware* planning. Empirically, we find that our method generates effective plans to solve long horizon navigation tasks, even in image-based domains, without a map and without odometry. Comparisons with state-of-the-art RL methods show that SoRB is substantially more successful in reaching distant goals. We also observe that the learned policy generalizes well to navigate in unseen environments. In summary, graph search over previously visited states is a simple tool for boosting the performance of a goal-conditioned RL algorithm.

## 2 Bridging Planning and Reinforcement Learning

Planning algorithms must be able to (1) sample valid states, (2) estimate the distance between reachable pairs of states, and (3) use a local policy to navigate between nearby states. These requirements are difficult to satisfy in complex tasks with high dimensional observations, such as images. For example, consider a robot arm stacking blocks using image observations. Sampling states requires generating photo-realistic images, and estimating distances and choosing actions requires reasoning about dozens of interactions between blocks. Our method will obtain distance estimates and a local policy using a RL algorithm. To sample states, we will simply use a replay buffer of previously visited states as a non-parametric generative model.

### 2.1 Building Block: Goal-Conditioned RL

A key building block of our method is a goal-conditioned policy and its associated value function. We consider a goal-reaching agent interacting with an environment. The agent observes its current state  $s \in \mathcal{S}$  and a goal state  $s_g \in \mathcal{S}$ . The initial state for each episode is sampled  $s_1 \sim \rho(s)$ , and dynamics are governed by the distribution  $p(s_{t+1} | s_t, a_t)$ . At every step, the agent samples an action  $a \sim \pi(a | s, s_g)$  and receives a corresponding reward  $r(s, a, s_g)$  that indicates whether the agent has reached the goal. The episode terminates as soon as the agent reaches the goal, or after  $T$  steps, whichever occurs first. The agent’s task is to maximize its cumulative, *undiscounted*, reward. We use an off-policy algorithm to learn such a policy, as well as its associated goal-conditioned Q-function

and value function:

$$Q(s, a, s_g) = \mathbb{E}_{\substack{s_1 \sim p(s), a_t \sim \pi(a_t | s_t, s_g) \\ s_{t+1} \sim p(s_{t+1} | s_t, a_t)}} \left[ \sum_{t=1}^T r(s_t, s_g, a_t) \right], \quad V(s, s_g) = \max_a Q(s, a, s_g)$$

We obtain a policy by acting greedily w.r.t. the Q-function:  $\pi(a | s, s_g) = \arg \max_a Q(s, a, s_g)$ . We choose an off-policy RL algorithm with goal relabelling (Andrychowicz et al., 2017; Kaelbling, 1993b) and distributional RL (Bellemare et al., 2017)) not only for improved data efficiency, but also to obtain good distance estimates (See Section 2.2). We will use DQN (Mnih et al., 2013) for discrete action environments and DDPG (Lillicrap et al., 2015) for continuous action environments. Both algorithms operate by minimizing the Bellman error over transitions sampled from a replay buffer  $\mathcal{B}$ .

## 2.2 Distances from Goal-Conditioned Reinforcement Learning

To ultimately perform planning, we need to compute the *shortest path distance* between pairs of states. Following Kaelbling (1993b), we define a reward function that returns -1 at every step:  $r(s, a, s_g) \triangleq -1$ . The episode ends when the agent is sufficiently close to the goal, as determined by a state-identity oracle. Using this reward function and termination condition, there is a close connection between the Q values and shortest paths. We define  $d_{sp}(s, s_g)$  to be the shortest path distance from state  $s$  to state  $s_g$ . That is,  $d_{sp}(s, s_g)$  is the expected number of steps to reach  $s_g$  from  $s$  under the optimal policy. The value of state  $s$  with respect to goal  $s_g$  is simply the negative shortest path distance:  $V(s, s_g) = -d_{sp}(s, s_g)$ . We likewise define  $d_{sp}(s, a, s_g)$  as the shortest path distance, conditioned on initially taking action  $a$ . Then Q values also equal a negative shortest path distance:  $Q(s, a, s_g) = -d_{sp}(s, a, s_g)$ . Thus, goal-conditioned RL on a suitable reward function yields a Q-function that allows us to estimate shortest-path distances.

## 2.3 The Replay Buffer as a Graph

We build a weighted, *directed* graph directly on top of states in our replay buffer, so each node corresponds to an observation (e.g., an image). We add edges between nodes with weight (i.e., length) equal to their predicted distance, but ignore edges that are longer than MAXDIST, a hyperparameter:

$$\mathcal{G} \triangleq (\mathcal{V}, \mathcal{E}, \mathcal{W}) \quad \text{where} \quad \mathcal{V} = \mathcal{B}, \quad \mathcal{E} = \mathcal{B} \times \mathcal{B} = \{e_{s_1 \rightarrow s_2} \mid s_1, s_2 \in \mathcal{B}\}$$

$$\mathcal{W}(e_{s_1 \rightarrow s_2}) = \begin{cases} d_\pi(s_1, s_2) & \text{if } d_\pi(s_1, s_2) < \text{MAXDIST} \\ \infty & \text{otherwise} \end{cases}$$

Given a start and goal state, we temporarily add each to the graph. We add directed edges from the start state to every other state, and from every other state to the goal state, using the same criteria as above. We use Dijkstra’s Algorithm to find the shortest path. See Appendix A for details.

## 2.4 Algorithm Summary

After learning a goal-conditioned Q-function, we perform graph search to find a set of waypoints and use the goal-conditioned policy to reach each. We view the combination of graph search and the underlying goal-conditioned policy as a new SEARCHPOLICY, shown in Algorithm 1. The algorithm starts by using graph search to obtain the shortest path  $s_{w_1}, s_{w_2}, \dots$  from the current state  $s$  to the goal state  $s_g$ , planning over the states in our replay buffer  $\mathcal{B}$ . We then estimate the distance from the current state to the first waypoint, as well as the distance from the current state to the goal. In most cases, we then condition the policy on the first waypoint,  $s_{w_1}$ . However, if the goal state is closer than the next waypoint and the goal state is not too far away, then we directly condition the policy on the final goal. If the replay buffer is empty or there is not a path in  $\mathcal{G}$  to the goal, then Algorithm 1 resorts to standard goal-conditioned RL.

**Algorithm 1** Inputs are the current state  $s$ , the goal state  $s_g$ , a buffer of observations  $\mathcal{B}$ , the learned policy  $\pi$  and its value function  $V$ . Returns an action  $a$ .

