A PROOFS AND ADDITIONAL ANALYSIS

A.1 DERIVING THE LOWER BOUND (PROOF OF LEMMA 1)

We first formulate the sampling procedure on starting states s_0, waypoints s_w, goals s_g and the corresponding time horizon variable t_1 and t_2. Then we derive a variational lower bound on the target log density of Eq. 2. We then show that optimizing the variational lower bound through an EM procedure is equivalent to breaking the goal-reaching task into a sequence of easier sub-problems. Finally, we wrapped up this section with a practical algorithm.

Data Generation Process. The generative model for which inference corresponds to our planning procedure can be formulated as follows. The episode starts by sampling an initial state $s_0 \sim p_0(s_0)$. Then it samples a geometric random variable $t_1 \sim Geom(1-\gamma)$ and roll out the policy $\pi(a \mid s, s_g)$ for exactly t_1 steps, starting from state s_0. We define s_w to be the state where we end up (i.e., $s_w \triangleq s_{t_1}$). Thus, s_w is sampled $s_w \sim p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+} \mid s_0)$. We then sample another geometric random variable $t_2 \sim Geom(1-\gamma)$ and roll out the policy $\pi(a \mid s, s_g)$ for exactly t_1 steps, starting from state s_w. We define s_g to be the state where we end up (i.e., $s_g \triangleq s_{t_1+t_2}$). Thus, s_g is sampled $s_g \sim p_{Geom}^{\pi(\cdot|\cdot,s_g)}(s_{t+} \mid s_w)$. Note that the time index of the final state s_g is a sample from a negative binomial distribution: $t_1 + t_2 \overset{d}{=} NB(p = 1-\gamma, n = 2)$. We can equivalently express the sampling of s_g as $s_g \sim p_{NegBinom}^{\pi(\cdot|\cdot,s_g)}(s_{t+} \mid s_0)$. We illustrate the data generative process in Fig. **??** (top).

Inference process. Under the formulation of the data generation process above, we then aim to answer the following question in the inference procedure: what intermediate states would a policy visit if it eventually reached the goal state s_g? Formally, we will estimate a distribution $q(s_w \mid s_0, s_g) \approx p(s_w \mid s_0, s_g)$. We illustrate the inference process in Fig. **??** (bottom).

We learn $q(s_w \mid s_0, s_g)$ by optimizing a evidence lower bound on our main objective (Eq. 2).

$$\log p_{\text{NEGBINOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+} = s_g \mid s_0) = \log \int p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+} = s_g \mid s_w) p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_w \mid s_0) ds_w \tag{5}$$

$$= \log \int p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+} = s_g \mid s_w) p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_w \mid s_0) \frac{q(s_w \mid s_g, s_0)}{q(s_w \mid s_g, s_0)} ds_w \tag{6}$$

$$\geq \int q(s_w \mid s_g, s_0) \left(\log p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+} = s_g \mid s_w) + \log p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_w \mid s_0) - \log q(s_w \mid s_g, s_0) \right) ds_w \tag{7}$$

$$\triangleq \mathcal{L}(\pi, q(s_w \mid s_g, s_0)). \tag{8}$$

Note that s_g is conditionally independent of s_0 given s_w (see Fig. **??** (top)), so the $p^\pi(s_{t+} = s_g \mid s_w)$ terms on the RHS need not be conditioned on s_0. The evidence lower bound, $\mathcal{L}$, depends on two quantities: the goal-conditioned policy and the distribution over waypoints. The objective for the goal-conditioned policy is to maximize the probabilities of reaching the waypoint and reaching the final state. The objective for the waypoint distribution is to select waypoints s_w that satisfy two important properties: the current policy should have a high probability of successfully navigating from the initial state to the waypoint and from the waypoint to the final goal. Note that the optimal choice for the waypoint distribution automatically depends on the current capabilities of the goal-conditioned policy.

Before optimizing the lower bound, we introduce a subtle modification to the lower bound:

$$\mathcal{L}_2(\pi, q(s_w \mid s_g, s_0)) \triangleq \int q(s_w \mid s_g, s_0) \left(\log p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+} = s_g \mid s_w) + \log p_{\text{GEOM}}^{\pi(\cdot|\cdot,\color{orange}{s_w})}(s_w \mid s_0) - \log q(s_w \mid s_g, s_0) \right) ds_w.$$

The difference, highlighted in orange, is that the probability of reaching the waypoint is computed for a goal-conditioned policy that is commanded to reach that waypoint, rather than the final goal. In Appendix **??**, we show that this new objective is also an evidence lower bound on the same goal-reaching objective (Eq. 2), but modified such that the sequence of *commanded* goals is treated as an additional latent variable.

