

922	Appendices	
923	A Additional Related Work	23
924	B Connection to the Likelihood of Success	23
925	C Proofs	23
926	C.1 Notation	23
927	C.2 Assumptions	24
928	C.3 Proof of Informal Theorem 4.2	24
929	D Additional Experimental Results	28
930	E Implementation Details	32
931	E.1 Probabilistic Value Estimation	32
932	F Hyperparameters	32

933 A Additional Related Work

934 **Self-play** Many games can be naturally formulated as sparse-reward reinforcement learning
 935 problems, where the agent receives a reward only upon winning. Seminal works have achieved
 936 superhuman ability using this approach [e.g., 45, 66, 67, 7]. A central technique, enabling these
 937 successes, is self-play [80, 64, 65, 66]. There are clear connections between these techniques and
 938 the directed exploration in DISCOVER. Self-play can be viewed as a form of goal selection [64, 65],
 939 where the current agent (or a simultaneously trained agent) is chosen as its own opponent. This
 940 opponent selection promotes *achievability*, as it is possible to beat a recent version of yourself, *novelty*
 941 as the opponent is of similar ability, and *relevance* as it is the currently strongest opponent to beat.

942 **Hierarchical RL** DISCOVER is closely related to the field of hierarchical RL [77, 16, 46], which
 943 aims to explore and plan at a higher level of abstraction. Recent methods have introduced frameworks
 944 in which multiple hierarchical levels are used to propose and learn skills [46, 25]. In this work,
 945 we focus on a simplified setting in which a single goal is selected for exploration in each episode.
 946 However, the general approach can be naturally extended to more complex hierarchical structures.

947 **Directed exploration in sequential decision making** DISCOVER is designed to efficiently address
 948 the exploration-exploitation dilemma, a key concept in sequential decision making. Fundamentally, it
 949 expresses an agent’s inevitable trade-off between the objectives of learning its environment and solving
 950 a task. A particularly common approach to balancing exploration and exploitation is grounded in the
 951 principle of optimism in the face of uncertainty [e.g., 69]. Thereby, the agent selects actions that max-
 952 imize an upper confidence bound (UCB) of the reward function, i.e., it selects actions which based on
 953 the agents’ imperfect knowledge *could* lead to a large reward. In many settings of sequential decision
 954 making, such as linear bandits, this approach achieves the rate-optimal regret $R_T \sim O(d\sqrt{T})$ [69, 2,
 955 1, 13]. The DISCOVER objective extends UCB to the problem of goal selection in RL. Beyond UCB,
 956 many other methods have been shown to effectively direct exploration towards “relevant” experience,
 957 such as in bandits [e.g., 60, 59, 34], in RL [e.g., 14, 73], or in active learning [e.g., 42, 37, 6].

958 B Connection to the Likelihood of Success

959 From a goal-reaching perspective, the undiscounted version of the DISCOVER criterion is tightly
 960 connected to the actual objective of the agent, i.e., reaching the target goal. This emerges naturally as
 961 when $\gamma \rightarrow 1$, $\epsilon \rightarrow 0$, and $\pi \rightarrow \pi^*$, the value function becomes a (negative) quasimetric [84]. Thus, it
 962 is non-positive, and respects the triangle inequality,

$$V^\pi(s_0, g) + V^\pi(g, g^*) \leq V^\pi(s_0, g^*), \quad (5)$$

963 for arbitrary $s_0 \in \mathcal{S}$, $g, g^* \in \mathcal{G}$. Note that the direction of the inequality is flipped as the value function
 964 represents a negative distance. For $\alpha = 0.5$, DISCOVER maximizes a probabilistic estimate of the left
 965 hand side of Equation (5), which is a tight lower-bound to $V^\pi(s_0, g^*)$. The value $V^\pi(s_0, g^*)$ is exactly
 966 the quantity of interest for the agent, as it represents the negative expected number of steps to reach the
 967 true goal g^* . Intuitively, DISCOVER selects goals that are optimistically going to guarantee the short-
 968 est path to the actual goal or, in the undiscounted case, the likelihood of reaching it within an episode.

969 C Proofs

970 In this section, we prove the theoretical guarantee informally stated in Informal Theorem 4.2. We
 971 begin by introducing some useful notation and the formal assumptions before proving Theorem C.9.

972 C.1 Notation

973 We define the reward function $r^*(g) = \alpha V^*(s_0, g) + (1 - \alpha)V^*(g, g^*)$ for any fixed $\alpha \in (0, \frac{1}{2})$.
 974 We define $d^*(g, g') = -V^*(g, g')$. We denote by $\mathcal{G}_t \subseteq \mathcal{G}$ the goals that are achievable by the agent
 975 in episode t with probability at least α , i.e., the policy is able to reach goals within \mathcal{G}_t with probability
 976 at least α . We use \log to denote the natural logarithm. For simplicity, we assume throughout that the
 977 initial state s_0 is fixed across all episodes.

978 C.2 Assumptions

979 **Assumption C.1** (Linear value function within feature space). For any $n \geq 1$ and for all $g \in \mathcal{G}_n$, the
 980 value functions $V^*(s_0, g)$ and $V^*(g, g^*)$ are linear in the features $\phi(\cdot), \varphi(\cdot) \in \mathbb{R}^d$ with $\phi(\cdot) \perp \varphi(\cdot)$,
 981 i.e.,

$$V^*(s_0, g) = \langle \phi(g), \theta \rangle \quad \text{and} \quad V^*(g, g^*) = \langle \varphi(g), \theta' \rangle$$

982 for some fixed $\theta, \theta' \in \mathbb{R}^d$ with $\|\theta\|_2, \|\theta'\|_2 \leq 1$.