---

```

function SEARCHPOLICY( $s, s_g, \mathcal{B}, V, \pi$ )
   $s_{w_1}, \dots \leftarrow \text{SHORTESTPATH}(s, s_g, \mathcal{B}, V)$ 
   $d_{s \rightarrow w_1} \leftarrow -V(s, s_{w_1})$ 
   $d_{s \rightarrow g} \leftarrow -V(s, s_g)$ 
  if  $d_{s \rightarrow w_1} < d_{s \rightarrow g}$  or  $d_{s \rightarrow g} > \text{MAXDIST}$ 
     $a \leftarrow \pi(a, \mid s, s_{w_1})$ 
  else
     $a \leftarrow \pi(a, \mid s, s_g)$ 
  return  $a$ 

```

---

### 3 Better Distance Estimates

The success of our SEARCHPOLICY depends heavily on the accuracy of our distance estimates. This section proposes two techniques to learn better distances with RL.

#### 3.1 Better Distances via Distributional Reinforcement Learning

Off-the-shelf Q-learning algorithms such as DQN (Mnih et al., 2013) or DDPG (Lillicrap et al., 2015) will fail to learn accurate distance estimates using the  $-1$  reward function. The true value for a state and goal that are unreachable is  $-\infty$ , which cannot be represented by a standard, feed-forward Q-network. Simply clipping the Q-value estimates to be within some range avoids the problem of ill-defined Q-values, but empirically we found it challenging to train clipped Q-networks. We adopt distributional Q-learning (Bellemare et al., 2017), noting that it has a convenient form when used with the  $-1$  reward function. Distributional RL discretizes the possible value estimates into a set of bins  $B = (B_1, B_2, \dots, B_N)$ . For learning distances, bins correspond to distances, so  $B_i$  indicates the event that the current state and goal are  $i$  steps away from one another. Our Q-function predicts a distribution  $Q(s_t, s_g, a_t) \in \mathcal{P}^N$  over these bins, where  $Q(s_t, s_g, a_t)_i$  is the predicted probability that states  $s_t$  and  $s_g$  are  $i$  steps away from one another. To avoid ill-defined Q-values, the final bin,  $B_N$  is a catch-all for predicted distances of at least  $N$ . Importantly, this gives us a well-defined method to represent large and infinite distances. Under this formulation, the targets  $Q^* \in \mathcal{P}^N$  for our Q-values have a simple form:

$$Q^* = \begin{cases} (1, 0, \dots, 0) & \text{if } s_t = g \\ (0, Q_1, \dots, Q_{N-2}, Q_{N-1} + Q_N) & \text{if } s_t \neq g \end{cases}$$

As illustrated in Figure 2, if the state and goal are equivalent, then the target places all probability mass in bin 0. Otherwise, the targets are a right-shift of the current predictions. To ensure the target values sum to one, the mass in bin  $N$  of the targets is the sum of bins  $N-1$  and  $N$  from the predicted values. Following Bellemare et al. (2017), we update our Q function by minimizing the KL divergence between our predictions  $Q^\theta$  and the target  $Q^*$ :

$$\min_{\theta} D_{\text{KL}}(Q^* \parallel Q^\theta) \quad (1)$$

#### 3.2 Robust Distances via Ensembles of Value Functions

Since we ultimately want to use estimated distances to perform search, it is crucial that we have accurate distances estimates. It is challenging to robustly estimate the distance between all  $|\mathcal{B}|^2$  pairs of states in our buffer  $\mathcal{B}$ , some of which may not have occurred during training. If we fail and spuriously predict that a pair of distant states are nearby, graph search will exploit this “wormhole” and yield a path which assumes that the agent can “teleport” from one distant state to another. We seek to use a bootstrap (Bickel et al., 1981) as a principled way to estimate uncertainty for our Q-values. Following prior work (Lakshminarayanan et al., 2017; Osband et al., 2016), we implement an approximation to the bootstrap. We train an ensemble of Q-networks, each with independent weights, but trained on the same data using the same loss (Eq. 1). When performing graph search, we aggregate predictions from each Q-network in our ensemble. Empirically, we found that ensembles were crucial for getting graph search to work on image-based tasks, but we observed little difference in whether we took the maximum predicted distance or the average predicted distance.

### 4 Related Work

*Planning Algorithms:* Planning algorithms (Choset et al., 2005; LaValle, 2006) efficiently solve long-horizon tasks, including those that stymie RL algorithms (see, e.g., Kavraci et al. (1996); Lau and

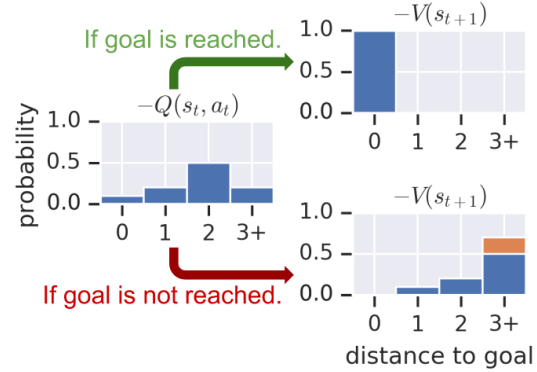


Figure 2: The Bellman update for distributional RL is simple when learning distances, simply corresponding to a left-shift of the Q-values at every step until the agent reaches the goal.

Kuffner (2005); Levine et al. (2011)). However, these techniques assume that we can (1) efficiently sample valid states, (2) estimate the distance between two states, and (3) acquire a local policy for reaching nearby states, all of which make it challenging to apply these techniques to high-dimensional tasks (e.g., with image observations). Our method removes these assumptions by (1) sampling states from the replay buffer and (2,3) learning the distance metric and policy with RL. Some prior works have also combined planning algorithms with RL (Chiang et al., 2019; Faust et al., 2018; Savinov et al., 2018a), finding that the combination yields agents adept at reaching distant goals. Perhaps the most similar work is Semi-Parametric Topological Memory (Savinov et al., 2018a), which also uses graph search to find waypoints for a learned policy. We compare to SPTM in Section 5.3.

**Goal-Conditioned RL:** Goal-conditioned policies (Kaelbling, 1993b; Pong et al., 2018; Schaul et al., 2015) take as input the current state and a goal state, and predict a sequence of actions to arrive at the goal. Our algorithm learns a goal-conditioned policy to reach waypoints along the planned path. Recent algorithms (Andrychowicz et al., 2017; Pong et al., 2018) combine off-policy RL algorithms with goal-relabelling to improve the sample complexity and robustness of goal-conditioned policies. Similar algorithms have been proposed for visual navigation (Anderson et al., 2018; Gupta et al., 2017; Mirowski et al., 2016; Zhu et al., 2017). A common theme in recent work is learning distance metrics to accelerate RL. While most methods (Florensa et al., 2019; Savinov et al., 2018b; Wu et al., 2018) simply perform RL on top of the learned representation, our method explicitly performs search using the learned metric.

