Unbiased Methods for Multi-Goal RL

Léonard Blier
Inria, Université Paris Saclay, FAIR

Yann Ollivier
FAIR

Abstract

In multi-goal reinforcement learning (RL) settings, the reward for each goal is sparse, and located in a small neighborhood of the goal. In large dimension, the probability of reaching a reward vanishes and the agent receives little learning signal. Methods such as *Hindsight Experience Replay* (HER) tackle this issue by also learning from realized but unplanned-for goals. But HER is known to introduce bias (Plappert et al., 2018), and can converge to low-return policies by overestimating chancy outcomes. First, we vindicate HER by proving that it is actually unbiased in *deterministic* environments, such as many optimal control settings. Next, for stochastic environments in continuous spaces, we tackle sparse rewards by directly taking the infinitely sparse reward limit. We fully formalize the problem of multi-goal RL with infinitely sparse Dirac rewards at each goal. We introduce unbiased deep Q-learning and actor-critic algorithms that can handle such infinitely sparse rewards, and test them in toy environments.

Most standard *reinforcement learning* (RL) methods fail when faced with very sparse reward signals. Multi-task reinforcement learning attempts to solve this problem by presenting agents with a diverse set of tasks and learn a task-dependent policy in the hope that the agent could leverage knowledge from some tasks on others (Jaderberg et al., 2016; Hausman et al., 2018; Nagabandi et al., 2019). *Multi-goal* reinforcement learning is a sub-field of multi-task RL, where the different tasks consist in reaching particular *goals* in the environment.

Universal Value Function Approximators (UVFA) (Schaul et al., 2015) extend the classical Q-learning and Temporal Difference (TD) algorithms to the multi-goal setting. It learns the goal-conditioned value-function $V^\pi(s, g)$ or Q-function $Q^*(s, a, g)$ for every state-goal pair, with function approximation, via a TD algorithm. Still, no learning occurs until a reward is observed, and UVFA fails in many high dimensional environments, when the probability of reaching the target goal is low and the agent almost never gets any learning signal.

Hindsight Experience Replay (HER) (Andrychowicz et al., 2017) is a possible solution to this issue. It leverages information between goals via the following principle: trajectories aiming at a goal g but reaching a goal g' can be used for learning exactly as if the trajectory had been aiming at g' from start. This strategy has proved successful in practice, but is known to be *biased* (Manela & Biess, 2021; Lanka & Wu, 2018). In their request for research for robotic multi-goal environments, Plappert et al. (2018) list the necessity for an unbiased version of HER, as such bias can lead to low-return policies.

Our contributions are:

- We show that in *deterministic* environments, HER is actually unbiased (Theorem 2). This case covers many standard control or robotic environments, in which HER is known to perform well. This result strengthens HER theoretically.
- We show that sparse rewards in a multi-goal setting can be handled, counter-intuitively, by dealing directly with the infinitely sparse reward limit: then the sparse reward contribution can be computed algebraically instead of sampled. The resulting Q-learning and actor-critic algorithms are unbiased even in the stochastic case and handle multi-goal RL without having

Preprint. Under review.

to observe sparse rewards, although their variance is higher than HER. For this, we fully formalize the problem of multi-goal RL with infinitely sparse rewards.

1 Multi-Goal Reinforcement Learning and Vanishing Rewards

Definition. We define a multi-goal RL environment as a variant of a Markov decision process (MDP) including a goal space. The MDP is defined by a state-space $\mathcal{S}$, an action space $\mathcal{A}$ (discrete or continuous), a discount factor γ, and a transition probability measure $P(\mathrm{d}s'|s,a)$ which describes the probability that the next state is s' after taking action a in state s; for stochastic continuous environments, this is generally a continuous probability distribution over s', hence the notation $\mathrm{d}s'$ which represents the probability to be in an infinitesimal set $\mathrm{d}s'$ around s'.

The goal space is a set $\mathcal{G}$ together with a function $\varphi\colon \mathcal{S} \to \mathcal{G}$ defining for every state s a corresponding goal $g = \varphi(s)$, which is the goal *achieved* by state s. The objective of the agent is to *reach* a goal g. This is usually formalized by defining a reward function $R_\varepsilon(s,g)$ as 1 when a given distance between the achieved goal $\varphi(s)$ and the target g is lower than a fixed value ε: $R_\varepsilon(s,g) := \mathbb{1}_{\|\varphi(s)-g\|\leqslant\varepsilon}$ for a fixed norm $\|.\|$ on $\mathcal{G}$. Thus, each goal $g \in \mathcal{G}$ defines an ordinary MDP with reward $R(s,g)$, and Q and value functions $Q^*_\varepsilon(s,a,g)$, $V^\pi_\varepsilon(s,g)$. A goal-conditioned policy $\pi(a|s,g)$ is a probability distribution over the action space $\mathcal{A}$ for every $(s,g) \in \mathcal{S}\times\mathcal{G}$.

We assume that, for a multi-goal policy $\pi(a|s,g)$, we are able to sample trajectories in the environment by sampling a goal $g \sim \rho_{\mathcal{G}}(\mathrm{d}g)$, a starting state $s_0 \sim \rho_0(\mathrm{d}s_0|g)$, and then by sampling at step t the action $a_t \sim \pi(a|s_t,g)$ and the next state $s_{t+1} \sim P(\mathrm{d}s'|s_t,a_t)$. We use the notation $P^\pi(\mathrm{d}s'|s,g) := \int_a \pi(a|s,g)P(\mathrm{d}s'|s,a)$.

Universal Value Function Approximations. UVFA (Schaul et al., 2015) allow for learning the value function $V^\pi_\varepsilon(s,g) = \mathbb{E}_{a_t\sim\pi(.|s_t,g),s_{t+1}\sim P(.|s_t,a_t)}\left[\sum_{t\geqslant 0}\gamma^t R_\varepsilon(s_t,g)|s_0=s\right]$ and the optimal Q-function $Q^*_\varepsilon(s,a,g)$. Formally, Q-learning with UVFA can be defined as standard Q-learning on the augmented state space $\tilde{\mathcal{S}} := \mathcal{S}\times\mathcal{G}$, with the transition distribution $\tilde{P}$ defined as follows: if action a is performed in state (s,g), the next state $\tilde{s}'$ is (s',g) with $s' \sim P(\mathrm{d}s'|s,a)$. The augmented environment is not a multi-goal environment, and the policy $\pi(a|s,g) = \pi(a|\tilde{s})$ becomes a standard non-goal-dependent policy in $\tilde{\mathcal{S}}$. The UVFA Q-learning update corresponds to standard parametric Q-learning on the augmented environment.

In practice, we consider a parametric function $Q_\theta(s,g)$, and we want to learn θ such that $Q_\theta(s,g)$ approximates $Q^*_\varepsilon(s,g)$. If θ is our current estimate and Q_{tar} a target Q-function the Q-learning UVFA stochastic update $\widehat{\delta\theta}_{\mathrm{UVFA}}$ is defined as follows. We consider an exploration policy $\pi_{\mathrm{expl}}(a|s,g)$. When a transition (s,a,s',g) is observed, with $a \sim \pi_{\mathrm{expl}}(.|s,g)$ and $s' \sim P(.|s,g)$, $\widehat{\delta\theta}_{\mathrm{UVFA}}$ is:

$$\widehat{\delta\theta}_{\mathrm{UVFA}}(s,a,s',g) := -\frac{1}{2}\partial_\theta\left(Q_\theta(s,a,g) - R_\varepsilon(s,g) - \gamma\sup_{a'}Q_{\mathrm{tar}}(s',a',g)\right)^2 \tag{1}$$

Then, we update θ with $\theta \leftarrow \theta + \eta\widehat{\delta\theta}_{\mathrm{UVFA}}$, where η is the learning rate. The update $\widehat{\delta\theta}_{\mathrm{UVFA}}$ is an unbiased estimate of $1/2\partial_\theta\|Q_\theta - T\cdot Q_{\mathrm{tar}}\|^2$ where T is the optimal Bellman operator, $T\cdot Q(s,a,g) = R_\varepsilon(s,g) + \gamma\mathbb{E}_{s'\sim P(.|s,a)}\left[\sup_{a'}Q(s',a',g)\right]$, whose unique fixed point is Q^*_ε. In particular, in the tabular setting, this guarantees that a function Q_∞ is a fixed point of UVFA if and only if $T\cdot Q_\infty = Q_\infty$, which means $Q_\infty = Q^*_\varepsilon$.

UVFA and vanishing rewards. A major problem with multi-goal setups is the low probability with which each specific goal g is achieved, since rewards are observed only in a ball of radius ε around the goal. In a continuous noisy environment of dimension n, reaching a goal up to precision ε becomes almost surely impossible when $\varepsilon \to 0$. With noise in dimension n, the probability to exactly reach a predefined goal g scales like $O(\varepsilon^n)$. In particular, the Q and value functions vanish like $O(\varepsilon^n)$ when ε is small. The situation is different in continuous deterministic environments. If it is possible to reach a goal exactly by selecting the right action, then the optimal Q-function Q^*_ε does not vanish, even if $\varepsilon = 0$.