983 **Assumption C.2** (Noisy feedback). In episode t , selecting any $g_t \in \mathcal{G}_t$, the agent receives noisy
 984 feedback $y_t = r^*(g_t) + \varepsilon_t$. We assume that the noise sequence $\{\varepsilon_t\}_{t=1}^\infty$ is conditionally R -sub-
 985 Gaussian for a fixed constant $R \geq 0$, i.e.,

$$\forall t \geq 0, \quad \forall \lambda \in \mathbb{R}, \quad \mathbb{E}[e^{\lambda \varepsilon_t} \mid \mathcal{F}_{t-1}] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right)$$

986 where \mathcal{F}_{t-1} is the σ -algebra generated by the random variables $\{g_s, \varepsilon_s\}_{s=1}^{t-1}$ and g_t .

987 **Assumption C.3** (Value function estimates). For any $h \geq 1, t \geq h$, the value function estimate
 988 relative to s_0 is given by

$$V_t(s_0, \cdot) \stackrel{\text{def}}{=} \phi(\cdot)^\top (\Sigma_t + \lambda I)^{-1} \sum_{s=h}^{t-1} \phi(g_s) y_s,$$

$$\sigma_t(s_0, \cdot) \stackrel{\text{def}}{=} \sqrt{\phi(\cdot)^\top (\Sigma_t + \lambda I)^{-1} \phi(\cdot)},$$

989 where $\Sigma_t = \sum_{s=h}^{t-1} \phi(g_s) \phi(g_s)^\top$ and $\lambda > 0$. The value function estimate relative to g^* is defined
 990 analogously with respect to the feature vector $\varphi(\cdot)$.

991 **Assumption C.4** (Goal space contains optimal paths⁴). For any goal $g' \in \mathcal{G}$, the optimal path from
 992 g' to g^* is contained in \mathcal{G} . Formally, there exists a $g \in \mathcal{G}$ such that $d^*(g', g^*) = d^*(g', g) + d^*(g, g^*)$
 993 for any $d^*(g', g) \in [0, d^*(g', g^*)]$.

994 **Assumption C.5** (Goal achievability). We denote by $\mathcal{G}_t \subseteq \mathcal{G}$ the goals that are achievable by the
 995 agent in episode t with probability at least $\alpha \in (0, 1)$. Moreover, for any $t \geq 1$, the \mathcal{G}_t contains all
 996 goals $g \in \mathcal{G}$ for which we have previously selected a goal, which is $(1 - \alpha)\kappa$ -close (under the optimal
 997 value function). We call $\kappa > 0$ the *expansion rate*. Formally,

$$\mathcal{G}_{t+1} \supseteq \{g \in \mathcal{G} \mid \exists t' \leq t : d^*(g_{t'}, g) \leq (1 - \alpha)\kappa\}.$$

998 Further, $\mathcal{G}_0 \supseteq \{s_0\}$, i.e., the initial state is always achievable.

999 **Relaxing Assumption C.2** One can consider any individual feedback y_t as being the result of an
 1000 oracle that achieves the commanded goal g_t with probability at least α within K episodes. With this
 1001 looser assumption, our bound in Informal Theorem 4.2 simply increases by a factor K .

1002 C.3 Proof of Informal Theorem 4.2

1003 We begin by restating the regret bound obtained for linear bandits in [13].

1004 **Proposition C.6.** Let Assumptions C.1 to C.3 hold. Fix any $\delta \in (0, 1), n \geq 1, \alpha \in (0, 1)$, and let

$$\beta_t = 1 + R\sqrt{2(d \log(t - n + 1) + 1 + \log(1/\delta))} \quad \text{for } t \geq n. \quad (6)$$

1005 We then have with probability $1 - \delta$ that the regret of selecting goals g_n, g_{n+1}, \dots with
 1006 $\text{DISCOVER}(\alpha, \beta_t)$ is bounded by

$$\sum_{t=h}^{h+T} \max_{g \in \mathcal{G}_n} r^*(g) - r^*(g_t) \leq O(d\sqrt{T} \log T \log(1/\delta)).$$

⁴This assumption simply states that the goal space \mathcal{G} is geodesically convex under the quasimetric d^* induced by the optimal value function. This is a standard “reachability” condition, which may be familiar to readers from control theory.