**Hierarchical RL:** Hierarchical RL algorithms automatically learn a set of primitive skills to help an agent learn complex tasks. One class of methods (Bacon et al., 2017; Frans et al., 2017; Kaelbling, 1993a; Kulkarni et al., 2016; Nachum et al., 2018; Parr and Russell, 1998; Precup, 2000; Sutton et al., 1999; Vezhnevets et al., 2017) jointly learn a low-level policy for performing each of the skills together with a high-level policy for sequencing these skills to complete a desired task. Another class of algorithms (Drummond, 2002; Fox et al., 2017; Şimşek et al., 2005) focus solely on automatically discovering these skills or subgoals. SoRB learns primitive skills that correspond to goal-reaching tasks, similar to Nachum et al. (2018). While jointly learning high-level and low-level policies can be unstable (see discussion in Nachum et al. (2018)), we sidestep the problem by using graph search as a fixed, high-level policy.

**Model Based RL:** RL methods are typically divided into model-free (Schulman et al., 2015a,b, 2017; Williams, 1992) and model-based (Lillicrap et al., 2015; Watkins and Dayan, 1992) approaches. Model-based approaches all perform some degree of planning, from predicting the value of some state (Mnih et al., 2013; Silver et al., 2016), obtaining representations by unrolling a learned dynamics model (Racanière et al., 2017), or learning a policy directly on a learned dynamics model (Agrawal et al., 2016; Chua et al., 2018; Finn and Levine, 2017; Kurutach et al., 2018; Nagabandi et al., 2018; Oh et al., 2015; Sutton, 1990). One line of work (Amos et al., 2018; Lee et al., 2018; Srinivas et al., 2018; Tamar et al., 2016) embeds a differentiable planner inside a policy, with the planner learned end-to-end with the rest of the policy. Other work (Lenz et al., 2015; Watter et al., 2015) explicitly learns a representation for use inside a standard planning algorithm. In contrast, SoRB learns to predict the distances between states, which can be viewed as a high-level inverse model. SoRB predicts a scalar (the distance) rather than actions or observations, making the prediction problem substantially easier. By planning over previously visited states, SoRB does not have to cope with infeasible states that can be predicted by forward models in state-space and latent-space.

model	real states	multi-step	prediction dimension
state-space	✓	✓	1000s+
latent-space	✗	✓	10s
inverse	✓	✗	10s
SoRB	✓	✓	1

Figure 3: Four classes of model-based RL methods. Dimensions in the last column correspond to typical robotics tasks with image/lidar observations.

## 5 Experiments

We compare SoRB to prior methods on two tasks: a simple 2D environment, and then a visual navigation task, where our method will plan over images. Ablation experiments will illustrate that accurate distances estimates are crucial to our algorithm’s success.



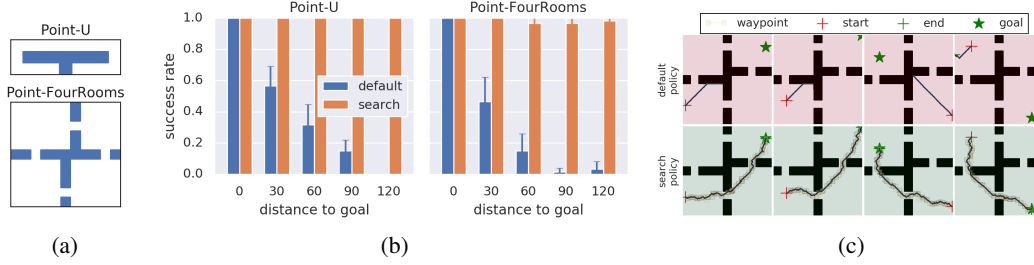


Figure 4: **Simple 2D Navigation:** (Left) Two simple navigation environments. (Center) An agent that combines a goal-conditioned policy with search is substantially more successful at reaching distant goals in these environments than using the goal-conditioned policy alone. (Right) A standard goal-conditioned policy (top) fails to reach distant goals. Applying graph search on top of that *same* policy (bottom) yields a sequence of intermediate waypoints (yellow squares) that enable the agent to successfully reach distant goals.

## 5.1 Didactic Example: 2D Navigation

We start by building intuition for our method by applying it to two simple 2D navigation tasks, shown in Figure 4a. The start and goal state are chosen randomly in free space, and reaching the goal often takes over 100 steps, even for the optimal policy. We used goal-conditioned RL to learn a policy for each environment, and then evaluated this policy on randomly sampled (start, goal) pairs of varying difficulty. To implement SoRB, we used exactly the same policy, both to perform graph search and then to reach each of the planned waypoints. In Figure 4b, we observe that the goal-conditioned policy can reach nearby goals, but fails to generalize to distant goals. In contrast, SoRB successfully reaches goals over 100 steps away, with little drop in success rate. Figure 4c compares rollouts from the goal-conditioned policy and our policy. Note that our policy takes actions that temporarily lead away from the goal so the agent can maneuver through a hallway to eventually reach the goal.

## 5.2 Planning over Images for Visual Navigation

We now examine how our method scales to high-dimensional observations in a visual navigation task, illustrated in Figure 5. We use 3D houses from the SUNCG dataset (Song et al., 2017), similar to the task described by Shah et al. (2018). The agent receives either RGB or depth images and takes actions to move North/South/East/West. Following Shah et al. (2018), we stitch four images into a panorama, so the resulting observation has dimension  $4 \times 24 \times 32 \times C$ , where  $C$  is the number of channels (3 for RGB, 1 for Depth). At the start of each episode, we randomly sample an initial state and goal state. We found that sampling nearby goals (within 4 steps) more often (80% of the time) improved the performance of goal-conditioned

RL. We use the same goal sampling distribution for all methods. The agent observes both the current image and the goal image, and should take actions that lead to the goal state. The episode terminates once the agent is within 1 meter of the goal. We also terminate if the agent has failed to reach the goal after 20 time steps, but treat the two types of termination differently when computing the TD error (see Pardo et al. (2017)). Note that it is challenging to specify a meaningful distance metric and local policy on pixel inputs, so it is difficult to apply standard planning algorithms to this task.

On this task, we evaluate four state-of-the-art prior methods: hindsight experience replay (HER) (Andrychowicz et al., 2017), distributional RL (C51) (Bellemare et al., 2017), semi-parametric topological memory (SPTM) (Savinov et al., 2018a), and value iteration networks (VIN) (Tamar et al., 2016). SoRB uses C51 as its underlying goal-conditioned policy. For VIN, we tuned the number of iterations as well as the number of hidden units in the recurrent layer. For SPTM, we performed a grid search over the threshold for adding edges, the threshold for choosing the next waypoint along



Figure 5: **Visual Navigation:** Given an initial state and goal state, our method automatically finds a sequence of intermediate waypoints. The agent then follows those waypoints to reach the goal.