With the UVFA update, the probability to observe a reward $\mathbb{1}_{\|\varphi(s)-g\|\leqslant\varepsilon}$ vanishes like $O(\varepsilon^n)$ for continuous exploration policies. So even if Q^*_ε itself does not vanish, the learning algorithm for

Q^*_ε may vanish. In practice, in an environment of dimension $n = 6$, UVFA is not able to learn anymore (experiment in Fig. 1). This vanishing issue cannot be solved solely by an exploration strategy: the issue is not the lack of diversity in visited states but rather the state space is too large to be visited by an exploration trajectory (Andrychowicz et al., 2017). Solving the issue of sparse rewards requires gathering some information even from *failing trajectories* which do not reach their initial goal, namely, leveraging the structure of multi-goal environments by using that every state achieves *some* goal. This the case in HER but not UVFA.

In this work we study algorithms which leverage the multi-goal structure and do not vanish even in the limit $\varepsilon \to 0$. We will focus on *unbiased* algorithms, which ensure that the true Q or value function is indeed a fixed point, by stochastic gradient arguments. UVFA is unbiased but vanishes when $\varepsilon \to 0$. HER does not vanish, but is known to be biased. In Section 2 we prove that HER is unbiased in deterministic environments. Sections 3.2 and 3.4 present non-vanishing, unbiased algorithms for stochastic environments; however, they are less efficient than HER in deterministic environments.

2 Hindsight Experience Replay in Stochastic or Deterministic Environments

Hindsight Experience Replay (HER) (Andrychowicz et al., 2017) is a way to solve the issue of sparse rewards for multi-goal environments by leveraging the mutual information between goals. The principle is the following: trajectories aiming at a goal g but reaching a goal g' can be used for learning exactly as if the trajectory had been aiming for g' from start. Formally, when observing a trajectory $\tau = (g, s_0, a_0, s_1, a_1, ...)$, HER samples two random integers $0 \leqslant K \leqslant L$, and performs a Q-learning update at step s_K, but for a re-sampled goal g' that is, with some probability, either $g' = g$ or $g' = \varphi(s_L)$, the goal achieved by the L-th state in the trajectory: $\widehat{\delta\theta}_{\mathrm{HER}}(\tau, K, L) := \frac{1}{2}\partial_\theta \left(Q_\theta(s_K, a_K, g') - R_\varepsilon(s_K, g') - \gamma \sup_{a'} Q_{\mathrm{tar}}(s_{K+1}, a', g')\right)^2$. In particular, the HER update does not vanish even for $\varepsilon = 0$: with nonzero probability, $K = L$ and $g' = \varphi(s_L)$, so that $R_\varepsilon(s_K, g') = 1$.

Bias of HER in stochastic environments. HER is known to be biased in a general setting (Manela & Biess, 2021; Lanka & Wu, 2018; Plappert et al., 2018), and this bias corresponds to a well-known psychological bias (Fischhoff, 1975). Here is a simple way to design *counter-examples* environments which exhibit this HER bias. Consider a finite multi goal environment and add a single action a^* which, from any state s, sends the agent to a uniform random state s' and then *freezes* it, which means the agent will always stay at s'.

Both in theory and practice, HER will learn to always select the action a^* (third plot in Fig. 1). The intuition is the following: when the agent acts with a^* and reaches a random state s', HER reinforces a^* as a good way to reach s' from s, while this was purely random. Formally, the following statement (proof in Appendix B.2) shows that HER will overestimate the value of action a^*. We say that Q_∞ is a *fixed point* of HER if $\mathbb{E}_{\tau,K,L}\left[\widehat{\delta\theta}_{\mathrm{HER}}(\tau, K, L)\right] = 0$ when $Q_\theta = Q_{\mathrm{tar}} = Q_\infty$.

THEOREM 1. *Let $\mathcal{M}$ be any finite multi-goal environment, and $\tilde{\mathcal{M}}$ the modified environment with the* freeze *action a^*. Then $\tilde{\mathcal{M}}$ is a* counter-example *to HER, which is biased in this environment. Namely, if Q_∞ is a fixed point of HER for $\tilde{\mathcal{M}}$, then for every* unfrozen *state s and goal g, Q_∞ will* overestimate *the value of a^*: $Q_\infty(s, a^*, g) > Q^*(s, a^*, g)$ where Q^* is the true value function.*

Generally, HER is overestimating chancy outcomes, by estimating that any action (even random) that led to some goal was a good way to reach that goal. This is clear in the example of the freeze-after-random-jump actions in Theorem 1. Thus, HER has no reason to learn reliably in a stochastic environment. Other hindsight methods such as (Rauber et al., 2019) experience a similar bias.

HER is unbiased in deterministic environments. Despite its bias, HER is efficient in practice, especially in continuous control environments. We vindicate HER theoretically by showing that HER is unbiased in *deterministic* environments. We say that an environment is *deterministic* is the next state s_{t+1} is uniquely determined by the current state s_t and an action a_t. This covers many usual environments such as robotic environments.

THEOREM 2. *In a deterministic multi-goal environment such that every target state is reachable from any starting state, HER is an unbiased Q-learning method. Namely, there is a Euclidean norm $\|.\|_{\mathrm{HER}}$*

such that if Q_θ is the current estimate of Q^*, the HER update $\widehat{\delta\theta_{HER}}$ is an unbiased stochastic estimate of the gradient step between Q_θ and the target function TQ_{tar}: $\mathbb{E}\left[\widehat{\delta\theta}_{HER}\right] = 1/2\partial_\theta \|Q_\theta - TQ_{\text{tar}}\|^2_{\text{HER}}$. In particular, the true Q-function Q^* is a fixed point of HER in expectation: if Q_θ and Q_{tar} are equal to Q^* then $\mathbb{E}\left[\widehat{\delta\theta}_{HER}\right] = 0$.

The proof and a more detailed statement are given in Appendix B.1. This result vindicates HER for deterministic environments: HER leverages the structure of multi-goal environments, is not vanishing when the rewards are sparse, and is mathematically well-grounded.

3 Multi-Goal RL via Infinitely Sparse Rewards

3.1 Taking the Infinitely Sparse Reward Limit

In Section 2, we saw that while HER is well-founded in deterministic environments, it is biased in the stochastic case and can learn low-return policies (Figure 1). We now introduce unbiased methods for multi-goal RL in the general setting, including stochastic environments.

In continuous state spaces, the reward is usually defined as $R_\varepsilon(s,g) = \mathbb{1}_{\|\varphi(s)-g\|\leqslant\varepsilon}$. When $\varepsilon \to 0$, the probability of reaching the reward with a stochastic policy goes to 0, and for any stochastic policy, the value function $V^\pi_\varepsilon(s,g)$ converges to 0 as well. To avoid this vanishing issue, we need a scaling factor, and consider the reward $\frac{1}{\lambda(\varepsilon)}R_\varepsilon(s,g)$, with $\lambda(\varepsilon)$ the volume of the ball of size ε in goal space. When $\varepsilon \to 0$, this rescaled reward *converges* to the *Dirac reward*:

$$R(s, \mathrm{d}g) := \delta_{\varphi(s)}(\mathrm{d}g), \tag{2}$$

where δ_x is the Dirac measure at x. Intuitively, the Dirac reward $R(s,\mathrm{d}g)$ is infinite if the goal is reached ($\varphi(s) = g$) and 0 elsewhere. Formally, the reward is not a function but a *measure* on the goal space $\mathcal{G}$ parametrized by the state s.

However, even after such a scaling, the UVFA update still vanishes with high probability for small ε (this just scales things by $1/\lambda(\varepsilon)$). We will build algorithms that work directly in the limit $\varepsilon = 0$: replacing the sparse reward $R_\varepsilon(s,g)$ by the *infinitely sparse* reward $R(s,\mathrm{d}g) = \delta_{\varphi(s)}(\mathrm{d}g)$ will allow us to leverage the Dirac structure to remove the vanishing rewards issue.

Computing the exact contribution of sparse rewards. We now explain how to leverage the multi-goal sparse reward structure. The key idea is that, with $\varepsilon = 0$, the contribution of the reward term in the Bellman equation can be computed exactly in expectation. Infinitely sparse rewards can be treated algebraically. This derivation is informal; the formal proof is in Appendix C.2.

Let us start with the expectation of the UVFA update (1) with $\varepsilon > 0$ and rewards rescaled by $1/\lambda(\varepsilon)$:

$$\begin{aligned}
\delta\theta_{\text{UVFA}} &= \mathbb{E}_{s,a,s',g}\left[\widehat{\delta\theta}_{\text{UVFA}}(s,a,s',g)\right] \\
&= -\frac{1}{2}\partial_\theta \mathbb{E}_{s,a,g,s'}\left[\left(Q_\theta(s,a,g) - \frac{1}{\lambda(\varepsilon)}R_\varepsilon(s,g) - \gamma \max_{a'} Q_{\text{tar}}(s',a',g)\right)^2\right] \\
&= \mathbb{E}_{s,a,g}\left[\partial_\theta Q_\theta(s,a,g)\frac{1}{\lambda(\varepsilon)}R_\varepsilon(s,g)\right] - \mathbb{E}_{s,a,g,s'}\left[\partial_\theta Q_\theta(s,a,g)\left(Q_\theta(s,a,g) - \gamma \max_{a'} Q_{\text{tar}}(s',a',g)\right)\right].
\end{aligned}$$

This update cannot be used for small ε, because $R_\varepsilon(s,g)$ is 0 most of the time, even though the expectation is nonzero and a huge $1/\lambda(\varepsilon)$ reward is observed with low probability.