1007 *Proof.* By Theorem 3 in [13] and using that the feature spaces ϕ and φ are orthogonal (cf. Assump-
 1008 tion C.1), we have

$$\sum_{t=h}^{h+T} \max_{g \in \mathcal{G}_n} r^*(g) - r^*(g_t) \leq O(\sqrt{T}(B\sqrt{\gamma_T} + \gamma_T + \sqrt{\gamma_T} \log(1/\delta)))$$

1009 with $B = 1$. Bounding $\gamma_T \leq O(d \log T)$ using Assumption C.1, completes the proof. \square

1010 **Lemma C.7.** *Let Assumption C.4 hold and fix any $\epsilon > 0, \alpha > 0$. Then, for all $g' \in \mathcal{G}$ with*
 1011 *$d^*(g', g^*) \geq \epsilon$ there exists a $g \in \mathcal{G}$ with $d^*(g, g') = \epsilon$ such that $r^*(g) - r^*(g') \geq (1 - 2\alpha)\epsilon$.*

1012 *Proof.* Consider the optimal path from g' to g^* , i.e., the goals g satisfying

$$d^*(g', g^*) = d^*(g', g) + d^*(g, g^*).$$

1013 By Assumption C.4, for any $d^*(g', g) \in [0, \epsilon]$, we have that $g \in \mathcal{G}$. We take the goal g such that
 1014 $d^*(g', g) = \epsilon$. We then obtain

$$\begin{aligned} r^*(g) - r^*(g') &= \alpha(V^*(s_0, g) - V^*(s_0, g')) + (1 - \alpha)(V^*(g, g^*) - V^*(g', g^*)) \\ &= \alpha(d^*(s_0, g') - d^*(s_0, g)) + (1 - \alpha)(d^*(g', g^*) - d^*(g, g^*)) \\ &= \alpha(d^*(s_0, g') - d^*(s_0, g)) + (1 - \alpha)\epsilon \\ &\geq \alpha(d^*(s_0, g') - (d^*(s_0, g') + d^*(g', g))) + (1 - \alpha)\epsilon \quad (\text{triangle inequality}) \\ &= -\alpha d^*(g', g) + (1 - \alpha)\epsilon \\ &= -\alpha\epsilon + (1 - \alpha)\epsilon \\ &= (1 - 2\alpha)\epsilon. \end{aligned}$$

1015 \square

1016 **Lemma C.8** (Improvement lemma). *Let Assumptions C.1 to C.4 hold with β_t as in Equation (6). Fix*
 1017 *any $\delta \in (0, 1)$, $n \geq 1$, $\epsilon > 0$, $\alpha \in (0, \frac{1}{2})$ and $0 < \Delta < (1 - 2\alpha)\epsilon$. With probability $1 - \delta$, there exist*
 1018 *a $t' \in \{n, \dots, n + T\}$ with $T = \tilde{\Theta}(\frac{d^2}{\alpha((1-2\alpha)\epsilon - \Delta)^2})$ and a $\tilde{g} \in \mathcal{G}$ with $d^*(g_{t'}, \tilde{g}) \leq \epsilon$ such that*

$$r^*(\tilde{g}) - \max_{g \in \mathcal{G}_n} r^*(g) \geq \Delta.$$

1019 *Proof.* With probability $1 - \frac{\delta}{2}$, the number of episodes T until the agent has achieved T_{ach} of its
 1020 goals is bounded as $T = \tilde{\Theta}(\frac{T_{\text{ach}}}{\alpha})$.⁵ We denote by $\mathcal{T} \subseteq \{n, \dots, n + T\}$ the set of episodes in which
 1021 the agent has achieved its goal. Further, by Proposition C.6, also with probability $1 - \frac{\delta}{2}$, we have

$$R_n \stackrel{\text{def}}{=} \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \max_{g \in \mathcal{G}_n} r^*(g) - r^*(g_t) \leq \tilde{O}(d/\sqrt{|\mathcal{T}|}).$$

1022 All further steps are conditional on the union of the above high probability events. Thus, the regret in
 1023 successful episodes is bounded by $R_n \leq \tilde{O}(d/\sqrt{T_{\text{ach}}})$.

1024 Observe that for some $T_{\text{ach}} = \tilde{\Theta}(d^2/((1 - 2\alpha)\epsilon - \Delta)^2)$, we have that $R_n \leq (1 - 2\alpha)\epsilon - \Delta$, where
 1025 the conditions $\alpha < \frac{1}{2}$ and $\Delta < (1 - 2\alpha)\epsilon$ ensure that the bound on the regret is positive. This then
 1026 implies that there exists a $t' \in \mathcal{T}$ such that

$$\max_{g \in \mathcal{G}_n} r^*(g) - r^*(g_{t'}) \leq (1 - 2\alpha)\epsilon - \Delta \quad (7)$$

1027 Further, by Lemma C.7, there exists a $\tilde{g} \in \mathcal{G}$ such that $d^*(g_{t'}, \tilde{g}) = \epsilon$ and

$$r^*(\tilde{g}) - r^*(g_{t'}) \geq (1 - 2\alpha)\epsilon. \quad (8)$$

1028 Combining the above, we obtain

$$\begin{aligned} r^*(\tilde{g}) - \max_{g \in \mathcal{G}_n} r^*(g) &\geq r^*(\tilde{g}) - [(1 - 2\alpha)\epsilon - \Delta + r^*(g_{t'})] && (\text{Equation (7)}) \\ &= r^*(\tilde{g}) - r^*(g_{t'}) - (1 - 2\alpha)\epsilon + \Delta \\ &\geq (1 - 2\alpha)\epsilon - (1 - 2\alpha)\epsilon + \Delta && (\text{Equation (8)}) \\ &= \Delta. \end{aligned}$$

1029 \square

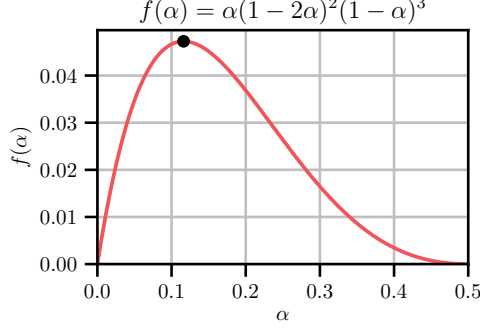


Figure 8: Plotting the effect of the parameter α on the (inverse) goal-achievement rate (cf. Theorem C.9). The target goal is reached fastest for $\alpha \approx 0.1$.