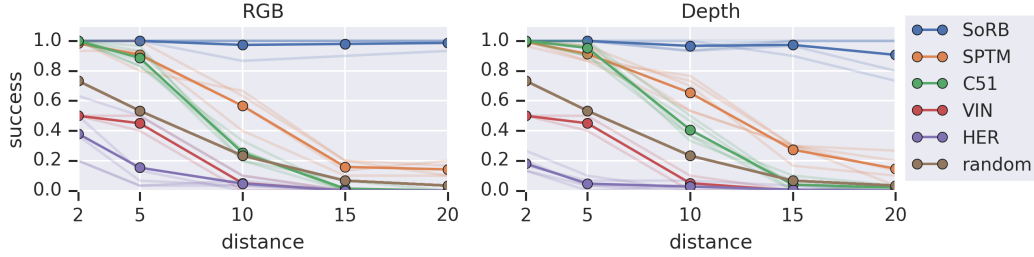


Figure 6: **Visual Navigation:** We compare our method (SoRB) to prior work on the visual navigation environment (Fig. 5), using RGB images (*Left*) and depth images (*Right*). We find that only our method succeeds in reaching distant goals. *Baselines:* SPTM (Savinov et al., 2018a), C51 (Bellemare et al., 2017), VIN (Tamar et al., 2016), HER (Andrychowicz et al., 2017).

the shortest path, and the parameters for sampling the training data. In total, we performed over 1000 experiments to tune baselines, more than an order of magnitude more than we used for tuning our own method. See Appendix F for details.

We evaluated each method on goals ranging from 2 to 20 steps from the start. For each distance, we randomly sampled 30 (start, goal) pairs, and recorded the average success rate, defined as reaching within 1 meter of the goal within 100 steps. We then repeated each experiment for 5 random seeds. In Figure 6, we plot each random seed as a transparent line; the solid line corresponds to the average across the 5 random seeds. While all prior methods degrade quickly as the distance to the goal increases, our method continues to succeed in reaching goals with probability around 90%. SPTM, the only prior method that also employs search, performs second best, but substantially worse than our method.

### 5.3 Comparison with Semi-Parametric Topological Memory

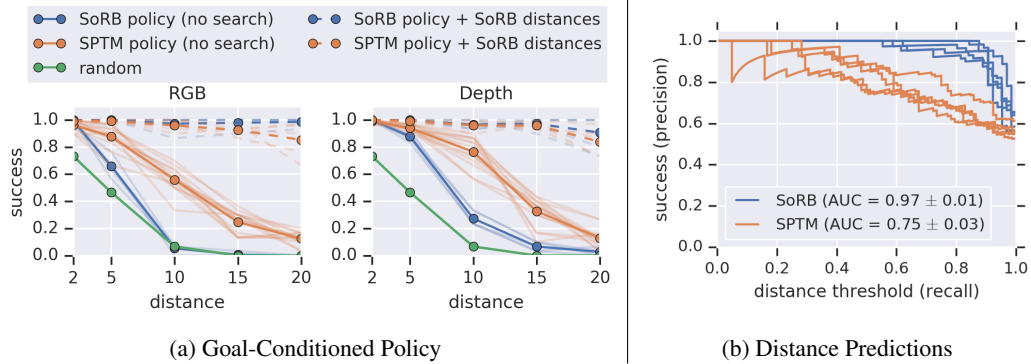


Figure 7: **SoRB vs SPTM:** Our method and Semi-Parametric Topological Memory (Savinov et al., 2018b) differ in the policy used and how distances are estimated. We find (*Left*) that both methods learn comparable policies, but (*Right*) our method learns more accurate distances. See text for details.

To understand why SoRB succeeds at reaching distant goals more frequently than SPTM, we examine the two key differences between the methods: (1) the *goal-conditioned policy* used to reach nearby goals and (2) the *distance metric* used to construct the graph. While SoRB acquires a goal-conditioned policy via goal-conditioned RL, SPTM obtains a policy by learning an inverse model with supervised learning. First, we compared the performance of the RL policy (used in SoRB) with the inverse model policy (used in SPTM). In Figure 7a, the solid colored lines show that, *without search*, the policy used by SPTM is more successful than the RL policy, but performance of both policies degrades as the distance to the goal increases. We also evaluate a variant of our method that uses the policy from SPTM to reach each waypoint, and find (dashed-lines) no difference in performance, likely because the policies are equally good at reaching nearby goals (within MAXDIST steps). We conclude that the difference in goal-conditioned policies cannot explain the difference in success rate.

The other key difference between SoRB and SPTM is their learned distance metrics. When using distances for graph search, it is critical for the predicted distance between two states to reflect whether the policy can successfully navigate between those states: the model should be more successful at reaching goals which it predicts are nearby. We can naturally measure this alignment using the area under a precision recall curve. Note that while SoRB predicts distances in the range  $[0, T]$ , SPTM predicts whether two states are reachable, so its predictions will be in the range  $[0, 1]$ . Nonetheless, precision-recall curves<sup>2</sup> only depend on the ordering of the predictions, not their absolute values. Figure 7b shows that the distances predicted by SoRB more accurately reflect whether the policy will reach the goal, as compared with SPTM. The average AUC across five random seeds is 22% higher for SoRB than SPTM. In retrospect, this finding is not surprising: while SPTM employs a learned, inverse model policy, it learns distances w.r.t. a random policy.

#### 5.4 Better Distance Estimates

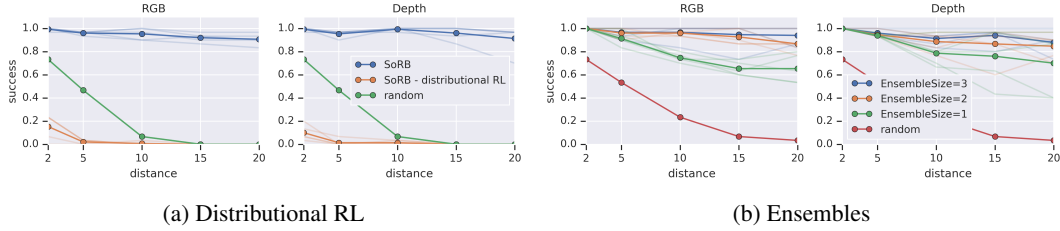


Figure 8: **Better Distance Estimates:** (Left) Without distributional RL, our method performs poorly. (Right) Ensembles contribute to a moderate increase in success rate, especially for distant goals.

We now examine the ingredients in SoRB that contribute to its accurate distance estimates: distributional RL and ensembles of value functions. In a first experiment, evaluated a variant of SoRB trained without distributional RL. As shown in Figure 8a, this variant performed worse than the random policy, clearly illustrating that distributional RL is a key component of SoRB. The second experiment studied the effect of using ensembles of value functions. Recalling that we introduced ensembles to avoid erroneous distance predictions for distant pairs of states, we expect that ensembles will contribute most towards success at reaching distant goals. Figure 8b confirms this prediction, illustrating that ensembles provide a 10 - 20% increase in success at reaching goals that are at least 10 steps away. We run additional ablation analysis in Appendix C.