A.2 THE OPTIMAL WAYPOINT DISTRIBUTION (PROOF OF LEMMA 2)

This section proves Lemma 2.

Proof. Recall that our goal is to solve the following maximization problem:

$$\max_{q(s_w|s_g,s_0)} \mathbb{E}_{q(s_w|s_g,s_0)}\left[\log p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+}=s_g \mid s_w) + \log p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_w \mid s_0) - \log q(s_w \mid s_g, s_0)\right].$$

Note that the waypoint distribution must integrate to one. The Lagrangian can be written as

$$\mathbb{E}_{q(s_w|s_g,s_0)}\left[\log p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+}=s_g \mid s_w) + \log p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_w \mid s_0) - \log q(s_w \mid s_g, s_0)\right] + \lambda\left(\int q(s_w \mid s_0, s_g)ds_w - 1\right),$$

where λ is a Lagrange multiplier. We then take the derivative with respect to $q(s_w \mid s_g, s_0)$:

$$\begin{aligned}\frac{d}{dq(s_w \mid s_0, s_g)} &= \frac{-q(s_w \mid s_0, s_g)}{q(s_w \mid s_0, s_g)} + \log p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+}=s_g \mid s_w) + \log p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_w \mid s_0) - \log q(s_w \mid s_g, s_0) + \lambda \\ &= -1 + \log p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+}=s_g \mid s_w) + \log p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_w \mid s_0) - \log q(s_w \mid s_g, s_0) + \lambda.\end{aligned}$$

We then set this derivative equal to zero and solve for $q(s_w \mid s_g, s_0)$:

$$q(s_w \mid s_g, s_0) = e^{\lambda-1} p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+}=s_g \mid s_w) p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_w \mid s_0).$$

Finally, we determine the value of λ such that $q(s_w \mid s_0, s_g)$ integrates to one. We can then express the optimal waypoint distribution as follows:

$$q^*(s_w \mid s_g, s_0) = \frac{p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_g \mid s_w) p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_w)}(s_w \mid s_0)}{\int p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_g \mid s'_w) p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_w)}(s'_w \mid s_0) ds'_w}.$$

□

A.3 ESTIMATING IMPORTANCE WEIGHTS (PROOF OF LEMMA 3)

This section proves Lemma 3.

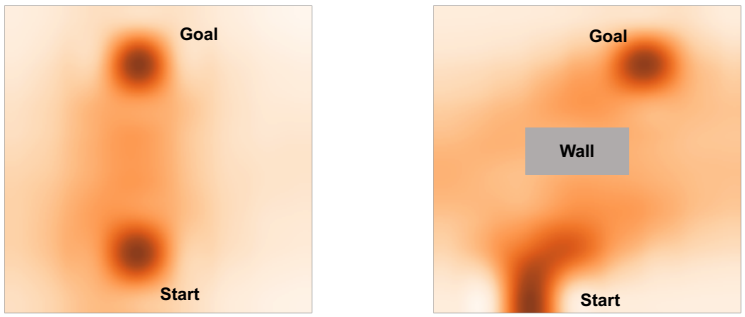
Proof. Define the normalizing constant as follows

$$Z(s_0, s_g) = \frac{b(s_g)}{\int p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_g \mid s'_w) p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_w)}(s'_w \mid s_0) ds'_w}.$$

Substituting $Z(s_0, s_g)$ into the RHS of Eq. 4 and simplifying the result, we show that it equals the LHS of Eq. 4:

$$\begin{aligned}&\frac{C_\theta(s_w, s_g)}{1 - C_\theta(s_w, s_g)} \frac{C_\theta(s_0, s_w)}{1 - C_\theta(s_0, s_w)} Z(s_0, s_g) \\ &\quad= \frac{C_\theta(s_w, s_g)}{1 - C_\theta(s_w, s_g)} \frac{C_\theta(s_0, s_w)}{1 - C_\theta(s_0, s_w)} \frac{b(s_g)}{\int p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_g \mid s'_w) p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_w)}(s'_w \mid s_0) ds'_w} \\ &\quad= \frac{p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+}=s_g \mid s_w)}{\cancel{b(s_g)}} \frac{p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_w)}(s_{t+}=s_w \mid s_0)}{b(s_w)} \frac{\cancel{b(s_g)}}{\int p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_g \mid s'_w) p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_w)}(s'_w \mid s_0) ds'_w} \\ &\quad= \frac{p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+}=s_g \mid s_w) p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_w)}(s_{t+}=s_w \mid s_0)}{\int p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_g \mid s'_w) p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_w)}(s'_w \mid s_0) ds'_w} \frac{1}{b(s_w)} \\ &\quad= \frac{q(s_w \mid s_0, s_g)}{b(s_w)}.\end{aligned}$$