Theorem C.9. Let Assumptions C.1 to C.5 hold with β_t as in Equation (6). Fix any $\delta \in (0, 1)$, $\alpha \in (0, \frac{1}{2})$, and let $D = d^*(s_0, g^*)$. Then, with probability $1 - \delta$, selecting goals g_t with $\text{DISCOVER}(\alpha, \beta_t)$, the number of episodes N until $g^* \in \mathcal{G}_N$ is bounded by $N \leq \tilde{O}(\frac{Dd^2}{\alpha(1-2\alpha)^2(1-\alpha)^3\kappa^3}) = \tilde{O}(\frac{Dd^2}{\kappa^3})$.

Proof. We first note that

$$\begin{aligned} r^*(g^*) - r^*(s_0) &= \alpha(V^*(s_0, g^*) - V^*(s_0, s_0)) + (1 - \alpha)(V^*(g^*, g^*) - V^*(s_0, g^*)) \\ &= \alpha V^*(s_0, g^*) - (1 - \alpha)V^*(s_0, g^*) \\ &= (2\alpha - 1)V^*(s_0, g^*) = (1 - 2\alpha)d^*(s_0, g^*) = (1 - 2\alpha)D. \end{aligned}$$

We prove the theorem by applying Lemma C.8 $M \stackrel{\text{def}}{=} \lceil \frac{(1-2\alpha)D}{\Delta} \rceil$ times, while setting $\epsilon = (1 - \alpha)\kappa$. First, for an arbitrary $0 \leq i \leq M - 1$, we assume for the goal set \mathcal{G}_{iT} with some $T = \tilde{\Theta}(\frac{d^2}{\alpha(1-2\alpha)(1-\alpha)\kappa - \Delta^2})$ that it holds that

$$\max_{g \in \mathcal{G}_{iT}} r^*(g) \geq r^*(s_0) + i\Delta.$$

Now, applying Lemma C.8 yields that after an additional T steps, with high probability, there exists a $t' \in \{iT, \dots, (i+1)T\}$ such that there is a $\tilde{g} \in \mathcal{G}$ with $d^*(g_{t'}, \tilde{g}) \leq (1 - \alpha)\kappa$ satisfying

$$r^*(\tilde{g}) - \max_{g \in \mathcal{G}_{iT}} r^*(g) \geq \Delta.$$

Hence, by Assumption C.5, we have that $\tilde{g} \in \mathcal{G}_{(i+1)T}$, and therefore,

$$\max_{g \in \mathcal{G}_{(i+1)T}} r^*(g) \geq r^*(\tilde{g}) \geq \Delta + \max_{g \in \mathcal{G}_{iT}} r^*(g) \geq r^*(s_0) + (i+1)\Delta.$$

Iterating this argument M times and applying a union bound, we obtain

$$\max_{g \in \mathcal{G}_{MT}} r^*(g) \geq r^*(s_0) + M\Delta \geq r^*(g^*).$$

The total number of episodes is

$$N \stackrel{\text{def}}{=} MT \leq \tilde{O}\left(\frac{(1 - 2\alpha)Dd^2}{\Delta\alpha(1 - 2\alpha)(1 - \alpha)\kappa - \Delta^2}\right).$$

We can optimize α and Δ to minimize N under the constraints $0 < \alpha < \frac{1}{2}$, $\Delta > 0$, and $\Delta < (1 - 2\alpha)(1 - \alpha)\kappa$. The optimal choices are $\Delta = \frac{1}{3}(1 - 2\alpha)(1 - \alpha)\kappa$ and $\alpha \approx 0.1$. Substituting, we obtain $N \leq \tilde{O}(\frac{Dd^2}{\alpha(1-2\alpha)^2(1-\alpha)^3\kappa^3}) = \tilde{O}(\frac{Dd^2}{\kappa^3})$. \square

Finally, we include a technical lemma that is used in the proof of Lemma C.8.

⁵See Lemma C.10.

1047 **Lemma C.10.** *Let $0 < \alpha \leq 1$, $0 < \delta < 1$ and suppose $T_{\text{ach}} \geq 8 \log(1/\delta)$. Furthermore, let*
1048 *$X_1, \dots, X_T \stackrel{\text{i.i.d.}}{\sim} \text{Bern}(\alpha)$ and $S_T = \sum_{t=1}^T X_t$. Then, for some $T = \tilde{\Theta}(\frac{T_{\text{ach}}}{\alpha})$, with probability*
1049 *$1 - \delta$, we have $S_T \geq T_{\text{ach}}$.*

1050 *Proof.* Set

$$\gamma = \sqrt{\frac{2 \log(1/\delta)}{T_{\text{ach}}}}, \quad T = \left\lceil \frac{(1+\gamma)^2 T_{\text{ach}}}{\alpha} \right\rceil = \tilde{\Theta}\left(\frac{T_{\text{ach}}}{\alpha}\right), \quad \mu = \mathbb{E}[S_T] = \alpha T.$$