#### 5.5 Generalizing to New Houses

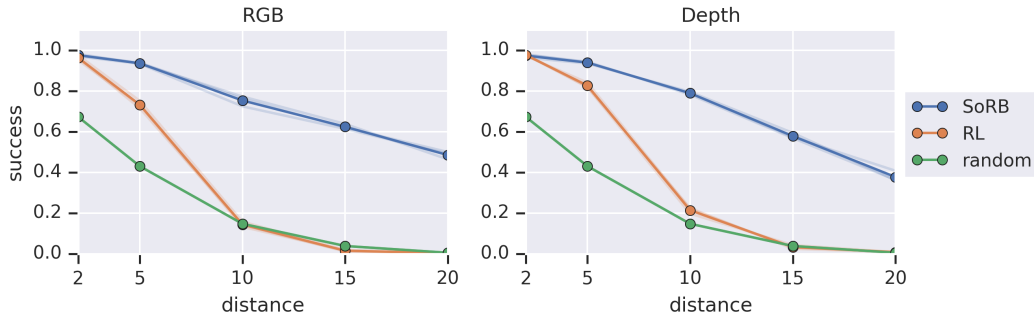


Figure 9: **Does SoRB Generalize?** After training on 100 SUNCG houses, we collect random data in held-out houses to use for search in those new environments. Whether using depth images or RGB images, SoRB generalizes well to new houses, reaching almost 80% of goals 10 steps away, while goal-conditioned RL reaches less than 20% of these goals. Transparent lines correspond to average success rate across 22 held-out houses for each of three random seeds.

We now study whether our method generalizes to new visual navigation environments. We train on 100 SUNCG houses, randomly sampling one per episode. We evaluated on a held-out test set of 22 SUNCG houses. In each house, we collect 1000 random observations and use those observations

<sup>2</sup>We negate the distance prediction from SoRB before computing the precision recall curve because small distances indicate that the policy should be more successful.



to perform search. We use the same goal-conditioned policy and associated distance function that we learned during training. As before, we measure the fraction of goals reached as we increase the distance to the goal. In Figure 9, we observe that SoRB reaches almost 80% of goals that are 10 steps away, about four times more than reached by the goal-conditioned RL agent. Our method succeeds in reaching 40% of goals 20 steps away, while goal-conditioned RL has a success rate near 0%. We repeated the experiment for three random seeds, retraining the policy from scratch each time. Note that there is no discernible difference between the three random seeds, plotted as transparent lines, indicating the robustness of our method to random initialization.

## 6 Discussion and Future Work

We presented SoRB, a method that combines planning via graph search and goal-conditioned RL. By exploiting the structure of goal-reaching tasks, we can obtain policies that generalize substantially better than those learned directly from RL. In our experiments, we show that SoRB can solve temporally extended navigation problems, traverse environments with image observations, and generalize to new houses in the SUNCg dataset. Our method relies heavily on goal-conditioned RL, and we expect advances in this area to make our method applicable to even more difficult tasks. While we used a stage-wise procedure, first learning the goal-conditioned policy and then applying graph search, in future work we aim to explore how graph search can improve the goal-conditioned policy itself, perhaps via policy distillation or obtaining better Q-value estimates. In addition, while the planning algorithm we use is simple (namely, Dijkstra), we believe that the key idea of using distance estimates obtained from RL algorithms for planning will open doors to incorporating more sophisticated planning techniques into RL.

**Acknowledgements:** We thank Vitchyr Pong, Xingyu Lin, and Shane Gu for helpful discussions on learning goal-conditioned value functions, Aleksandra Faust and Brian Okorn for feedback on connections to planning, and Nikolay Savinov for feedback on the SPTM baseline. RS is supported by NSF grant IIS1763562, ONR grant N000141812861, AFRL CogDeCON, and Apple. Any opinions, findings and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of NSF, AFRL, ONR, or Apple.

## References

- Agrawal, P., Nair, A. V., Abbeel, P., Malik, J., and Levine, S. (2016). Learning to poke by poking: Experiential learning of intuitive physics. In *Advances in Neural Information Processing Systems*, pages 5074–5082.
- Amos, B., Jimenez, I., Sacks, J., Boots, B., and Kolter, J. Z. (2018). Differentiable mpc for end-to-end planning and control. In *Advances in Neural Information Processing Systems*, pages 8289–8300.
- Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., and van den Hengel, A. (2018). Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683.
- Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Abbeel, O. P., and Zaremba, W. (2017). Hindsight experience replay. In *Advances in Neural Information Processing Systems*, pages 5048–5058.
- Bacon, P.-L., Harb, J., and Precup, D. (2017). The option-critic architecture. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- Bellemare, M. G., Dabney, W., and Munos, R. (2017). A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 449–458. JMLR.org.
- Bickel, P. J., Freedman, D. A., et al. (1981). Some asymptotic theory for the bootstrap. *The annals of statistics*, 9(6):1196–1217.
- Chiang, H.-T. L., Faust, A., Fiser, M., and Francis, A. (2019). Learning navigation behaviors end-to-end with autorl. *IEEE Robotics and Automation Letters*, 4(2):2007–2014.
- Choset, H. M., Hutchinson, S., Lynch, K. M., Kantor, G., Burgard, W., Kavraki, L. E., and Thrun, S. (2005). *Principles of robot motion: theory, algorithms, and implementation*. MIT press.