But when $\varepsilon \to 0$, the rescaled reward $\frac{1}{\lambda(\varepsilon)}R_\varepsilon(s,g)$ converges to the Dirac reward $\delta_{\varphi(s)}$. Therefore, we can rewrite this first term as

$$\begin{aligned}
\frac{1}{\lambda(\varepsilon)}\mathbb{E}_{s,a,g}\left[\partial_\theta Q_\theta(s,a,g)R_\varepsilon(s,g)\right] &\to_{\varepsilon\to 0} \mathbb{E}_{s,a,g}\left[\partial_\theta Q_\theta(s,a,g)\delta_{\varphi(s)}(\mathrm{d}g)\right] \\
&= \mathbb{E}_{s,a}\left[\partial_\theta Q_\theta(s,a,\varphi(s))\right].
\end{aligned}$$

In this expression, sparse reward issues are avoided, just by taking the goal $g = \varphi(s)$ associated with the currently visited state s. Instead of waiting to reach a goal to update the Q-function, this updates the Q-function for the currently realized goal.

The resulting algorithm, δ-DQN, is described in Theorem 4 and Algorithm 1. The proper mathematical treatment (below and in the Appendix) of the Dirac limit shows that this actually estimates, not Q itself, but the *density* $q(s, a, g)$ of the distribution of realized goals with respect to the goal sampling distribution $\rho_{\mathcal{G}}(\mathrm{d}g)$ of the environment. This density q can be used to rank actions (indeed, the scaling by $\rho_{\mathcal{G}}$ between q and Q only depends on the goal g, so for a fixed goal, states and actions are ranked the same way). In general, working with probability densities is the only way that makes sense in the presence of noise, as the probability to exactly reach a goal will be 0.

A similar treatment holds for policy gradient (Sections 3.3–3.4).

3.2 Unbiased Multi-Goal Q-learning with Infinitely Sparse Rewards

Our goal here is to formally define multi-goal Q-learning with infinitely sparse rewards. In general, the probability of reaching any goal *exactly* is 0: instead we will learn the probability *distribution* of the goals reached by a policy, and compute the probability to reach each infinitesimal element $\mathrm{d}g$ in goal space. This is done by treating everything as measures over $\mathcal{G}$: the reward $\delta_{\varphi(s)}(\mathrm{d}g)$ is a measure, and the value functions $V^\pi(s, \mathrm{d}g)$ or optimal action-value function $Q^*(s, a, \mathrm{d}g)$ are measures on $\mathcal{G}$ as well. In the following, we define these objects in detail, and show how to learn them in practice.

First, we define an *optimal Bellman operator* T on *action-value measures*, and the optimal action-value measure $Q^*(s, a, \mathrm{d}g)$. Then, we derive δ-DQN, a deep Q-learning algorithm with infinitely sparse rewards for multi-goal RL.

Optimal Bellman equation and optimal Q-function. We first define $Q^*(s, a, \mathrm{d}g)$, the *optimal action-value measure*, the mathematical object corresponding to the usual optimal Q-function Q^* but infinitely sparse rewards. The following theorem defines the optimal Bellman operator for action-value measures, and $Q^*(s, a, \mathrm{d}g)$ as its fixed point. It is formallly stated in Appendix C.1.

DEFINITION-THEOREM 3. *Let $Q(s, a, \mathrm{d}g)$ a measure on $\mathcal{G}$ parametrized by $s, a \in \mathcal{S} \times \mathcal{A}$. We define the optimal Bellman operator T which sends Q to $T \cdot Q$ with*

$$(T \cdot Q)(s, a, \mathrm{d}g) := \delta_{\varphi(s)}(\mathrm{d}g) + \gamma\, \mathbb{E}_{s' \sim P(\mathrm{d}s'|s,a)} \sup_{a'} Q(s', a', \mathrm{d}g) \tag{3}$$

where $\delta_{\varphi(s)}$ is the Dirac measure at $\varphi(s) \in \mathcal{G}$. We define the optimal action-value measure Q^* *as follows. Set $Q_0(s, a, \mathrm{d}g) := 0$, and $Q_{n+1} := TQ_n$. Then $Q_n(s, a, \mathrm{d}g)$ converges to some $Q^*(s, a, \mathrm{d}g)$. Moreover, this $Q^*(s, a, \mathrm{d}g)$ solves the fixed point equation $TQ^* = Q^*$.*

Q-learning with function approximations, with infinitely sparse rewards. From the fixed point equation for Q^*, we would like to learn a model of $Q^*(s, a, \mathrm{d}g)$ with function approximation. We will represent measures over goals via their *density* with respect to the goal sampling function $\rho_{\mathcal{G}}$ of the environment. Namely, we will approximate $Q^*(s, a, \mathrm{d}g)$ by a model $Q_\theta(s, a, \mathrm{d}g) = q_\theta(s, a, g)\rho_{\mathcal{G}}(\mathrm{d}g)$ where $q_\theta(s, a, g)$ is an ordinary function, and learn q_θ. Hence, q_θ may be approximated by any parametric model, such as a neural network.

The following theorem properly defines an unbiased stochastic δ-DQN update with infinitely sparse rewards for the density $q_\theta(s, a, g)$:

THEOREM 4. *Let $Q_\theta(s, a, \mathrm{d}g) = q_\theta(s, a, g)\rho_{\mathcal{G}}(\mathrm{d}g)$ be a current estimate of $Q^*(s, a, \mathrm{d}g)$. Let likewise $Q_{\mathrm{tar}}(s, a, \mathrm{d}g) = q_{\mathrm{tar}}(s, a, g)\rho_{\mathcal{G}}(\mathrm{d}g)$ be a target Q-function, and consider the following update to bring Q_θ closer to TQ_{tar} with T the optimal Bellman operator.*

Let (s, a, s') be a sample of the environment such that $s' \sim P(s'|s, a)$ and $g \sim \rho_{\mathcal{G}}$ is sampled independently. Let $\widehat{\delta\theta}_{\delta\text{-DQN}}(s, a, s', g)$ be

$$\widehat{\delta\theta}_{\delta\text{-DQN}}(s, a, s', g) := \partial_\theta q_\theta(s, a, \varphi(s)) + \partial_\theta q_\theta(s, a, g) \left(\gamma \max_{a'} q_{\mathrm{tar}}(s', a', g) - q_\theta(s, a, g)\right) \tag{4}$$

Then $\widehat{\delta\theta}_{\delta\text{-DQN}}$ is an unbiased estimate of the Bellman error: $\mathbb{E}\left[\widehat{\delta\theta}_{\delta\text{-DQN}}\right] = \frac{1}{2}\partial_\theta \|Q_\theta - TQ_{\mathrm{tar}}\|^2$, where the Euclidean norm $\|\cdot\|$ on measures is defined in Theorem 11 (Appendix C.2).

In particular, the true optimal state-action measure Q^* is a fixed point of this update: if $Q_\theta = Q_{\mathrm{tar}} = Q^*$ then $\mathbb{E}\left[\widehat{\delta\theta}_{\delta\text{-DQN}}\right] = 0$.

Algorithm 1 δ-DQN

Input: Randomly initialized model $q_\theta(s, a, g)$; φ; exploration policy $\pi_{\text{expl}}(a|s, g)$; goal function φ; memory buffer `TransitionMemory`, T the maximum trajectory length
repeat
  for K trajectories **do**
    Get a goal g and an initial state s_0
    for $0 \leqslant t \leqslant T$ steps do **do**
      Sample $a_t \sim \pi_{\text{expl}}(.|s_t, g)$, execute a_t and observe s_{t+1}
      Store in the transition memory the transition `TransitionMemory` $\leftarrow (s_t, a_t, s_{t+1})$
    end for
    for L gradient steps **do**
      Sample $(s, a, s') \sim$ `TransitionMemory` and $g \sim \rho_{\mathcal{G}}$
      $\widehat{\delta\theta}_{\delta\text{-DQN}} := \partial_\theta q_\theta(s, a, \varphi(s)) + \partial_\theta q_\theta(s, a, g)\,(\gamma \max_{a'} q_\theta(s', a', g) - q_\theta(s, a, g))$.
      Stochastic gradient step: $\theta \leftarrow \theta + \eta \widehat{\delta\theta}_{\delta\text{-DQN}}$.
    end for
  end for
until end of learning

This update leads to δ-DQN (Algorithm 1, which corresponds to standard DQN with infinitely sparse rewards. For continuous actions, δ-DQN can be modified similarly to DDPG (Lillicrap et al., 2016).

Example: the tabular case. The tabular case highlights the difference between UVFA and δ-DQN. When a transition (s, a, s', g) is observed, the UVFA update is:

$$Q(s, a, g) \leftarrow Q(s, a, g) + \eta \left(\mathbb{1}_{\varphi(s)=g} + \gamma \max_{a'} Q(s', a', g) - Q(s, a, g) \right) \tag{5}$$

where η is the learning rate. The only modified value is $Q(s, a, g)$.