1051 Note that $\mu \geq (1 + \gamma)^2 T_{\text{ach}}$ and $0 < \gamma < 1$ (for $T_{\text{ach}} \geq 8 \log(1/\delta)$). Hence,

$$T_{\text{ach}} = (1 - \epsilon) \mu, \quad \epsilon = 1 - \frac{T_{\text{ach}}}{\mu} = \frac{\gamma(2+\gamma)}{(1+\gamma)^2} \geq \frac{\gamma}{1+\gamma} \in (0, 1).$$

1052 By the multiplicative Chernoff bound,

$$\Pr[S_T < T_{\text{ach}}] = \Pr[S_T < (1 - \epsilon)\mu] \leq \exp\left(-\frac{\epsilon^2 \mu}{2}\right) \leq \exp\left(-\frac{\gamma^2 T_{\text{ach}}}{2}\right) = \exp(-\log(1/\delta)) = \delta.$$

1053 □

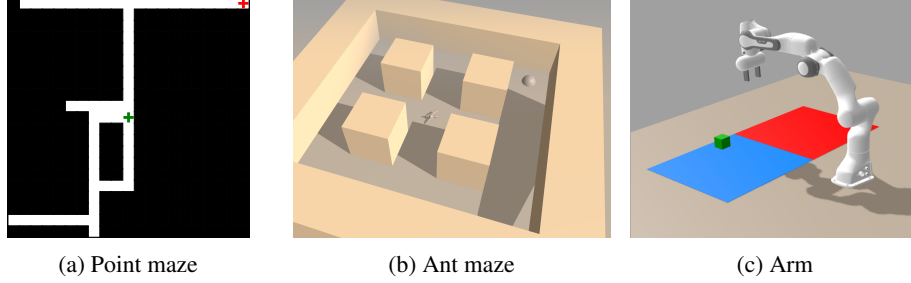


Figure 9: Sparse-reward environments from the JaxGCRL [9] library used in our evaluation. We additionally implement the pointmaze environment (left), which allows for arbitrary dimensionality. The maze is created by randomly generating paths in the environment until the target is found sufficiently often.

D Additional Experimental Results

In this section, we present additional experiments and ablations. The environments used for this evaluation are visualized in Figure 9 [9].

In high-dimensional search spaces, even direction estimates with high variance are useful. In complex environments, obtaining accurate direction estimates can be challenging. To evaluate the utility of directed goal selection in scenarios where direction estimates are imprecise, we add Gaussian noise to the target goal location. We then compare the number of environment steps required to reach a 10% target goal achievement rate using the hand-designed goal selection strategy from Figure 6, under varying levels of noise variance. The results in Figure 10 show that even with substantial noise, pointmaze environments that are unsolvable by undirected methods, remain solvable by DISCOVER. This demonstrates that even imprecise directional estimates can significantly aid target goal discovery in complex goal spaces.

Ablation of the online parameter adaptation strategy

In Figure 11, we evaluate the effect of using the simple proposed adaptation strategy with different target goal achievement ratio p_A^* , and compare with fixing the α_t parameters to 0 (Target Relevance + Novelty) and 0.5 (Fixed DISCOVER). Furthermore, we report the average α_t for DISCOVER for all easy and hard tasks respectively in Figure 12. The comparison of the success rates on the antmaze environments demonstrates that the adaptation strategy with any target goal achievement works better than fixing the parameters. This can be observed from the goal achievement rates. If we fix $\alpha_t = 0.5$, we choose goals that are "too easy" and therefore don't explore sufficiently. On the other hand, by fixing $\alpha_t = 0$ we select goals that are "too hard", which also leads too poor improvement. By using the simple adaptation strategy, we roughly achieve the target goal achievement specified. The optimal performance is achieved for the target goal achievement $p_A^* = 0.5$, which is in line with what other methods found [53, 41, 85]. The average α_t , which is found by the adaptation strategy, initially goes up to 0.15, which is roughly what we found in the theoretical analysis in the linear bandit setting (cf. Theorem C.9), and then starts to decay. The decay can be explained by the fact that once we can reach the target goal, we don't need to optimize for achievability anymore.

Influence of the term $\sigma(g, g^*)$ In Figure 13, we study how the standard deviation $\sigma(g, g^*)$ from goal g to the target g^* influences the training. This term theoretically is part of the UCB term, directing the agent towards the target goal. To this end, we fix the contribution of $\sigma(s_0, g)$ (i.e., set

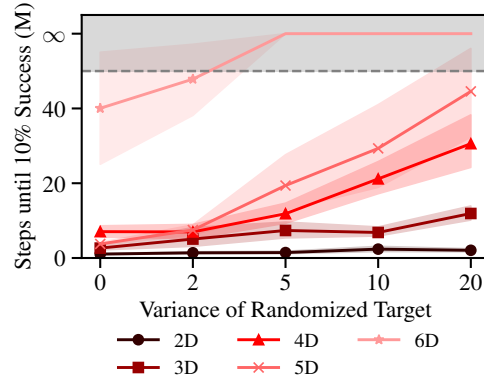


Figure 10: Evaluation of robustness directed goal selection with respect to noisy target estimates. The variance refers to the variance of the random Gaussian noise that is added to the target goal before selecting the goals using the hand-designed goal selection strategy, which uses the L_2 distance to estimate direction.