- Chua, K., Calandra, R., McAllister, R., and Levine, S. (2018). Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems*, pages 4759–4770.
- Drummond, C. (2002). Accelerating reinforcement learning by composing solutions of automatically identified subtasks. *Journal of Artificial Intelligence Research*, 16:59–104.
- Faust, A., Ramirez, O., Fiser, M., Oslund, K., Francis, A., Davidson, J., and Tapia, L. (2018). Prm-rl: Long-range robotic navigation tasks by combining reinforcement learning and sampling-based planning. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pages 5113–5120, Brisbane, Australia.
- Finn, C. and Levine, S. (2017). Deep visual foresight for planning robot motion. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2786–2793. IEEE.
- Florensa, C., Degraeve, J., Heess, N., Springenberg, J. T., and Riedmiller, M. (2019). Self-supervised learning of image embedding for continuous control. *arXiv preprint arXiv:1901.00943*.
- Fox, R., Krishnan, S., Stoica, I., and Goldberg, K. (2017). Multi-level discovery of deep options. *arXiv preprint arXiv:1703.08294*.
- Frans, K., Ho, J., Chen, X., Abbeel, P., and Schulman, J. (2017). Meta learning shared hierarchies. *arXiv preprint arXiv:1710.09767*.
- Gupta, S., Davidson, J., Levine, S., Sukthankar, R., and Malik, J. (2017). Cognitive mapping and planning for visual navigation. *arXiv preprint arXiv:1702.03920*, 3.
- Hadar, J. and Russell, W. R. (1969). Rules for ordering uncertain prospects. *The American Economic Review*, 59(1):25–34.
- Kaelbling, L. P. (1993a). Hierarchical learning in stochastic domains: Preliminary results. In *Proceedings of the tenth international conference on machine learning*, volume 951, pages 167–173.
- Kaelbling, L. P. (1993b). Learning to achieve goals. In *IJCAI*, pages 1094–1099. Citeseer.
- Kavraki, L., Svestka, P., and Overmars, M. H. (1996). Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE transactions on robotics and automation*, 12(4):566–580.
- Kulkarni, T. D., Narasimhan, K., Saeedi, A., and Tenenbaum, J. (2016). Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. In *Advances in neural information processing systems*, pages 3675–3683.
- Kurutach, T., Clavera, I., Duan, Y., Tamar, A., and Abbeel, P. (2018). Model-ensemble trust-region policy optimization. *arXiv preprint arXiv:1802.10592*.
- Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413.
- Lau, M. and Kuffner, J. J. (2005). Behavior planning for character animation. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 271–280. ACM.
- LaValle, S. M. (2006). *Planning algorithms*. Cambridge university press.
- Lee, L., Parisotto, E., Chaplot, D. S., Xing, E., and Salakhutdinov, R. (2018). Gated path planning networks. *arXiv preprint arXiv:1806.06408*.
- Lenz, I., Knepper, R. A., and Saxena, A. (2015). Deepmpc: Learning deep latent features for model predictive control. In *Robotics: Science and Systems*. Rome, Italy.
- Levine, S., Lee, Y., Koltun, V., and Popović, Z. (2011). Space-time planning with parameterized locomotion controllers. *ACM Transactions on Graphics (TOG)*, 30(3):23.
- Levy, A., Platt, R., and Saenko, K. (2019). Hierarchical reinforcement learning with hindsight. In *International Conference on Learning Representations*.
- Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.
- Lynch, C., Khansari, M., Xiao, T., Kumar, V., Tompson, J., Levine, S., and Sermanet, P. (2019). Learning latent plans from play. *arXiv preprint arXiv:1903.01973*.

- Mirowski, P., Pascanu, R., Viola, F., Soyer, H., Ballard, A. J., Banino, A., Denil, M., Goroshin, R., Sifre, L., Kavukcuoglu, K., et al. (2016). Learning to navigate in complex environments. *arXiv preprint arXiv:1611.03673*.
- Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.
- Nachum, O., Gu, S. S., Lee, H., and Levine, S. (2018). Data-efficient hierarchical reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 3307–3317.
- Nagabandi, A., Kahn, G., Fearing, R. S., and Levine, S. (2018). Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7559–7566. IEEE.
- Nair, A., McGrew, B., Andrychowicz, M., Zaremba, W., and Abbeel, P. (2018). Overcoming exploration in reinforcement learning with demonstrations. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6292–6299. IEEE.
- Oh, J., Guo, X., Lee, H., Lewis, R. L., and Singh, S. (2015). Action-conditional video prediction using deep networks in atari games. In *Advances in neural information processing systems*, pages 2863–2871.
- Osband, I., Blundell, C., Pritzel, A., and Van Roy, B. (2016). Deep exploration via bootstrapped dqn. In *Advances in neural information processing systems*, pages 4026–4034.
- Pardo, F., Tavakoli, A., Levdi, V., and Kormushev, P. (2017). Time limits in reinforcement learning. *arXiv preprint arXiv:1712.00378*.
- Parr, R. and Russell, S. J. (1998). Reinforcement learning with hierarchies of machines. In *Advances in neural information processing systems*, pages 1043–1049.
- Pong, V., Gu, S., Dalal, M., and Levine, S. (2018). Temporal difference models: Model-free deep rl for model-based control. *arXiv preprint arXiv:1802.09081*.
- Precup, D. (2000). *Temporal abstraction in reinforcement learning*. University of Massachusetts Amherst.
- Racanière, S., Weber, T., Reichert, D., Buesing, L., Guez, A., Rezende, D. J., Badia, A. P., Vinyals, O., Heess, N., Li, Y., et al. (2017). Imagination-augmented agents for deep reinforcement learning. In *Advances in neural information processing systems*, pages 5690–5701.
- Savinov, N., Dosovitskiy, A., and Koltun, V. (2018a). Semi-parametric topological memory for navigation. *arXiv preprint arXiv:1803.00653*.
- Savinov, N., Raichuk, A., Marinier, R., Vincent, D., Pollefeys, M., Lillicrap, T., and Gelly, S. (2018b). Episodic curiosity through reachability. *arXiv preprint arXiv:1810.02274*.
- Schaul, T., Horgan, D., Gregor, K., and Silver, D. (2015). Universal value function approximators. In *International Conference on Machine Learning*, pages 1312–1320.
- Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. (2015a). Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897.
- Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. (2015b). High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Shah, P., Fiser, M., Faust, A., Kew, J. C., and Hakkani-Tur, D. (2018). Follownet: Robot navigation by following natural language directions with deep reinforcement learning. *arXiv preprint arXiv:1805.06150*.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484.
- Şimşek, Ö., Wolfe, A. P., and Barto, A. G. (2005). Identifying useful subgoals in reinforcement learning by local graph partitioning. In *Proceedings of the 22nd international conference on Machine learning*, pages 816–823. ACM.

- Song, S., Yu, F., Zeng, A., Chang, A. X., Savva, M., and Funkhouser, T. (2017). Semantic scene completion from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1746–1754.
- Srinivas, A., Jabri, A., Abbeel, P., Levine, S., and Finn, C. (2018). Universal planning networks. *arXiv preprint arXiv:1804.00645*.
- Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Machine Learning Proceedings 1990*, pages 216–224. Elsevier.
- Sutton, R. S., Precup, D., and Singh, S. (1999). Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211.
- Tamar, A., Wu, Y., Thomas, G., Levine, S., and Abbeel, P. (2016). Value iteration networks. In *Advances in Neural Information Processing Systems*, pages 2154–2162.
- Vezhnevets, A. S., Osindero, S., Schaul, T., Heess, N., Jaderberg, M., Silver, D., and Kavukcuoglu, K. (2017). Feudal networks for hierarchical reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3540–3549. JMLR. org.
- Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine learning*, 8(3-4):279–292.
- Watter, M., Springenberg, J., Boedecker, J., and Riedmiller, M. (2015). Embed to control: A locally linear latent dynamics model for control from raw images. In *Advances in neural information processing systems*, pages 2746–2754.
- Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.
- Wu, Y., Tucker, G., and Nachum, O. (2018). The laplacian in rl: Learning representations with efficient approximations. *arXiv preprint arXiv:1810.04586*.
- Zhu, Y., Mottaghi, R., Kolve, E., Lim, J. J., Gupta, A., Fei-Fei, L., and Farhadi, A. (2017). Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 3357–3364. IEEE.