With δ-DQN, we learn the density q of $Q(s, a, \mathrm{d}g)$ with respect to $\rho_{\mathcal{G}}$. Assume that $\rho_{\mathcal{G}}(g)$ is the uniform measure over the finite goal space $\mathcal{G}$. Then we learn $q(s, a, g) = |\mathcal{G}| \times Q(s, a, g)$. For a tabular model, the δ-DQN update in Equation (87) is

$$q(s, a, \varphi(s)) \leftarrow q(s, a, \varphi(s)) + \eta \tag{6}$$

$$q(s, a, g) \leftarrow q(s, a, g) + \eta \left(\gamma \max_{a'} q(s', a', g) - q(s, a, g) \right). \tag{7}$$

Here two values are updated: in addition to (s, a, g), the trajectory visiting s is also used to update the value for the goal $\varphi(s)$. The first part always increases q at the goal $\varphi(s)$ achieved by s; the second part at (s, a, g) has no reward contribution, and decreases q at (s, a, g) by a factor $(1 - \eta)$ while propagating the value from s'. In expectation, the decrease at (s, a, g) compensates the increase at $(s, a, \varphi(s))$: this compensation is exact when q is the exact solution.

As a comparison, the tabular HER update works as follows: when observing a trajectory $(g, s_0, a_0, s_1, ...)$, a transition $(s, a, s', g) = (s_K, a_K, s_{K+1}, g)$ for some $K \geqslant 0$ is selected; then HER samples $L \geqslant K$, defines $g' := \varphi(s_L)$ as the re-sampled goal, then applies the UVFA update (5) but with (s, a, s', g') instead of (s, a, s', g). When $L = K$, the goal sampled by HER is $g' = \varphi(s)$: this is somewhat similar to δ-DQN, except δ-DQN resamples an independant goal instead of g' for the second term instead. Despite this similarity, HER is biased in stochastic environments and can converge to a low-return policy, while δ-DQN is unbiased.

3.3 Unbiased Policy Evaluation with Infinitely Sparse Rewards

Similarly to δ-DQN, there exists an actor-critic algorithm for multi-goal environments with infinitely sparse rewards. We start with policy evaluation, then derive the policy gradient algorithm.

Learning the value function $V^\pi(s, \mathrm{d}g)$ directly without bias poses technical issues due to the double dependency of $V^\pi(s, \mathrm{d}g)$ on g (first via the location of the reward, second, via the goal-dependent policy $\pi(.|., g)$). This is discussed in Appendix D.2.

Instead, we learn a richer object, $M^\pi(s, g, \mathrm{d}g')$, the value function of s if the reward is a Dirac at g' but the agent follows the policy $\pi(a|s, g)$ for goal g. This is defined as the measure over goals

$$M^\pi(s, g, \mathrm{d}g') := \mathbb{E}_{a_t \sim \pi(.|s_t, g), s_{t+1} \sim P(.|s_t, a_t)} \left[\sum_{t \geqslant 0} \gamma^t \delta_{\varphi(s_t)}(\mathrm{d}g') | s_0 = s \right] \tag{8}$$

$M^\pi(s, g, \mathrm{d}g')$ represents the *successor goal measure*, and is related to the successor state measure (Blier et al., 2021). Compared to $V^\pi(s, \mathrm{d}g)$, $M^\pi(s, g, \mathrm{d}g')$ splits the two effects of the goal g in two variables g and g'. $V^\pi(s, \mathrm{d}g)$ can be derived from $M^\pi(s, g, \mathrm{d}g')$ as $V^\pi(s, \mathrm{d}g) = M^\pi(s, g, \mathrm{d}g)$ (see Appendix E.2). M^π is a fixed point of the Bellman operator T^π defined as:

$$(T^\pi \cdot M)(s, g, \mathrm{d}g') := \delta_{\varphi(s)}(\mathrm{d}g') + \gamma \mathbb{E}_{a \sim \pi(a|s,g), s' \sim P(\mathrm{d}s'|s,a)} \left[M(s', g, \mathrm{d}g') \right] \tag{9}$$

A rigorous proof of the existence of M^π as well as its fixed point Bellman equation is given in Theorem 13 in the supplementary. Similarly to the δ-DQN update obtained in Theorem 4, we can now derive an unbiased δ-TD update for M^π, leveraging the structure of the Dirac reward and removing the issue of vanishing rewards. As for $Q^*(s, a, \mathrm{d}g)$, because $M^\pi(s, g, \mathrm{d}g')$ is a measure, we learn a model $m_\theta(s, g, g')$ of its density with respect to $\rho_\mathcal{G}$, namely, $M_\theta(s, g, \mathrm{d}g') = m_\theta(s, g, g')\rho_\mathcal{G}(\mathrm{d}g')$.

THEOREM 5. *Let $M_\theta(s, g, \mathrm{d}g') = m_\theta(s, g, g')\rho_\mathcal{G}(\mathrm{d}g')$ be a current estimate of $M^\pi(s, g, \mathrm{d}g')$. Let likewise $M_{\mathrm{tar}}(s, g, \mathrm{d}g') = m_{\mathrm{tar}}(s, g, g')\rho(\mathrm{d}g')$ be a target M, and consider the following update to bring Q_θ closer to $T^\pi Q_{\mathrm{tar}}$ with T^π the Bellman operator.*

Let (s, a, s', g, g') be samples of the environment such that $a \sim \pi(a|s, g)$, $s' \sim P(s'|s, a)$ and $g' \sim \rho_\mathcal{G}$ is a goal sampled independently. Let $\widehat{\delta\theta}_{\delta\text{-TD}}$ be

$$\widehat{\delta\theta}_{\delta\text{-TD}}(s, a, s', g, g') := \partial_\theta m_\theta(s, g, \varphi(s)) + \partial_\theta m_\theta(s, g, g') \left(\gamma m_{\mathrm{tar}}(s', g, g') - m_\theta(s, g, g') \right) \tag{10}$$

Then $\widehat{\delta\theta}_{\delta\text{-TD}}$ is an unbiased estimate of the Bellman error: $\mathbb{E}_{s,a,s',g,g'} \left[\widehat{\delta\theta}_{\delta\text{-TD}}(s, a, s', g, g') \right] = \frac{1}{2}\partial_\theta \|M_\theta - T^\pi M_{\mathrm{tar}}\|^2$, where the norm $\|\cdot\|$ on measures is defined in Theorem 13 (Appendix D.2).

In particular, the true successor goal measure $M^\pi(s, g, \mathrm{d}g')$ is a fixed point of this udpate: if $M_\theta = M_{\mathrm{tar}} = M^\pi$, then $\mathbb{E}\left[\widehat{\delta\theta}_{\delta\text{-TD}}\right] = 0$.

Similarly to update $\widehat{\delta\theta}_{\delta\text{-DQN}}$, the update $\widehat{\delta\theta}_{\delta\text{-TD}}$ has two parts: the first part $\partial_\theta m_\theta(s, g, \varphi(s))$ represents the reward update for the goal achieved in the current state s, and removes the vanishing reward issue. The second part propagates the rewards along transitions.

We can also define a horizon-n δ-TD(n) update if we have access to longer sub-trajectories $\tau = (g, s_0, a_0, s_1, ...)$. The update at a state s_k in the trajectory is (Appendix, Theorem 13)

$$\widehat{\delta\theta}_{\delta\text{-TD}(n)}(\tau, k, g') := \sum_{l=0}^{n-1} \gamma^l \partial_\theta m_\theta(s_k, g, \varphi(s_{k+l})) + \partial_\theta m_\theta(s_k, g, g') \left(\gamma^n m_\theta(s_{k+n}, g, g') - m_\theta(s_k, g, g') \right) \tag{11}$$

where $g' \sim \rho_\mathcal{G}$ is sampled independently. The first part increases the value estimate at state s_k for every of the n goals $\varphi(s_k), ..., \varphi(s_{k+n-1})$ achieved in the next n steps: this corresponds to the n-step return with Dirac rewards. The second part propagates the value along transitions. This is similar to HER in that future goals achieved along the trajectory are explicitly used, and could thus improve sample efficiency. However, computational complexity is an issue. In non-multi-goal environments, algorithms such as PPO (Schulman et al., 2017) compute the TD(n) update at every step of the trajectory. This is computable with $O(n)$ forward passes through the value model v_θ, because it only requires to compute $v_\theta(s_0), \ldots, v_\theta(s_n)$. Here we have to compute $m_\theta(s_k, g, \varphi(s_{k+l}))$ for every k and l, leading to an $O(n^2)$ complexity (though this could potentially be sub-sampled as in HER). This makes it slow in practice, and δ-TD(n) was not tested experimentally here.