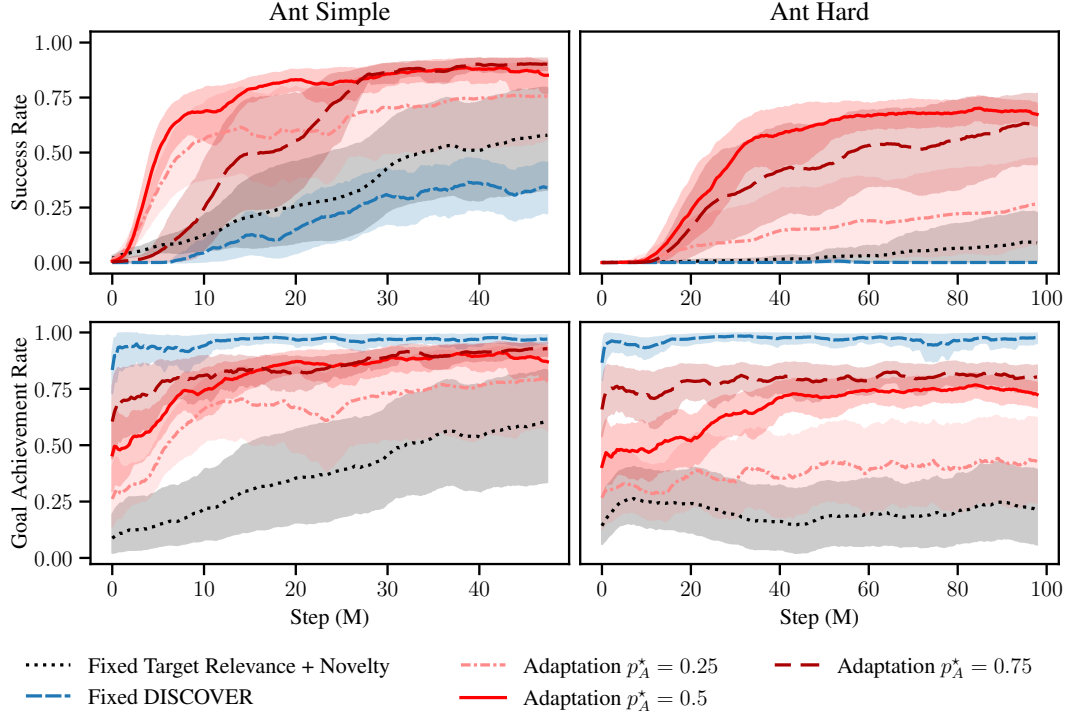


Figure 11: Comparison of how the adaptation strategy influences the goal achievement and success rates. We compare two constant strategies (Fixed DISCOVER and Fixed Target Relevance + Novelty) with an adaptation rule for different goal achievement targets.

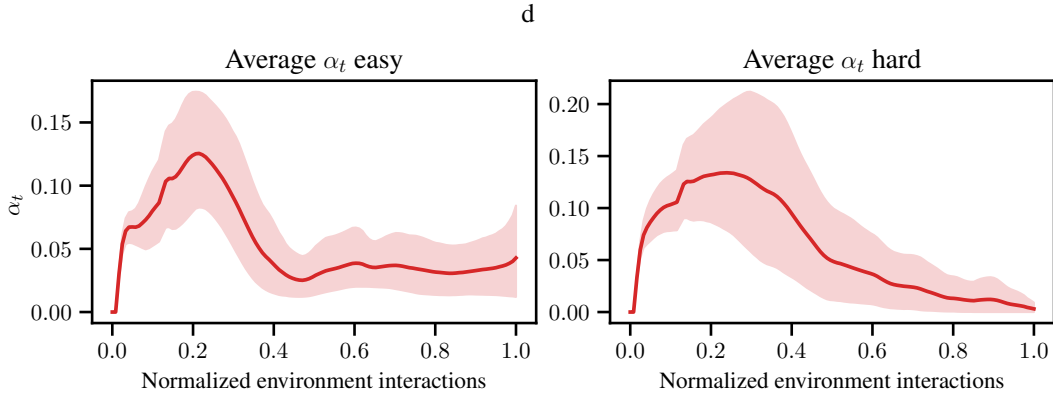


Figure 12: We plot the average α_t over the training, as adapted by the previously introduced online adaptation strategy for the DISCOVER goal selection strategy. The α_t are averaged over the three main environments.