□



(a) Optimal state distributon for Push

(b) Optimal state distributon for PushWall

Figure 8: *(Left)*An agent must navigate from the `start` state to the `goal` state. The heatmap visualizes the marginal state distribution of the optimal policy.

A.4 The Marginal State Distribution is a Better Initial State Distribution (Proof of Lemma 4)

We now provide a proof of Lemma 4.

Proof. We first apply Jensen's inequality

$$\mathbb{E}_{s_w \sim p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+}|s_0,s_g)}\left[p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+}=s_g \mid s_w)^2\right] \geq \mathbb{E}_{s_w \sim p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+}|s_0,s_g)}\left[p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+}=s_g \mid s_w)\right]^2$$
$$= p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+}=s_g)^2. \tag{9}$$

We then rearrange the LHS of Eq. 9:

$$\mathbb{E}_{s_w \sim p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+}|s_0,s_g)}\left[p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+}=s_g \mid s_w)^2\right]$$
$$= \mathbb{E}_{s_w \sim p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+}|s_0,s_g)}\left[p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+}=s_g \mid s)\frac{p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_w \mid s_{t+}=s_g)p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+}=s_g)}{p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_w)}\right]$$
$$= \mathbb{E}_{s_w \sim p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+}|s_0,s_g)}[p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+}=s_g \mid s_w)]p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+}=s_g). \tag{10}$$

Substituting Eq. 10 into Eq. 9 and dividing both sides by $p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+}=s_g)$, we obtain the desired result. □

B Visualization of State Density Map of Optimal Policy

We conduct this experiment on a 2D navigation task shown in Fig. 4 (left), where we have also visualized the original initial state distribution, the state distribution of an optimal policy, and the goal state. To conduct this experiment, we apply a state-of-the-art goal-conditioned RL algorithm (C-learning) in the two settings with different initial state distributions. For fair evaluation, we evaluate the policies learned in both settings using the original initial state distribution. The results shown in Fig. 4 (right) show that starting from the optimal initial state distribution results in YYx faster learning.

C Analysis

Our analysis provides a theoretical justification for why planning accelerates the acquisition of goal-reaching behaviors. We show two complementary claims. First, we show that a policy that performs planning is more likely to reach the goal than a policy that does not do planning. Our second result is that the planning process accelerates learning. This second claim is distinct because it analyses learning progress.

Lemma 4 (The marginal state distribution is a better initial state distribution)**.** *Let policy $\pi(a \mid s, s_g)$, initial state s_0, and goal state s_g be given. The discounted probability of reaching goal s_g is larger if the policy is initialized at $s_w \sim p(s_w \mid s_0, s_g)$, as compared to a policy that is initialized at s_0:*

$$p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+} = s_g \mid s_0) \leq \mathbb{E}_{s_w \sim p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+}|s_0,s_g)} \left[p_{\text{GEOM}}^{\pi(\cdot|\cdot,s_g)}(s_{t+} = s_g \mid s_w) \right]$$

See Appendix A.4 for the proof. We emphasize that this result ignores the complexity of reaching the waypoints s_w. This result therefore only applies in the idealized situation where the agent can perfectly reach the waypoint. Nonetheless, it provides intuition for why $p(s_w \mid s_0, s_g)$ is a good initial state distribution.

This first result says that the policy has a higher probability of success if it starts at the marginal state distribution of the optimal policy. Our next result shows that using the marginal state distribution of the optimal policy also accelerates learning:

Theorem 1. *Let the initial state s_0 and goal state s_g be fixed, and let π^* be the optimal goal-reaching policy:*

$$\pi^* \in \arg\max p^\pi(s_{t+} = s_g). \tag{11}$$

Consider doing projected[3] *gradient ascent on the objective function $p^\pi(s_{t+} = s_g)$ using initial state distribution μ to try to find the optimal policy for state s_0. Then gradient ascent with a step size of $\eta = (1-\gamma)^3/(2\gamma|\mathcal{A}|)$ finds an ϵ-optimal policy in time $T = \frac{64\gamma|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^6\epsilon^2} \| \frac{p_{\text{GEOM}}^{\pi^*(\cdot|\cdot,s_g)}(s_w|s_0,s_g)}{p_0(s_0)} \|_\infty^2$:*

$$\min_{t<T} p_{\text{GEOM}}^{\pi^*(\cdot|\cdot,s_g)}(s_{t+} = s_g \mid s_0) - p_{\text{GEOM}}^{\pi^t(\cdot|\cdot,s_g)}(s_{t+} = s_g \mid s_g) \leq \epsilon. \tag{12}$$

Proof. The proof is a direct application of Theorem 4.1 from Agarwal et al. (2021). □

This result is important because it directly relates the sample complexity to the mismatch between the initial state distribution $\rho_0(s_0)$ and the state distribution of an optimal policy, $p_{\text{GEOM}}^{\pi^*(\cdot|\cdot,s_g)}$. Our method implicitly sets the initial state distribution equal to the marginal state distribution of the optimal policy, thereby minimizing this upper bound on sample complexity. It is in this sense that we say our method samples optimal waypoints.

D Learning Dynamics

D.1 Learning Dynamics.

To further gain intuition into the mechanics of our method, we visualize how the distribution over waypoints changes during training of the 2D navigation of the four rooms environment. Fig. 9 shows the sampled waypoints. The value functions (i.e., future state classifiers) are randomly initialized at the start of training, so the waypoints sampled at the start of training are roughly uniform over the state space. As training progresses, the distribution over waypoints converges to the states that an optimal policy would visit enroute to the goal. While we have only shown one goal here, our method trains the policy for reaching all goals. This set of experiments provides important intuition of how C-Planning works as the distribution of waypoints shrinks from a uniform distribution to the single path connecting start state and goal state. This also provides experimental support for Eq. 2 as the distributions of waypoints qualitatively match well with the optimal solution.

[3] The projection is onto the space of feasible policies.

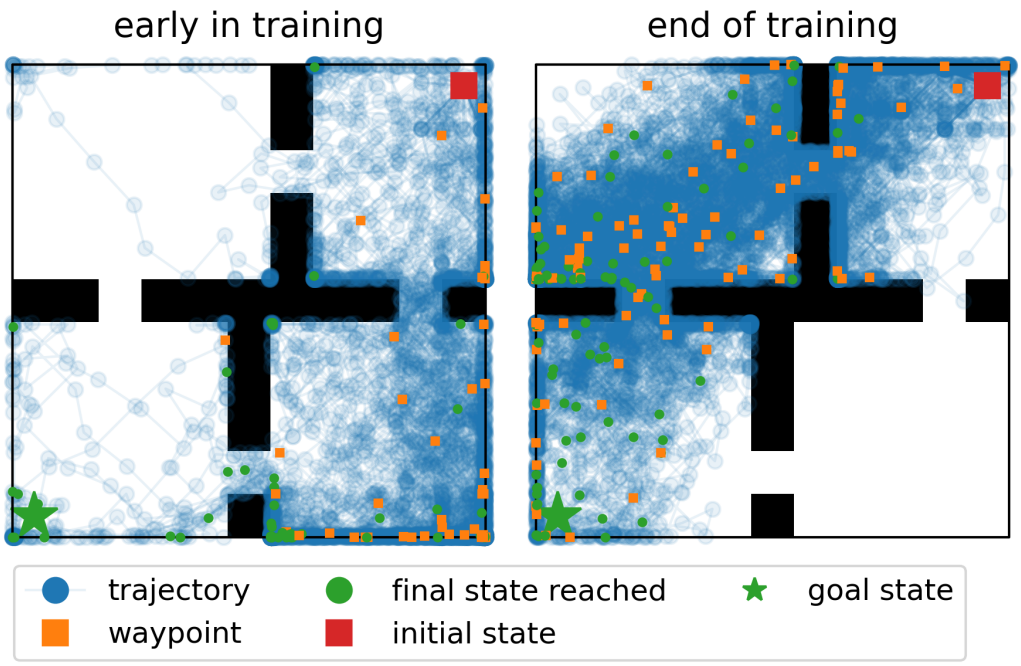


Figure 9: **Learning Dynamics:** *(Left)* Among those states that the agent is able to successfully reach early in training, the states closest to the goal are in the corners of the two rooms on the right. *(Right)* At convergence, waypoints are evenly distributed along states visited by the optimal policy, as predicted by our theory.