A SUPPLEMENTARY EXPERIMENTAL DETAILS

A.1 IMPLEMENTATION DETAILS IN CONTINUOUS ROBOTICS BENCHMARKS

This part includes details necessary to reproduce results in the continuous robotics environments. The code will be released with the camera-ready version of the paper.

Soft Actor Critic Details. Our algorithms are based on the *Soft Actor Critic* algorithm (Haarnoja et al., 2018). Notably, we use the double Q-Networks trick to help tackle the overestimation bias (Fujimoto et al., 2018). In our experiments, we do not automatically tune the entropy hyperparameter α. In fact, we found that fixing its value to $\alpha = 0.2$ is sufficient for the purpose of this paper.

Delayed Networks. As the results for the GSAC learned stationary policies show that the performance tend to decrease for high values of delay D, we opt for $D = 2$ in the continuous robotics experiments. Our main objective is to study the effect of shallow delayed geometric discounts on classic robotics environments.

Networks Architecture. Each of the discount factors γ_1 and γ_2 is associated with a different critic and target critics networks. All these networks, as well as the policy, are 1-hidden layer networks of hidden size 256. They use $ReLU$ activations and the Xavier initialization. We use Adam optimizers, with learning rates 3×10^{-4}. The list of hyperparameters is provided in Table 1.

Table 1: Sensorimotor learning hyperparameters used in DECSTR.

Hyperparam.	Description	Values.
lr_actor	Actor learning rate	3×10^{-4}
lr_critic	Critic learning rate	3×10^{-4}
τ	Polyak coefficient for target critics smoothing	0.95
$batch\_size$	Size of the batch during updates	256
$hidden\_size$	Dimension of the networks' hidden layers	256
γ_0	Discount factor associated with the first delay	0.99
γ_1	Discount factor associated with the second delay	0.99
$delayed\_update\_ratio$	# of first critic updates before single second critic update	1
$update\_per\_step$	# of networks updates loop per a single environment step	1
$target\_update$	Target networks soft updates per step	1
α	Entropy coefficient used in SAC	0.2
$automatic\_entropy$	Automatically tune the entropy coefficient	$False$

A.2 ABLATION ANALYSIS (DISCRETE CORRIDOR ENVIRONMENT)

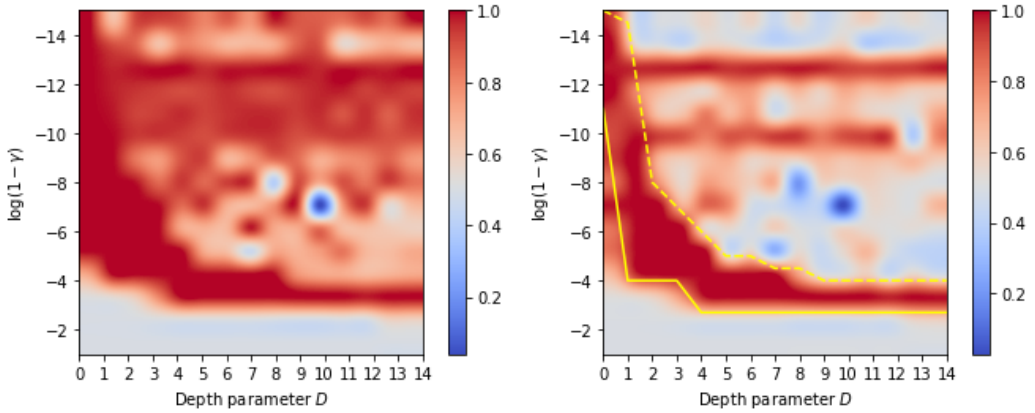
In this section, we consider a simple 2000 states long corridor environment with a deceptive reward of 0.9 on one extremity, a desirable reward of 1 on the other and an adversarial reward of -1 in the middle (states 990 to 1010). For this environment, we consider that a policy is successful if it ends up reaching the best reward and that it failed in any other scenario (including the case where it reaches the deceptive reward). The success rate of a policy is consequently the proportion of states from which it reaches the best reward.