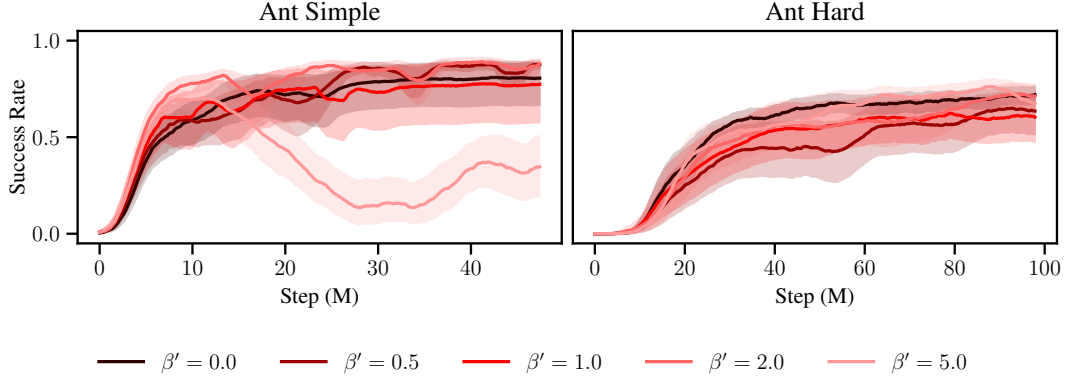


Figure 13: Comparison of how changing the coefficient of $\sigma(g, g^*)$ influences the performance, when training on the two antmaze configurations. We use the configuration with $\beta' = 0$ in all other experiments.

1094 $\beta_t = \frac{1}{\alpha_t}$) and consider a separate fixed $\beta'_t = \frac{\beta'}{1-\alpha}$, which determines the contribution of $\sigma(g, g^*)$.
 1095 The plot shows that no value for the β' parameter has a significant positive effect on the performance.
 1096 For this reason, we substitute $\sigma(s_0, g)$ for $\sigma(g, g^*)$ in our other experiments with DISCOVER.
 1097

1098 **Exploration of DISCOVER + pre-trained prior** We
 1099 visualize the exploration of the DISCOVER + pre-trained
 1100 prior goal selection strategy in Figure 14. In comparison
 1101 to DISCOVER starting from a randomly initialized agent,
 1102 it only explores in the correct direction, avoiding obstacles.
 1103 This demonstrates that access to prior can further improve
 1104 performance of DISCOVER.

1105 **Investigation of the Role of the DISCOVER components for exploration** We visualize the different components of the DISCOVER objective over the course of training in Figure 15. The first term $V(s_0, g)$ has high-value close to the initial state. By maximizing it, we will pick a goal that is close to the start and likely to be *achievable*, which matches the intuition. The second term $V(g, g^*)$ represents the value from a goal g to the target goal g^* . The plots show that in the first episodes the value is small everywhere and only once goals are discovered that are closer to the final target we observe higher values. Over the course of training, its role of encouraging to pick goals *relevant* to the final target becomes more evident. This term therefore directs the goal selection towards the final goal. Finally, the standard deviation $\sigma(s_0, g)$ has the largest value at the border of the current achievable goal set and therefore encourages selecting *novel* goals. In general, the components of the DISCOVER objective during the training match the previously presented intuition and can efficiently guide the goal selection towards the desired target.

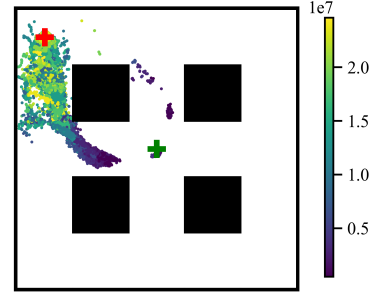


Figure 14: Selected goals by “DISCOVER + pre-trained prior” in the antmaze environment.