3.4 Multi-Goal Policy Gradient

We now derive the actor-critic algorithm. The classical approach with reward R_ε for $\varepsilon > 0$ considers the expected return $J_\varepsilon(\pi) = \mathbb{E}_{g \sim p_\mathcal{G}, s_0 \sim p_0(.|g)} \left[\sum_{t \geqslant 0} \gamma^t R_\varepsilon(s_t, g) | s_0 = s \right] =$

Algorithm 2 One-step δ-Actor-Critic

Input: Model $m_{\theta_M}(s, g)$; policy π_θ; goal function φ; T the maximum trajectory length
Get a goal g and an initial state s_0 from the environment
for $0 \leqslant t \leqslant T$ steps **do**
  Sample $a_t \sim \pi(a|s_t, g)$
  Execute action a_t and observe the next state s_{t+1}
  Sample an independent goal $g' \sim \rho_{\mathcal{G}}(\mathrm{d}g')$
  $\widehat{\delta\theta}_{\delta\text{-TD}} := \partial_\theta m_{\theta_M}(s_t, g, \varphi(s_t)) + \partial_\theta m_{\theta_M}(s_t, g, g')\,(\gamma m_{\theta_M}(s_{t+1}, g, g') - m_{\theta_M}(s_t, g, g'))$
  $\widehat{\delta\theta}_{\delta\text{-AC}} = \gamma^t \times \partial_\theta \log \pi_{\theta_\pi}(a_t|s_t, g)\,(\gamma m(s_{t+1}, g, g) - m(s_t, g, g))$
  $\theta_M \leftarrow \theta_M + \eta_M \widehat{\delta\theta}_{\delta\text{-TD}}$
  $\theta_\pi \leftarrow \theta_\pi + \eta_\pi \widehat{\delta\theta}_{\delta\text{-AC}}$
end for

$\int_{s_0,g} V_\varepsilon^\pi(s, g)\,\rho_{\mathcal{G}}(\mathrm{d}g)\rho_0(\mathrm{d}s_0|g)$ with a sampled goal $g \sim \rho_{\mathcal{G}}(\mathrm{d}g)$ and sampled initial state $s_0 \sim \rho_0(\mathrm{d}s_0|g)$, and subsequent actions sampled from the policy for g. As in δ-DQN, we want to derive an algorithm solving the vanishing reward issue directly for $\varepsilon = 0$. We first show that the limit makes sense, then derive the corresponding update in (13) below.

THEOREM 6. *Under* continuity assumptions *(Assumption 1 in the Appendix), there is a function* $J(\pi)$ *such that, for every parametric policy* $\pi_\theta(a|s, g)$*:*

$$\frac{1}{\lambda(\varepsilon)} J_\varepsilon(\pi_\theta) \to_{\varepsilon\to 0} J(\pi_\theta) \qquad \text{and} \qquad \frac{1}{\lambda(\varepsilon)} \partial_\theta J_\varepsilon(\pi_\theta) \to_{\varepsilon\to 0} \partial_\theta J(\pi_\theta) \tag{12}$$

where $\lambda(\varepsilon)$ *is the volume of a ball of size* ε *in goal space. We call* $J(\pi)$ *the expected return with infinitely sparse rewards. Moreover,* $J(\pi) := \int_{s_0,g} V^\pi(s_0, \mathrm{d}g)\,p_{\mathcal{G}}(g)\,\rho_0(\mathrm{d}s_0|g)$ *where* $p_{\mathcal{G}}(g)$ *is the density of* $\rho_{\mathcal{G}}(\mathrm{d}g)$ *with respect to Lebesgue measure on goals.*

We now derive an estimate of $\partial_\theta J(\pi_\theta)$ for a parametric policy $\pi_\theta(a|s, g)$. We assume access to transition samples (s, a, s', g) such that $a \sim \pi(.|s, g)$, $s' \sim P(\mathrm{d}s'|s, a)$ and s is sampled from the goal-dependant *discounted visitation frequencies* $\nu^\pi(\mathrm{d}s|g) = (1-\gamma)\sum_{t\geqslant 0} \gamma^t \rho_0(\mathrm{d}s_0|g)(P^\pi)^t(\mathrm{d}s|s_0, g)$: namely, states s on a trajectory sampled from π with goal g.

We can define the actor critic update with infinitely sparse rewards by using the model $m(s, g, g)$ as an estimate of the values, and applying the ordinary policy gradient theorem (Sutton & Barto, 2018) on the extended space $\mathcal{S} \times \mathcal{G}$ to include the goals (see Appendix E.4). This leads to

$$\widehat{\delta\theta}_{\delta\text{-AC}}(s, a, s', g) := \partial_\theta \log \pi_\theta(a|s, g)\,(\gamma m_{\theta_M}(s', g, g) - m_{\theta_M}(s, g, g)) \tag{13}$$

where $m_{\theta_M}(s, g, g')$ is the model of the value density learned in Section 3.3. This is justified by the following statement, which is an informal version of Theorem 20 in Appendix E.4: namely, if the value function model m_{θ_M} is correct, then this actor-critic update is an unbiased estimate of $\partial_\theta J(\pi_\theta)$.

INFORMAL THEOREM 7. *If* $m_{\theta_M}(s, g, g)\lambda(\mathrm{d}g)$ *approximates* $V^\pi(s, \mathrm{d}g)$ *as a measure, then* $\mathbb{E}_{s,a,s',g}\left[\widehat{\delta\theta}_{\delta\text{-AC}}(s, a, s', g)\right]$ *approximates* $\partial_\theta J(\pi_\theta)$.

This update, together with the one for m in Theorem 5, make up the δ-Actor-Critic algorithm (Algorithm 2). We can similarly define a PPO algorithm (Appendix A), used in the experiments.

4 Experiments

The `Torus` environment. We first define the `Torus`(n) environment, which is a continuous version of the *flipping coin* environment introduced in (Andrychowicz et al., 2017). The state space is the n-th dimensional torus, represented as $\mathcal{S} = [0, 1)^n$, and can be obtained from the n-dimensional hypercube by gluing the opposite faces together. The action space is $\mathcal{A} = \{1, \ldots, n\} \times \{-\alpha, \alpha\}$ and action $a = (i, u)$ in state s moves the position on the axis i of a quantity u, then the environment adds a Gaussian noise. Formally $s' \sim \big((s + u.e_i + \mathcal{N}(0, \sigma^2)) \mod 1\big)$, where $(e_j)_{1\leqslant j\leqslant n}$ is the canonical basis $(e_i)_k = \mathbb{1}_{i=k}$. We consider the environment in dimensions $n = 4$ and $n = 6$. We

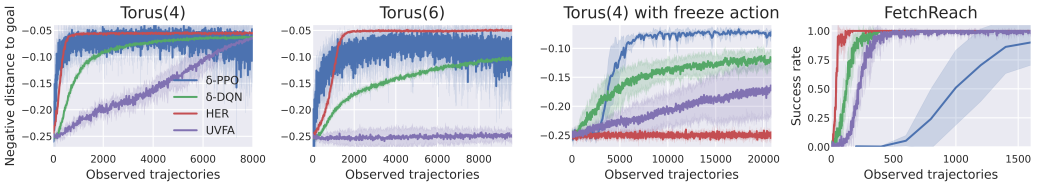

Figure 1: We compare UVFA, HER, δ-DQN in toy environments. We observe different regimes: with a highly stochastic environment (`Torus` with freeze action), HER is unable to learn because of its bias, whereas UVFA and δ-DQN are. When the state dimension becomes too large (`Torus`(6)), UVFA is unable to learn because of the vanishing reward issue. In environments in which HER is able to learn, it is the most efficient method, and δ-DQN is always performing better than UVFA.

also consider the modified environment with the *freeze* action described in Section 2. For every environment, we observe trajectories of length 200, and the reported metric is the rescaled negative L1 distance to the goal at the end of trajectory $-\frac{1}{n}\|s-g\|_1$. The experimental details are in Appendix A.

We compare UVFA, HER, δ-DQN, and δ-PPO (defined in Appendix A based on δ-AC). Each algorithm fails in some environment: additional experiments in the Appendix show that δ-DQN and δ-PPO are both failing to learn when the dimension of the torus increases, while HER is still able to learn. This is discussed in Section 5. While UVFA, HER and δ-DQN are similar algorithms and can be compared as actor-critic methods handle the trajectory samples in a different way from Q-learning methods. Still, we observe that δ-PPO learns successfully in the same environments as δ-DQN, and also failing when δ-DQN does.

The FetchReach environment. The FetchReach environment (Plappert et al., 2018) is a robotic arm environment in which the objective is for the extremity of the arm to reach a given 3D position. The environment is deterministic, so HER is expected to perform well. Here, all methods learn successfully. We also experimented δ-DQN and δ-PPO on more complex environments of the same robotic suite, such as FetchPush, but both methods fail in this setting, while HER was successful.

5 Limitations and Future Work

The algorithms using infinitely sparse rewards always perform better than UVFA, and perform better than HER in environments designed to exhibit the HER bias issue. But they do not perform as well as HER in some standard environments, and are unable to learn at all in more complex environments such as FetchPush. We discuss two technical limitations of δ-DQN and δ-Actor-Critic.

The first issue is the function approximation. Learning the models $Q_\theta(s,a,\mathrm{d}g) = q_\theta(s,a,g)\rho(\mathrm{d}g)$ of Q^* and $M^\pi_\theta(s,g_1,\mathrm{d}g_2) = m_\theta(s,g_1,g_2)\rho(\mathrm{d}g_2)$ of M^π requires approximating a Dirac distribution (when $g_2 = \varphi(s)$) with a continuous density. The theorems justify this, but in practice the functions m_θ and q_θ have to reach multiple orders of magnitude (high values close to the goal, low everywhere else), and the values need to be accurate in these two regimes. Representing multiple orders of magnitude in neural networks may require a well-suited family of parametric functions.

A second issue is variance. The Dirac rewards remove the *infinite* variance of vanishing rewards in UVFA when $\varepsilon \to 0$. But the variance of the remaining term can be high. Consider the tabular case (6)–(7): δ-DQN learns significantly faster than UVFA on the *diagonal* $Q(s,a,g)$ when $g = s$, thanks to the Diracs. But this does not change the way the reward is propagated to other states, due to the independent sampling of g in (7). Selecting goals g more correlated to the state s as in HER could also be helpful, but this is not obvious to do without re-introducing HER-style bias.