we investigate the performances of the obtained policies using GSAC[2] as we vary the delay parameter $D \in \{0, \dots, 14\}$ and the discounts $\gamma_{i \leq D} = \gamma \in [1 - 10^{-1}, 1 - 10^{-14}]$. For each couple of values, we evaluate the best (Figure 7a) and the average (Figure 7b) success rate of the learned policies in 10 randomly initialised runs of GSAC. The obtained performances are reported in Figure 7 as heat-maps where higher success rates (close to 1) are associated with red and lower ones with blue.

Naturally, for low discount parameter γ, the success rate is around 0.5 as states on each side of the adversarial reward are encouraged to leave that area in the direction of the closest positive reward. Interestingly, there is a limiting curve (continuous yellow line in Figure 7b) above which the best stationary policy in the sense of $\mathcal{L}_D$ has a success rate of 1, and a second line (dashed yellow line) above which numerical instabilities induce poor numerical performances.

[2] A discrete version of the algorithm with down dynamics

Notice that if we consider the vertical line in Figure 7 corresponding to $D = 0$, we recover the Blackwell criterion: there exist a critical value of the discount above of which agents are capable to reach the desirable reward. Intuitively, these observations generalize this criterion to the delayed geometrically discounted framework: There exists a critical frontier that depends on both the delay D and the discounts $\gamma_{i \leq D}$, above of which optimal stationary policies in the sense of the delayed criterion $\mathcal{L}_D$ is also optimal in the sense of the average criterion

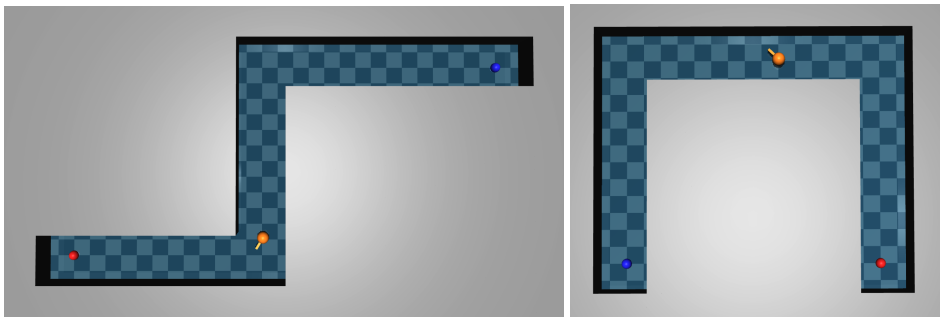


(a) Max performance (b) Average performance

Figure 7: Success rate as a function of the Delay D and the discount values

A.3 Additional Results (continuous hard exploration navigation problems)

In this section, we advance additional results in the continuous settings. We introduce two different maze environments based on the MUJOCO robotics engine: *SMaze-v0* and *UMaze-v0* (Figure 8). In both environment, the agent is a sphere whose action space is 2-dimensional. The attributes of its state space correspond to both its positions and its velocity. The environment also contain two rewarding states: a deceptive rewarding state of +0.8 (blue dot) and a maximal rewarding state of +1.0 (red dot).