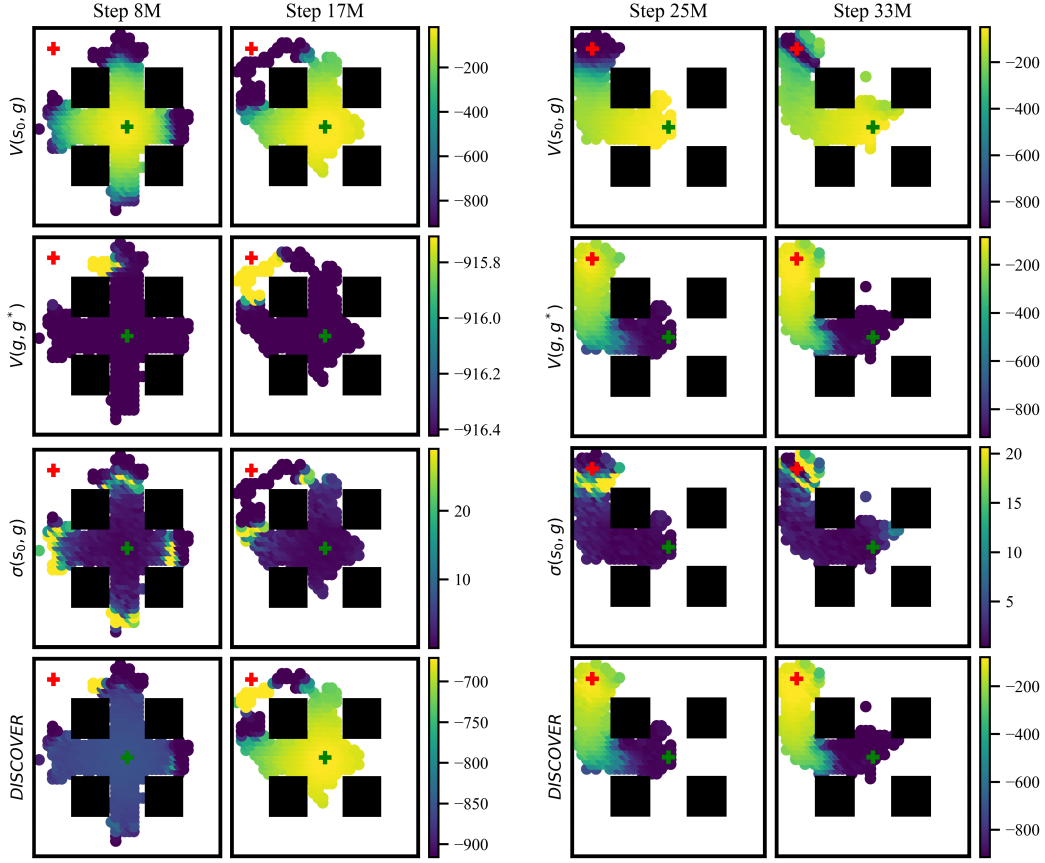


Figure 15: Visualization of the components of the DISCOVER objective at different points of the training. We plot the value functions in the regions, where achieved goals are sampled and therefore goals can be selected. The final DISCOVER objective combines the visualized terms with the current adaptation parameters α_t, β_t .

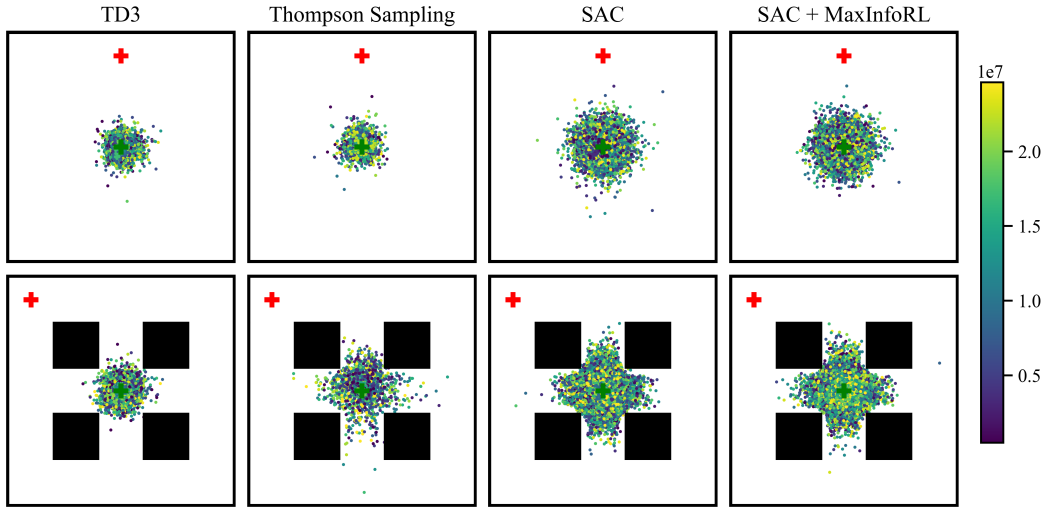


Figure 16: Visualization of the exploration of the standard non-goal-conditioned RL methods in the antmaze environments. We run MaxInfoRL [72] with SAC.

E Implementation Details

We consistently substitute $\sigma(g, g^*)$ by $\sigma(s_0, g)$ in all our experiments. In our evaluated environments, the empirical performance of DISCOVER is largely irrespective of $\sigma(g, g^*)$ (cf. Figure 13).

E.1 Probabilistic Value Estimation

A crucial component of DISCOVER is a probabilistic model of the value function, which can enable uncertainty-aware strategies. Fortunately, there are many options for probabilistic models of the value function [39, 48, 47]. For simplicity, we employ an ensemble of value functions [48] and quantify uncertainty via disagreement. This strategy has been used before to select goals, which have high exploration potential and therefore provide novel experiences [87]. Intuitively, these ensemble provide valid uncertainty estimates, as we use different random initializations for the networks and train on different data. If the networks have been trained on a certain training sample sufficiently often, the different ensemble members will converge to the same value, while if a sample hasn't been observed yet or only a few times the discrepancy will be higher. The mean and standard deviations used for the DISCOVER objective are computed as follows:

$$V(s, g) = \frac{1}{N} \sum_{i=1}^N V_i(s, g) \quad \sigma^2(s, g) = \frac{1}{N} \sum_{i=1}^N (V_i(s, g) - V(s, g))^2 \quad (9)$$

This can be seen as a straightforward extension of the standard twin critic approach [24]. We find that a slightly higher number of critic improves accuracy of uncertainty estimates. We further find that by training each critic against a random minimum of two target critics we obtain sufficient diversity for good uncertainty estimates as well as circumvent the maximization bias [82]. Additionally, we use a softplus activation at the output of each critic to limit the values to negative values.

Scaling critic ensembles to domains with large models (e.g., language) is challenging. An exciting direction for future work is to explore DISCOVER with other tools for uncertainty quantification, such as epistemic neural networks [50].

F Hyperparameters

Hyperparameter	Value
Offline RL algorithm	TD3
Ensemble size	6
Discount factor	0.99
Batch size	256
Learning rate	$3 \cdot 10^{-4}$
Policy update delay	2
Target critic Polyak factor	0.005
Relabel strategy	Uniform future: 70%, original: 30%
Target critic computation	Minimum of two random target critics
Size of critic ensemble	6
Discounting factor	0.99
Initial apdation parameter α_0	0
Horizon	100-250
k_{adapt}	64-128

Table 1: Hyperparameters for training in JaxGCRL Environments.