6 Conclusion

We have proved that there exist unbiased goal-oriented RL algorithms which do not vanish when rewards become sparse: it is possible to deal with sparse rewards in RL directly via the infinitely sparse reward limit, although this does not solve all variance issues. We have also proved that another multi-goal method, HER, is unbiased and has the correct fixed point in all deterministic environments.

Acknowledgments

We would like to thank Ahmed Touati for his technical help, and Corentin Tallec, Alessandro Lazaric, Nicolas Usunier and Jonathan Laurent for their helpful comments and advice.

References

Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Abbeel, O. P., and Zaremba, W. Hindsight experience replay. In *Advances in neural information processing systems*, pp. 5048–5058, 2017.

Blier, L., Tallec, C., and Ollivier, Y. Learning successor states and goal-dependent values: A mathematical viewpoint. *ArXiv*, abs/2101.07123, 2021.

Bogachev, V. I. *Measure theory*, volume 1. Springer Science & Business Media, 2007.

Fischhoff, B. Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human perception and performance*, 1 (3):288, 1975.

Hausman, K., Springenberg, J. T., Wang, Z., Heess, N., and Riedmiller, M. Learning an embedding space for transferable robot skills. In *International Conference on Learning Representations*, 2018.

Jaderberg, M., Mnih, V., Czarnecki, W. M., Schaul, T., Leibo, J. Z., Silver, D., and Kavukcuoglu, K. Reinforcement learning with unsupervised auxiliary tasks. *arXiv preprint arXiv:1611.05397*, 2016.

Lanka, S. and Wu, T. Archer: Aggressive rewards to counter bias in hindsight experience replay. *ArXiv*, abs/1809.02070, 2018.

Lillicrap, T., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *CoRR*, abs/1509.02971, 2016.

Manela, B. and Biess, A. Bias-reduced hindsight experience replay with virtual goal prioritization. *Neurocomputing*, 451:305–315, 2021.

Nagabandi, A., Clavera, I., Liu, S., Fearing, R., Abbeel, P., Levine, S., and Finn, C. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. *arXiv: Learning*, 2019.

Plappert, M., Andrychowicz, M., Ray, A., McGrew, B., Baker, B., Powell, G., Schneider, J., Tobin, J., Chociej, M., Welinder, P., et al. Multi-goal reinforcement learning: Challenging robotics environments and request for research. *arXiv preprint arXiv:1802.09464*, 2018.

Rauber, P., Mutz, F. W., and Schmidhuber, J. Hindsight policy gradients. *ArXiv*, abs/1711.06006, 2019.

Schaul, T., Horgan, D., Gregor, K., and Silver, D. Universal value function approximators. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1312–1320, Lille, France, 07–09 Jul 2015. PMLR. URL `http://proceedings.mlr.press/v37/schaul15.html`.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347, 2017.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018. 2nd edition.

Wang, Z., Schaul, T., Hessel, M., Hasselt, H. V., Lanctot, M., and Freitas, N. D. Dueling network architectures for deep reinforcement learning. *ArXiv*, abs/1511.06581, 2016.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]

(c) Did you discuss any potential negative societal impacts of your work? [N/A]
(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Notation	Definition
$\mathcal{S}$	State space
$\mathcal{A}$	Action space
γ	Discount factor, $0 \leqslant \gamma < 1$
$P(\mathrm{d}s'\|s,a)$	Transition probability measure, over $\mathcal{S}$, for every $s,a \in \mathcal{S}\times\mathcal{A}$.
$\mathcal{G}$	Goal space
n	Dimension of the goal space
φ	Goal function $\varphi : \mathcal{S} \to \mathcal{G}$. The goal $\varphi(s)$ is the goal *achieved* in s.
$\rho_{\mathcal{G}}(\mathrm{d}g)$	Goal sampling distribution.
$p_{\mathcal{G}}(g)$	If it exists, the density of $\rho_{\mathcal{G}}$ with respect to λ.
$\rho_0(\mathrm{d}s_0\|g)$	Initial state sampling distribution
$p_0(s_0\|g)$	If it exists, the density of ρ_0 with respect to λ.
$\lambda(\mathrm{d}.)$	Lebesgue measure
ε	Threshold for the sparse reward. $\varepsilon > 0$
$\mathbb{1}_A(x)$	Function equal to 1 if $x \in A$, and 0 is $x \notin A$.
$R_\varepsilon(s,g)$	Sparse reward around goal g: $R_\varepsilon(s,g) = \mathbb{1}_{\|\varphi(s)-g\|\leqslant\varepsilon}(s))$
$\lambda(\varepsilon)$	Volume of a sphere of radius ε: $\lambda(\varepsilon) = \lambda(\{x \quad \text{s.t.} \quad \|x\| \leqslant \varepsilon\})$
π	Goal dependent policy. If $\mathcal{A}$ is discrete, $\pi(a\|s,g)$ is the probability of selecting action a. If $\mathcal{A}$ is continuous, it is the density of selecting a with respect to Lebesgue measure.
$P^\pi(\mathrm{d}s'\|s,g)$	Transition probability measure for policy π: $P^\pi(\mathrm{d}s'\|s,g) = \int_a \pi(a\|s,g)P(\mathrm{d}s'\|s,a)$.
τ	Trajectory: $\tau = (g, s_0, a_0, s_1, ...)$, with $g \sim \rho_{\mathcal{G}}$, $s_0 \sim \rho_0(.\|g)$, $a_t \sim \pi(.\|s_t,g)$, $s_{t+1} \sim P(.\|s_t,a_t)$.
$V_\varepsilon^\pi(s,g)$	Value function for reward ε: $V_\varepsilon^\pi(s,g) = \mathbb{E}_{s_{t+1}\sim P^\pi(.\|s_t,g)}\left[\sum_{t\geqslant 0}\gamma^t R_\varepsilon(s_t,g)\|s_0=s\right]$
$Q_\varepsilon^\pi(s,a,g)$	Action-value function for reward ε: $Q_\varepsilon^\pi(s,a,g) = \mathbb{E}_{a_t\sim\pi(.\|s_t,g),s_{t+1}\sim P(.\|s_t,a_t)}\left[\sum_{t\geqslant 0}\gamma^t R_\varepsilon(s_t,g)\|s_0=s,a_0=a\right]$
π^*	Optimal policy
$Q_\varepsilon^*(s,a,g)$	Optimal action-value function for reward R_ε: $Q^* = Q^{\pi^*}$.
$\tilde{\mathcal{S}}, \tilde{P}$	Augmented MDP: $\tilde{\mathcal{S}} = \mathcal{S}\times\mathcal{G}$ and for every $\tilde{s} = (s,g)$, action a, next state $\tilde{s}$ is sampled as $\tilde{s} = (s',g)$ where $s' \sim P(\mathrm{d}s'\|s,a)$.
Q_θ, V_θ	Models of Q^*, V^π parametrized by θ.
Q_{tar}, V_{tar}	Target values
Q_∞	Fixed point of an algorithm
π_{expl}	Exploration policy
T	Optimal Bellman operator, defined for function $Q(s,a,g)$ or measures $Q(s,a,\mathrm{d}g)$ • Functions: $(T\cdot Q)(s,a,g) = R_\varepsilon(s,g) + \gamma\mathbb{E}_{s'\sim P(.\|s,a)}\left[\sup_{a'} Q(s',a',g)\right]$. • Measures: $(T\cdot Q)(s,a,\mathrm{d}g) = \delta_{\varphi(s)}(\mathrm{d}g) + \gamma\mathbb{E}_{s'\sim P(.\|s,a)}\left[\sup_{a'} Q(s',a',\mathrm{d}g)\right]$.
$\widehat{\delta\theta}_{\mathrm{UVFA}}(s,a,s',g)$	Stochastic *Universal Value Function Approximators* update for Q-learning
	Hindsight Experience Replay
K	Step s_K of the update for a trajectory $\tau = (g, s_0, a_0, s_1, ...)$
g'	Re-sampled goal by HER
L	Step of the resampling goal: $g' = \varphi(s_L)$.
$\widehat{\delta\theta}_{\mathrm{HER}}(\tau,K,L)$	HER stochastic update
$\|.\|_{\mathrm{HER}}$	Norm such that $\widehat{\delta\theta}_{\mathrm{HER}}(\tau,K,L)$ is an unbiased estimate of Bellman error
a^*	freeze-after-random-jump additional action

Continuing on next page...