## A Efficient Shortest Path Computation

Our policy solves a shortest path problem every time it recomputes a new waypoint. Naïvely running Dijkstra’s algorithm to compute a shortest path among the states in our active set  $\mathcal{B}$  requires  $O(|\mathcal{B}|^2)$  queries of our value function. While the search algorithm itself is fast, it is expensive to evaluate the value function on each pair of states at every time step. In our implementation (Algorithm 2), we amortize this computation across many calls to the policy. We periodically evaluate the value function on each pair of nodes in the replay buffer, and then used the Floyd Warshall algorithm to compute the shortest path between all pairs. This takes  $O(|\mathcal{B}|^3)$  time, but only  $O(|\mathcal{B}|^2)$  calls to the value function. Let  $D \in \mathbb{R}^{|\mathcal{B}| \times |\mathcal{B}|}$  be the resulting matrix storing the shortest path distances between all pairs of states in the active set. Now, given a start state  $s$  and goal state  $g$ , the shortest path distance is

$$d_{\text{sp}}(s, g) = \min \left( \min_{u, v \in \mathcal{T}} d(s, u) + D[u, v] + d(v, g), d(s, g) \right)$$

This computation requires  $O(|\mathcal{B}|)$  calls to the value function, substantially better than the  $O(|\mathcal{B}|^2)$  calls required with the naïve implementation.

## B Environments

We used two simple navigation environments, Point-U and Point-FourRooms, shown in Figure 4a. In both environments, the observations are the location of the agent,  $s = (x, y) \in \mathbb{R}^2$ . The agent’s actions  $a = (dx, dy) \in [-1, 1]^2$  are added to the agents current position at every time step. We tuned the environments so that the goal-conditioned algorithm (which we will use as a baseline) would perform as well as possible. Observing that the agent would get stuck at corners, we modified the environment to automatically add Gaussian noise to the agents action. The resulting dynamics were

$$s_{t+1} = \text{proj}(s_t + a_t + \epsilon_t) \quad \text{where} \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2)$$

where  $\text{proj}()$  handles collisions with walls by projecting the state to the nearest free state. We used  $\sigma^2 = 1.0$  for Point-U, and  $\sigma^2 = 0.1$  for the (larger) Point-FourRooms environment.

### B.1 Visual Navigation

We ran most experiments on SUNCG house 0bda523d58df2ce52d0a1d90ba21f95c. We repeated all experiments on SUNCG house 0601a680273d980b791505cab993096a, with nearly identical results. We manually choose houses using the following criteria (1) single story, (2) no humans, and (3) included multiple rooms to make planning challenging. During training, we sampled “nearby” goal states (within 4 steps) for 80% of episodes, and sampled goals uniformly at random for the remaining 20% of episodes. We tuned these parameters to make goal-conditioned RL work as well as possible. We implemented goal-relabelling (Andrychowicz et al., 2017; Kaelbling, 1993b), choosing between the (1) originally sampled goal, the (2) current state, and (3) a future state in the same trajectory, each with probability 33%. The agent’s actions space was to move North/South/East/West. Observations were panoramic images, created by concatenating the first-person views from each of the cardinal directions. We used ensembles of 3 value functions, each with entirely independent weights. For all neural networks conditioned on both the current observation and the goal observation, we concatenated the current observation and goal observation along their last channel. For RGB images, this resulted in an input with dimensions  $H \times W \times 6$ . For depth images, the concatenated input had dimension  $H \times W \times 2$ .

---

**Algorithm 2** Inputs are the current state  $s$ , the goal state  $g$ , the replay buffer  $\mathcal{B}$ , and the value function  $V$ . Returns the length and first waypoint of the shortest path.

---

```

function SHORTESTPATH( $s, s_g, \mathcal{B}, V$ )
    // Matrices:  $D_\pi, D_{\mathcal{B} \rightarrow \mathcal{B}}, D_{s \rightarrow s_g} \in \mathbb{R}^{|\mathcal{B}| \times |\mathcal{B}|}$ 
    // Vectors:  $D_{s \rightarrow \mathcal{B}}, D_{\mathcal{B} \rightarrow g} \in \mathbb{R}^{|\mathcal{B}|}$ 
     $D_\pi \leftarrow -V(\mathcal{B}, \mathcal{B})$  ▷ cached
     $D_{\mathcal{B} \rightarrow \mathcal{B}} \leftarrow \text{FLOYDWARSHALL}(D_\pi)$  ▷ cached
     $D_{s \rightarrow \mathcal{B}} \leftarrow -V(s, \mathcal{B})$ 
     $D_{\mathcal{B} \rightarrow g} \leftarrow -V(\mathcal{B}, g)$ 
     $D_{s \rightarrow g} \leftarrow D_{s \rightarrow \mathcal{B}} + D_{\mathcal{B} \rightarrow \mathcal{B}} + (D_{\mathcal{B} \rightarrow g})^T$ 
     $s_{w1} \leftarrow \arg \min_{u, v \in \mathcal{B}} D_{s \rightarrow g}$ 
    return  $s_{w1}$ 

```

---



## C Ablation Experiments

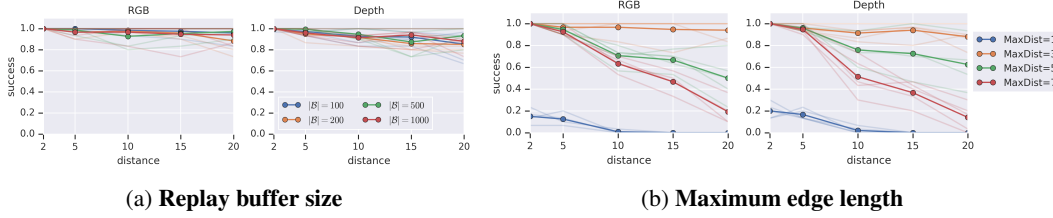


Figure 10: **Sensitivity to Hyperparameters:** (*Left*) While we used a buffer of 1000 observations for most of our experiments, decreasing the buffer size has little effect on the method’s success rate. (*Right*) When constructing our graph, we ignore edges that are longer than some distance,  $\text{MAXDIST}$ . We find that this hyperparameter is important to the success of our method.