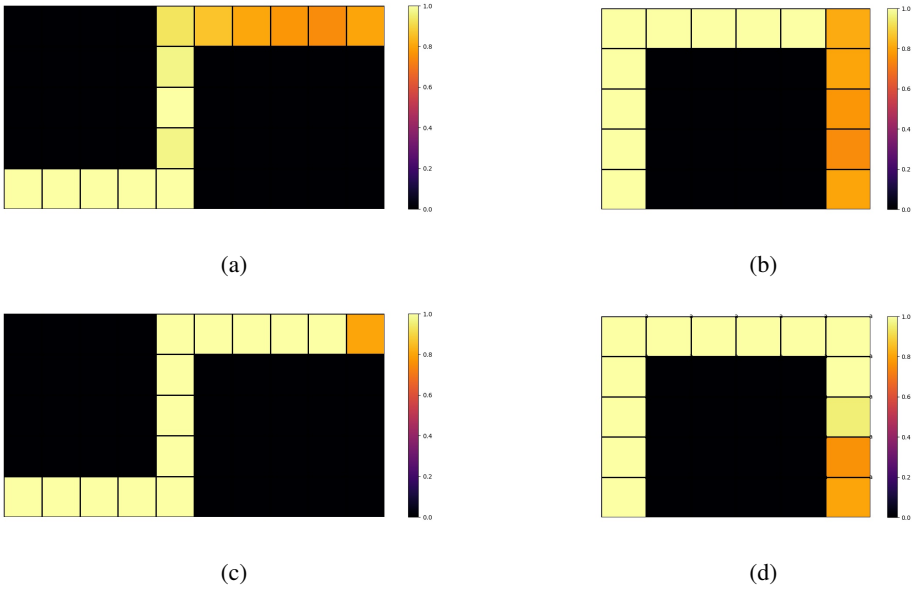
(a) SMaze-v0 (b) UMaze-v0

Figure 8: Continuous Maze Environments

Evaluation. To evaluate ours agents, we discretize the considered mazes into many cells (See Figure 9 for an illustration). At the beginning on each test episode, we initialize the agents in all the discretized cells. Following the experimental section of the main paper, we compare our proposed GSAC algorithm to the classic SAC algorithm. Our goal is to see if these agents are able to distinguish deceptive from real rewards from the cell they were initialized in.

Results. The grid plot on Figure 9 highlights the average rewards obtained by the agents when initialized in different cells. Experiments were conducted over 5 seeds. Depending on the cell in

which they were initialized, both agents choose to opt either for the deceptive or the real reward. However, the GSAC agent is capable of choosing the real reward even if it was initialized close to the deceptive one. Meanwhile, the SAC agent is more myopic, as the number of cells that leads it to the deceptive rewards are more than the ones encountered in GSAC.



(a) (b)

(c) (d)

Figure 9: Grid plot of the average rewards per cell initialization for SAC within (a) the *SMaze-v0* environment, (b) the *UMaze-v0* environment; GSAC within (c) the *SMaze-v0* environment and (d) the *UMaze-v0* environment

B PROOFS OF THE TECHNICAL RESULTS

B.1 USEFUL INTERMEDIATE RESULTS

We start by introducing some useful intermediate results that will be used later on to prove propositions 1 and 2.

USEFUL RESULTS FOR PROPOSITION 1

To derive the desired property of the value function, it is useful to derive a relationship between the coefficients $\Phi_D(t)$:

Lemma 1 *For any integer $D > 0$, the following equalities hold:*

$$\Phi_D(t) = \sum_{k=0}^{t} \gamma_D^k \Phi_{D-1}(t-k) = \Phi_{D-1}(t) + \gamma_D \Phi_D(t-1) = \sum_{d=0}^{D} \gamma_d \Phi_d(t-1)$$

Now, consider the state-value function $V_d^\pi(s)$ defined as:

$$V_D^\pi(s) := \mathbb{E}_\pi \Big[\sum_{t=0}^{\infty} \Phi_D(t) r_t | s_0 = s \Big]$$

Lemma 1 can be used to derive a relationship between the value functions $(V_d^\pi)_{d=0}^D$ for any depth parameter $D \in \mathbb{N}$:

Proposition 3 *For any state $s \in \mathcal{S}$ and for any integer $D \in \mathbb{N}$, we have:*

$$V_D^\pi(s) = \mathbb{E}_{\substack{a \sim \pi(s) \\ s' \sim \mathcal{P}(s,a)}} \Big[c(s,a) + \sum_{d=0}^{D} \gamma_d V_d^\pi(s') \Big]$$

USEFUL RESULTS FOR PROPOSITION 2

This proposition is proved using an induction reasoning. For this reason, we start by consider the simpler case of $H = 1$:

Proposition 4 *For any state $s_0 \in \mathcal{S}$, the following identity holds:*

$$V^*_\eta(s_0) = \max_{a_0} \Big\{ \big[\sum_{d=0}^{D} w_d \big] c(s_0, a_0) + \mathbb{E}_{s_1} \big[V^*_{f(\eta)}(s_1) \big] \Big\} \tag{11}$$

B.2 PROOFS

In this section we provide the proofs of the technical results

PROOF OF PROPOSITION 1

Recall that:

$$Q^\pi_D(s, a) = \mathbb{E}_{s' \sim \mathcal{P}(s,a)} \Big[V^\pi_D(s') \Big]$$

Then the statement from Proposition 3 can be reformulated as

$$Q^\pi_D(s, a) = \mathbb{E}_{\substack{s' \sim \mathcal{P}(s,a) \\ a' \sim \pi(s')}} \Big[c(s, a) + \sum_{d=0}^{D} \gamma_d Q^\pi_d(s', a') \Big] \tag{12}$$

Which means that Q^π_D is a fixed point of T^D_π. Given that this operator is a γ_D contraction with $\gamma_D \in (0, 1)$, it follows that it is the unique fixed point.

PROOF OF LEMMA 1:

The proof relies on algebraic manipulations of the indices:

$$\begin{aligned}
\Phi_D(t) &:= \sum_{\substack{\{a_d \in \mathbb{N}\}_{i=0}^{D} \\ \text{such that } \sum_d a_d = t}} \prod_{d=0}^{D} \gamma_d^{a_d} \\
&= \sum_{k=0}^{t} \gamma_D^k \Big[\sum_{\substack{\{a_d \in \mathbb{N}\}_{i=0}^{D-1} \\ \text{such that } \sum_d a_d = t-k}} \prod_{d=0}^{D-1} \gamma_d^{a_d} \Big] = \sum_{k=0}^{t} \gamma_D^k \Phi_{D-1}(t-k)
\end{aligned}$$

This concludes the proof of the first equality. Similarly, the second equality is achieved through similar algebraic treatments:

$$\begin{aligned}
\Phi_D(t) &= \sum_{k=0}^{t} \gamma_D^k \Phi_{D-1}(t-k) \\
&= \Phi_{D-1}(t) + \sum_{k=1}^{t} \gamma_D^k \Phi_{D-1}(t-k) \\
&= \Phi_{D-1}(t) + \gamma_D \sum_{k=0}^{t-1} \gamma_D^k \Phi_{D-1}((t-1)-k) = \Phi_{D-1}(t) + \gamma_D \Phi_D(t-1)
\end{aligned}$$

This concludes the proof of the second equality. The last one can be deduced directly using induction. In fact, the induction is a direct consequence of the second equality, and the basis case is trivially verified as:

$$\Phi_0(t) = \gamma_0^t = \gamma_0 \Phi_0(t-1)$$

PROOF OF PROPOSITION 3:

The proof relies on some algebraic manipulation as well as the last equality from Lemma 1.

$$
\begin{aligned}
V_D^\pi(s) &= \mathbb{E}_\pi\Big[\sum_{t=0}^{\infty}\Phi_D(t)c(s_t,a_t)|s_0=s\Big] \\
&= \mathbb{E}_\pi\Big[c(s_0,a_0)+\sum_{t=1}^{\infty}\Phi_D(t)c(s_t,a_t)|s_0=s\Big] \\
&= \mathbb{E}_\pi\Big[c(s_0,a_0)+\sum_{d=0}^{D}\gamma_d\sum_{t=1}^{\infty}\Phi_d(t-1)c(s_t,a_t)|s_0=s\Big] \\
&= \mathbb{E}_\pi\Big[c(s_0,a_0)+\sum_{d=0}^{D}\gamma_d\sum_{t=0}^{\infty}\Phi_d(t)c(s_{t+1},a_{t+1})|s_0=s\Big] \\
&= \mathbb{E}_{\substack{a\sim\pi(s)\\ s'\sim\mathcal{P}(s,a)}}\Big[c(s,a)+\sum_{d=0}^{D}\gamma_d V_d^\pi(s')\Big]
\end{aligned}
$$

where the last equality relies on the Markov property of MDPs.