Continued from previous page

Notation	Definition
	Infinitely Sparse Rewards
$\delta_x(\mathrm{d}x')$	Dirac measure located in g
$R(s, \mathrm{d}g)$	Infinitely spare Dirac reward: $R(s, \mathrm{d}g) = \delta_{\varphi(s)}(\mathrm{d}g)$.
$Q^*(s, a, \mathrm{d}g)$	Optimal action-value *measure*
$M^\pi(s, g, \mathrm{d}g')$	Successor goal *measure*: $M^\pi(s, g, \mathrm{d}g') = \mathbb{E}_{a_t \sim \pi(.\|s_t, g), s_{t+1} \sim P(.\|s_t, a_t)} \left[\sum_{t \geqslant 0} \gamma^t \delta_{\varphi(s_t)}(\mathrm{d}g') \| s_0 = s\right]$
$V^\pi(s, \mathrm{d}g)$	Value *measure*: $V^\pi(s, \mathrm{d}g) = M^\pi(s, g, \mathrm{d}g)$
$q_\theta(s, a, g)$	Model of the density of $Q^*(s, a, \mathrm{d}g)$ with respect to $\rho_{\mathcal{G}}$ parametrized by θ
$Q_\theta(s, a, \mathrm{d}g)$	Model of $Q^*(s, a, \mathrm{d}g)$ defined via its density: $Q_\theta(s, a, \mathrm{d}g) = q_\theta(s, a, g)\rho_{\mathcal{G}}(\mathrm{d}g)$
q_{tar}	Target values
$\widehat{\delta\theta}_{\delta\text{-DQN}}(s, a, s', g)$	Stochastic update of $q_\theta(s, a, g)$ for δ-DQN
η	Learning rate
T^π	Bellman operator: $(T^\pi \cdot M)(s, g, \mathrm{d}g') = \delta_{\varphi(s)}(\mathrm{d}g') + \gamma \mathbb{E}_{s' \sim P^\pi(.\|s,g)} [M^\pi(s', g, \mathrm{d}g')]$.
$m_\theta(s, g, g')$	Model of the density of $M^\pi(s, g, \mathrm{d}g')$ with respect to $\rho_{\mathcal{G}}(\mathrm{d}g')$ parametrized by θ
$M_\theta(s, g, \mathrm{d}g')$	Model of $M^\pi(s, g, \mathrm{d}g')$ defined via its density: $M_\theta(s, g, \mathrm{d}g') = m_\theta(s, g, g')\rho_{\mathcal{G}}(\mathrm{d}g)$
$\widehat{\delta\theta}_{\delta\text{-TD}}(s, a, s', g, g')$	Stochastic update of $m_\theta(s, g, g')$ for δ-TD
$\widehat{\delta\theta}_{\delta\text{-TD}(n)}(\tau, k, g')$	Stochastic update of $m_\theta(s, g, g')$ for δ-TD(n)
$J_\varepsilon(\pi)$	Expected return $J_\varepsilon(\pi) = \mathbb{E}_{g \sim \rho_{\mathcal{G}}, s_0 \sim \rho_0(.\|g)} \left[\sum_{t \geqslant 0} \gamma^t R_\varepsilon(s_t, g) \| s_0 = s\right]$
$J(\pi)$	Expected return with infinitely sparse rewards
$\nu^\pi(\mathrm{d}s\|g, s_0)$	Discounted visitation frequencies: $\nu^\pi(\mathrm{d}s\|s_0, g) = (1-\gamma)\sum_{t \geqslant 0} \gamma^t (P^\pi)^t(\mathrm{d}s\|s_0, g)$
θ_M	In policy gradient, parameter of the critic $m_{\theta_M}(s, g, g')$
$\widehat{\delta\theta}_{\delta\text{-AC}}(s, a, s', g)$	Stochastic update for δ-AC

Table 1: Notation in the main text

A Experiments Details

In this section, we present the experiment details of Section 4. Every experiment was performed on a single GPU.

The `Torus(n)` environment The state space of the `Torus`(n) environment is the n-th dimensional torus, $\mathcal{S} = [0, 1)^n$, and can be obtained from the n-dimensional hypercube by gluing the opposite faces together. If the current state is $s = (s_1, ..., s_n)$, we define the observation of the agent as $(\cos(2\pi s_1), ..., \cos(2\pi s_n), \sin(2\pi s_1), ..., \sin(2\pi s_n)) \in [-1, 1]^{2n}$. We use this representation in order to remove the discontinuity of the representation $[0, 1)^n$. This representation contains all the information of the state s and the environment is still fully observable (and not partially observable). The action space is $\mathcal{A} = \{1, \dots, n\} \times \{-\alpha, \alpha\}$ and action $a = (i, u)$ in state s moves the position on the axis i of a quantity u, then the environment adds a Gaussian noise. Formally $s' \sim \big((s + u.e_i + \mathcal{N}(0, \sigma^2)) \mod 1\big)$, where $(e_j)_{1 \leqslant j \leqslant n}$ is the canonical basis $(e_i)_k = \mathbb{1}_{i=k}$. In practice, we take $\alpha = 0.1$, and $\sigma = \frac{0.1}{n}$. The reward is $R_\varepsilon(s, g) = \mathbb{1}_{\|s-g\| \leqslant \varepsilon}$ where $\|.\|$ is the rescaled L1 distance in the Torus: $\|s - g\| = \frac{1}{n}\sum_{i=1}^n \min((s_i - g_i) \mod 1, |((s_i - g_i) \mod 1) - 1|)$. In practice, we use $\varepsilon = 0.05$. At the beginning of an episode, we sample a goal uniformly in the environment, then we observe trajectories of length 200. We set $\gamma = .995$.

FetchReach FetchReach is a standard environment from Plappert et al. (2018). The objective is to reach a *goal* position in 3 dimension with the end of the robotic arm. The observation space $\mathcal{S}$ is of dimension 10 and contains positions and velocities, such that the environment is Markov, fully observable, and deterministic. The action space $\mathcal{A}$ is continuous and of dimension 4. The goal space

$\mathcal{G}$ is of dimension 3, and the goal represent the position of the end of the robotic arm. Trajectories are of length 50.

Q-learning experiments Here we describe experiments with UVFA, HER and δ-DQN, which have similar structure. For every algorithm, we use the same neural network to learn $Q_\theta(s, a, g)$ or $q_\theta(s, a, g)$. Simlarly to DDPG (Lillicrap et al., 2016), if the action space $\mathcal{A}$ is continuous, we additionally learn a deterministic policy $\pi_\theta : \mathcal{S} \times \mathcal{G} \to \mathcal{A}$. We use a dueling architecture (Wang et al., 2016): we learn a *value* network $v_\theta(s, g)$ and an *advantage* network $\mathrm{adv}_\theta(s, a, g)$. We then define $q_\theta(s, a, g) = v_\theta(s, g) + \widetilde{\mathrm{adv}}_\theta(s, a, g)$, where $\widetilde{\mathrm{adv}}_\theta(s, a, g)$ is the *rescaled* advantage, and is defined as $\widetilde{\mathrm{adv}}_\theta(s, a, g) = \mathrm{adv}_\theta(s, a, g) - \frac{1}{|\mathcal{A}|}\sum_{a' \in \mathcal{A}} \mathrm{adv}_\theta(s, a', g)$ if $\mathcal{A}$ is finite, and $\widetilde{\mathrm{adv}}_\theta(s, a, g) = \mathrm{adv}_\theta(s, a, g) - \mathrm{adv}_\theta(s, \pi(s, g), g)$ if $\mathcal{A}$ is continuous. The networks for v_θ, a_θ and π_θ are 3-hidden layers MLP of width 256 and ReLU activations. The inputs of v_θ and π_θ are the concatenation of s and g. If $\mathcal{A}$ is continuous, the input of adv_θ is the concatenation of s, a, g. If $\mathcal{A}$ is discrete, the input of adv_θ is the concatenation of s and g, and its output is of dimension $|\mathcal{A}|$, every dimension corresponding to an action.

Most hypereparameters are shared among the three methods: we observe batchs of trajectories of size 16 for the Torus experiments, and of size 2 for the FetchReach environment. At every epoch, we observe a batch of trajectories and store it in a memory buffer of size 10^6 transitions. We use an ε-greedy exploration strategy, with $\varepsilon = 0.2$. At every epoch, we sample 100 batches from the replay buffer for the Torus experiments, and 50 for the FetchReach environment. For HER, we use the `future` sampling strategy for goals: when sampling a transition (s, a, s', g), with probability 0.2 we define $g' = g$, and with probabiliyt 0.8 we sample g' uniformly in the future of s. For δ-DQN in the Torus environment, we sample independant goals with $\rho_\mathcal{G}$ uniform distribution in the Torus. In FetchReach, we do not assume we have access to the goal sampling distribution. Therefore, we re-sample independant goals from the memory buffer. For every method, observations and goals are normalized. We use a target network with parameter θ_{tar} and update the target as $\theta_{\mathrm{tar}} \leftarrow (1 - \alpha)\theta_{\mathrm{tar}} + \alpha\theta$ with $\alpha = 0.05$ after every epoch. Every model is trained with the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

For every method and environment, the most sensitive hyperparameters were selected with a grid-search. For HER, UVFA and δ-DQN, we selected the learning rate of the optimizer from a range $\{1e-6, 3e-6, 1e-5, 3e-5, 1e-4, 3e-4, 1e-3\}$. For HER and UVFA, we additionally selected R a reward scaling factor, in $\{1e-2, 1e-1, 1, 10, 100, 1000, 1e4\}$. For δ-DQN, we also selected a parameter c_δ corresponding to the scaling of the reward: the scaled infinitely sparse reward is $R(s, \mathrm{d}g) = c_\delta \delta_{\varphi(s)}(\mathrm{d}g)$. We experimented all the possible hyperparameters of this grid separately on every environment on a single run and selected the best hyperparameters. The values in Figure 1 are the mean performance evaluated with 5 different random seeds, and the confidence intervals represent the standard deviation of the reported metric accross the 5 independent runs. In practice, the reward scaling factor for UVFA is 10 for all the Torus environments and 100 for FetchReach. The reward factor is 1 for HER for all the Torus environments and 10 for FetchReach. The learning rate for UVFA is $1e-4$ for all the Torus environments and $1e-3$ for FetchReach. The learning rate for HER is $3e-4$ for all the Torus environments and $1e-3$ for HER. For δ-DQN, the learning rate is $1e-5$ for all the Torus environments, and $1e-4$ for the FetchReach environment. The reward scaling coefficient c_δ is $1e-2$ for every environments.