Because SoRB plans over a fixed replay buffer, one potential concern is that performance might degrade if the replay buffer is too small. To test this concern, we ran an experiment varying the size of the replay buffer. As shown in Figure 10a, decreasing the replay buffer by a factor of 10x led to no discernible drop on performance. While we do expect performance to drop if we further decrease the size of the replay buffer, the requirement of storing 100 states (even high-resolution images) seems relatively minor. In a second ablation experiment, we varied the  $\text{MAXDIST}$  hyperparameter that governs when we stop adding new edges to the graph. As shown in Figure 10b, SoRB is sensitive to this hyperparameter, with values too large and too smaller leading to worse performance. When the  $\text{MAXDIST}$  parameter is too small, graph search fails to find a path to the goal state. As we increase  $\text{MAXDIST}$ , we increase the probability of underestimating the distance between pairs of states. We expect that improvements in uncertainty quantification in RL will improve the stability of our method w.r.t. this hyperparameter.

## D Tricks for Learning Distances with RL

1. *Small learning rates:* Especially for the image-based tasks, we found that RL completely failed with using a critic learning rate larger than  $1e-4$ . Smaller learning rates work too, but take longer to converge.
2. *Distributional RL:* The value function update for distributional RL has a particularly nice form when values correspond to distances. Additionally, distributional RL implicitly clips the values, preventing the critic to predict that unreachable states are infinitely far away.
3. *Termination Condition:* Carefully consider whether to set `done = True` at the end of each episode. In our setting the agent received a reward of -1 at each time step, so the value of each state was negative. An optimal agent therefore attempts to terminate the episode as quickly as possible. We only set `done = True` when the agent reached the goal state, not when the maximum number of time steps was reached or when it reached some other absorbing state.
4. *Ensembles of Value Functions:* Predicted distances from a single value function can be inaccurate for unseen (state, goal) pairs. When performing search using these predicted distances, these inaccurately-short predictions result in “wormholes” through the environment, where the agent mistakenly believes that two distant states are actually nearby. To mitigate this, we trained multiple, independent critics in parallel on the same data, and then aggregated predictions from each before doing search. Surprisingly, we found that taking the average predicted distance over the ensemble worked as well as taking the maximum predicted distance. We tried accelerating training by using shared convolutional layers for all critics in the ensemble, but found that this resulted in highly-correlated distant predictions that exhibited the “wormhole” problem.
5. *Normalizing Observations:* For the visual navigation experiments, we normalized the observations to be in the interval  $[0, 1]$  by dividing by the maximum pixel intensity (32 for depth, 255 for RGB). Normalization was most important for the generalization experiment with RGB observations.

## E Failed Experiments

1. *Goal Relabelling*: As mentioned above, we tried to combine our method with off-policy goal relabelling (Andrychowicz et al., 2017; Pong et al., 2018). Surprisingly, we found that this hurt performance of the non-search policy, and had no effect on the search policy.
2. *Lower-bounds on Q-values*: We attempted to use the search path to obtain a lower bound on the target Q-values during training. In the Bellman update, we replaced the distance predicted by the target Q-values with the minimum of (1) the distance predicted by the target Q-network and (2) the distance of the shortest path found by search. This can be interpreted as a generalization of the single-step lower bound from Kaelbling (1993b). Initial experiments showed this approach slowed down learning, and in some cases prevented the algorithm from converging. We hypothesize that Q-learning is much more sensitive to error in the *relative values* of two actions, rather than the *absolute value* of any particular action. While our lower-bound method likely decreased the absolute error, it did not decrease the relative error (and may have even increased it).
3. *TD3-style Ensemble Aggregation*: In our main experiments, we aggregated distance predictions from the ensemble of distributional critics by first computing the expected distance of each critic, and then taking the maximum predicted distance. This approach ignores the fact that our critics are distributional. Inspired by the stability of TD3, we attempted to apply a similar approach to aggregating predictions from the ensemble of distributional critics. The naïve approach of taking the minimum for each atom does not work because the resulting distribution will not sum to one. Instead, we first compute the cumulative density function (CDF) of each critic and then take the pointwise maximum over the CDFs. Note that critics correspond to negative distance, so the maximum corresponds to being pessimistic. Finally, we convert the resulting CDF back into a PDF and return the corresponding expected distance. While this method has neat connections to second-order stochastic dominance and risk-averse expected utility maximizers (Hadar and Russell, 1969), we found that it worked poorly in practice.

## F Hyperparameters

Unless otherwise noted, all baselines use the same hyperparameters as our method. Unless otherwise noted, parameters were not tuned.

### F.1 Search on the Replay Buffer

Parameter	Value	Comments
learning rate	1e-4	Lower values also work, but training takes longer. Same for actor and critic.
training iterations	1e6 environment steps	Performance changed little after 200k steps.
batch size	64	
train steps per environment step	1:1	
random steps at start of training	1000	
NN architecture (images)	Conv(16, 8, 4) + Conv(32, 4, 4) + FC(256)	Same for depth and RGB images.
optimizer	Adam	We used the default Tensorflow settings for $\beta_1, \beta_2, \epsilon$ . Same for actor and critic.
MaxDist	3	See Figure 10
replay buffer size (training)	100k	
replay buffer size (search)	1k	See Figure 10
gamma / discount	1	
$\epsilon$	0.1	Exploration parameter for discrete actions, used for visual navigation.
OU-stddev, OU-damping	1.0, 2.0	Exploration parameters for continuous actions, used for didactic 2D navigation
reward scale factor	0.1	Tuned for the DDPG baseline on the 2D navigation task.
target network update frequency	every 5 steps	
target network update rate ( $\tau$ )	0.05	

Table 1: Hyperparameters for SoRB

## F.2 Value Iteration Networks

Parameter	Value	Comments
number of iterations	50	Tuned over [1, 2, 5, 10, 20, 50]. Little effect.
hidden units in VI block	100	Tuned over [10, 30, 100, 300]. Little effect

Table 2: Hyperparameters for VIN (Tamar et al., 2016)

## F.3 Semi-Parametric Topological Memory

We first tuned the  $l$  parameter on goal-reaching without search. Setting  $l$  to the best found value, we performed a massive (over 1000 experiments) grid search over  $M$ ,  $s_{\text{reach}}$ , and the threshold for adding edges.

Parameter	Value	Comments
threshold for adding edges	0.9	Tuned over [0.1, 0.2, 0.5, 0.7, 0.9]
$s_{\text{reach}}$ , threshold for choosing the next waypoint along the shortest path	0.5	Tuned over [0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 1.0]
NN architecture	Conv(16, 8, 4) + Conv(32, 4, 4) + FC(256)	Same architecture (but different weights) for the retrieval and locomotor networks.
$l$ , threshold for sampling nearby states in trajectory	8	Tuned over [1, 2, 4, 8]
$M$ , margin between “close” and “far” states	1	Tuned over [1, 2, 4]

Table 3: Hyperparameters for SPTM (Savinov et al., 2018a)