PROOF OF PROPOSITION 4:

The proof relies on the linearity of the expectation as well as proposition 3. Let's denote in this proof with the policy π the sequence of action $a_0, a_1, \dots, a_\infty$ and with the transposed policy $T\pi$ the sequence of actions $a_1, a_2, \dots, a_\infty$. The following property then holds

$$
\begin{aligned}
V_\eta^\pi(s) &= \sum_{d=0}^{D} w_d V_d^\pi(s) = \sum_{d=0}^{D} w_d \mathbb{E}_{\substack{a\sim\pi(s)\\ s'\sim\mathcal{P}(s,a)}}\Big[c(s,a)+\sum_{i=0}^{d}\gamma_i V_i^{T\pi}(s')\Big] \\
&= \mathbb{E}_{a\sim\pi(s)}\Big[\big(\sum_{d=0}^{D} w_d\big)c(s,a)\Big] + \mathbb{E}_{\substack{a\sim\pi(s)\\ s'\sim\mathcal{P}(s,a)}}\Big[\sum_{d=0}^{D} w_d\sum_{i=0}^{d}\gamma_i V_i^{T\pi}(s')\Big] \\
&= \mathbb{E}_{a\sim\pi(s)}\Big[\big(\sum_{d=0}^{D} w_d\big)c(s,a)\Big] + \mathbb{E}_{\substack{a\sim\pi(s)\\ s'\sim\mathcal{P}(s,a)}}\Big[\sum_{d=0}^{D}\gamma_d\big(\sum_{i=d}^{D} w_i\big)V_d^{T\pi}(s')\Big] \\
&= \mathbb{E}_{\substack{a\sim\pi(s)\\ s'\sim\mathcal{P}(s,a)}}\Big[\big(\sum_{d=0}^{D} w_d\big)c(s,a)+\sum_{d=0}^{D}\gamma_d\big(\sum_{i=d}^{D} w_i\big)V_d^{T\pi}(s')\Big] \\
&= \mathbb{E}_{\substack{a\sim\pi(s)\\ s'\sim\mathcal{P}(s,a)}}\Big[\big(\sum_{d=0}^{D} w_d\big)c(s,a)+V_{f(\eta)}^{T\pi}(s')\Big]
\end{aligned}
$$

Using this equality and the Bellman property, it follows that the maximum value function V_η^* verifies the following:

$$
\begin{aligned}
V_\eta^*(s_0) &= \max_\pi V_\eta^\pi(s_0) = \max_\pi \mathbb{E}_{s_1\sim\mathcal{P}(s_0,a_0)}\Big[\big(\sum_{d=0}^{D} w_d\big)c(s_0,a_0)+V_{f(\eta)}^{T\pi}(s_1)\Big] \\
&= \max_{a_0,T\pi}\Big\{\big(\sum_{d=0}^{D} w_d\big)c(s_0,a_0)+\mathbb{E}_{s_1\sim\mathcal{P}(s_0,a_0)}\Big[V_{f(\eta)}^{T\pi}(s_1)\Big]\Big\} \\
&= \max_{a_0}\Big\{\big(\sum_{d=0}^{D} w_d\big)c(s_0,a_0)+\max_{T\pi}\mathbb{E}_{s_1\sim\mathcal{P}(s_0,a_0)}\Big[V_{f(\eta)}^{T\pi}(s_1)\Big]\Big\} \\
&= \max_{a_0}\Big\{\big[\sum_{d=0}^{D} w_d\big]c(s_0,a_0)+\mathbb{E}_{s_1\sim\mathcal{P}(s_0,a_0)}\big[V_{f(\eta)}^{*}(s_1)\big]\Big\}
\end{aligned}
$$