δ-PPO experiments The δ-PPO is defined from δ-AC similarly to PPO (Schulman et al., 2017) from actor critic methods. We learn the model $m_\theta(s, g, g')$ of the density of $M^\pi(s, g, \mathrm{d}g')$ with respect to $\rho_\mathcal{G}$, and $\pi_\theta(a|s, g)$ a parametric policy. We used a shared architecture: we define $h_\theta(s, g, g')$ a network computing a hidden representation of dimension H. Then, we define two linear layers L^m_θ and $L^\pi\theta$, and define $m_\theta(s, g, g') = L^m_\theta(h_\theta(s, g, g'))$ and $\pi_\theta(a|s, g) = L^\pi_\theta(h_\theta(s, g, g'))$. In practice, h_θ is a 2-hidden layers MLP with ReLU activations (except at the last layer), with width $H = 256$ for the internal and output layers.

A step of δ-PPO is defined as follow. We first gather a buffer of trajectories with the current policy π_θ. Then, we define $\theta' := \theta$. For every transition (s, a, s', g) in the buffer and every epoch $e \leqslant E$,

we sample an independant goal g' and compute:

$$\widehat{\delta\theta}_M \leftarrow \widehat{\delta\theta}_{\delta\text{-TD}}(s, a, s', g, g') \tag{14}$$

$$\text{adv} \leftarrow \gamma m_{\theta_M}(s', g, g) - m_{\theta_M}(s, g, g) \tag{15}$$

$$r(\theta') \leftarrow \frac{\pi_{\theta'}(a|s, g)}{\pi_\theta(a|s, g)} \tag{16}$$

$$\tilde{r}(\theta') \leftarrow \text{clip}(r, 1 - u, 1 + u) \tag{17}$$

$$\widehat{\delta\theta}_\pi \leftarrow \partial_{\theta'} \left(\min\left(\text{adv} \times r(\theta'), \text{adv} \times \tilde{r}(\theta')\right)\right) \tag{18}$$

$$\widehat{\delta\theta} \leftarrow \widehat{\delta\theta}_\pi + c_M \times \widehat{\delta\theta}_M \tag{19}$$

where c_M allow to scale the two updates. Then we use $\widehat{\delta\theta}$ and with Adam optimizer to obtain a new value for θ'. We did not use an entropy regularizer aw we observed that the diversity of actions was not an issue in practice.

For the Torus environment, the independent goals g' are sampled fron $\rho_{\mathcal{G}}$ the uniform distribution of goals in the environment. For FetchReach, we do not assume we know $\rho_{\mathcal{G}}$ and sample goals from the buffer.

In practice, at every step of the δ-PPO algorithm we observe a batch of 2 trajectories for Torus(4) and Torus(6), 100 for the Torus(4) with the freeze action a^*, and 200 for FetchReach. Three hyperparameters were selected independently for every environment via a grid search: E the number of epochs per δ-PPO step, the learning rate of Adam optimizer, and the coefficient c_M. We performed a grid search with a single run per tuple of parameters. Then, the reported results in Figure 1 are averaged over 5 different random seeds with the selected hyperparameters. The number of epoch E per step was selected as lowest number which achieved close-to-optimal performance accross the range $\{1, 2, 5, 10, 20, 50, 100\}$. In practice, $E = 20$ in the Torus(4) and Torus(6) environments, $E = 10$ in the Torus(4) with the freeze action a^*, and $E = 50$ for FetchReach. The learning rate was selected in the set $\{1e{-}6, 3e{-}6, 1e{-}5, 3e{-}5, 1e{-}4, 3e{-}4, 1e{-}3\}$, and in practice is $1e{-}4$ for every environment. The coefficient c_M was selected in $\{1e-4, 1e-3, 1e-2, 1e-1, 1e0, 1e1, 1e2, 1e3\}$ and in practice is $1e-3$ for every Torus environment and $1e-1$ for the FetchReach environment.

Additional experiments We experimented δ-DQN and δ-PPO in more complex environments such Torus of higher dimension, or other environments of OpenAI Robotic suite (Plappert et al., 2018). In the Torus environment, both methods fail when the dimension increases above 15 while HER is still able to learn. More importantly, δ-PPO and δ-DQN did not learn at all in environments such as FetchPush (which is easy to solve with HER) or HandReach, which has similar structure but higher dimension than FetchReach. In the FetchPush environment, the objective is to push a cube with a robotic arm to a given goal. We observed that the issue of our methods was not an exploration issue, since the robotic arm oftens reaches and pushes the cube randomly. We tried to increase the generalization accross goals with the δ-TD(n) update, but it was to computationally expensive, as explained in Section 3.3. Limitations of δ-DQN and δ-PPO which could explain these results are discussed in Section 5.

B Proofs of Theorems on HER

B.1 HER is Unbiased in Deterministic Environments

We prove that HER is an unbiased method in deterministic environments. In order to define HER, we assume access to samples of trajectories $(g, s_0, a_0, s_1, a_1, ...) \sim \rho(g, s_0, a_0, s_1, a_1, ...)$ with $g \sim \rho_{\mathcal{G}}(\mathrm{d}g)$, $s_0 \sim \rho_0(\mathrm{d}s_0|g)$, and for every $k \geqslant 0$, $a_k \sim \pi_{\text{expl}}(a|s_k, g)$ where π_{expl} is an exploration policy, $s_{k+1} \sim P(\mathrm{d}s|s_k, a_k)$. For simplicity, we will assume the trajectories are infinite.

Here we consider HER with the `future` strategy descibed in the original paper: goals are re-sampled from a trajectory as goal reached later in the trajectory. We formalize HER as follows: we sample a trajectory $\tau = (g, s_0, a_0, s_1, a_1, ...) \sim \rho(g, s_0, a_0, s_1, a_1, ...)$, a Bernoulli variable $U \sim \mathcal{B}(\alpha)$, and two independent integer random variables K, L, from distributions p_K and p_L, such that for every k, l, $p_K(k) > 0$ and $p_L(l) > 0$. The bernoulli variable U represents the random choice of using the standard Q-learning update, or the HER update with a resampled goal. The random variable

K represents the timestep of the transition we will use for the Q-learning update, and L represents the timestep used to sample a new goal g' for the `future` sampling strategy. Then, the update $\widehat{\delta\theta_{\text{HER}}}(\tau, U, K, L)$ is defined as:

- If $U = 0$:

$$\widehat{\delta\theta_{\text{HER}}}(\tau, U = 0, K, L) := \partial_\theta \frac{1}{2}(Q_\theta(s_K, a_K, g) - R(s_K, g) - \gamma \sup_{a'} Q(s_{K+1}, a', g))^2,$$

  which corresponds to the usual Q-learning update as defined in UVFA (Schaul et al., 2015).

- If $U = 1$ we set $g' = \varphi(s_{K+L+1})$ and:

$$\widehat{\delta\theta_{\text{HER}}}(\tau, U = 1, K, L) := \partial_\theta \frac{1}{2}(Q_\theta(s_K, a_K, g') - R(s_K, g') - \gamma \sup_{a'} Q(s_K, a', g'))^2,$$

  which corresponds to a Q-learning update for a re-sampled goal $g' = \varphi(s_{K+L})$, a goal achieved later in the trajectory.

We say that environment is a *continuous deterministic environment* if there is a continuous function $\psi : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ such that for every $(s, a) \in \mathcal{S} \times \mathcal{A}$, $P(\mathrm{d}s'|s, a) = \delta_{\psi(s,a)}(\mathrm{d}s')$. In particular, any discrete deterministic environment is a continuous deterministic environment for the discrete topology. Therefore, the following theorem can be applied to discrete environments.

THEOREM 8 (FORMAL STATEMENT OF THEOREM 2). *We assume the environment is a continuous deterministic environment. We also assume that for every pair of states (s, s'), s' is reachable from s, which means there is a sequence of actions $(a_1, ..., a_k)$ such that applying these actions from s leads to s'. Finally, we assume that the support of the exploration policy $\pi_{\text{expl}}(a|s, g)$ is the entire action space $\mathcal{A}$ for every s, g.*

Then, there is an euclidean norm $\|.\|$ such that, for every θ, the HER update with the `future` sampling strategy at θ, $\widehat{\delta\theta_{HER}}$ is an unbiased estimate of the gradient step between Q_θ and the target function $Q_{target} := T_{\max} Q_\theta$:

$$\mathbb{E}\left[\widehat{\delta\theta_{HER}}\right] = \partial_\theta \frac{1}{2}\|Q_\theta - Q^{\text{tar}}\|^2 \tag{20}$$

If the state space $\mathcal{S}$ is finite, HER has a single fixed point Q_∞, which is equal to Q^*.

The euclidean norm $\|.\|$ in the theorem will depend on the exploration policy $\pi_{\text{expl}}(a|s, g)$. Therefore, if the exploration policy is changing during learning, the norm will will be changing as well.

Proof. The principle of the proof is the following. We study the sampling distribution of transitions $\mu_{\text{HER}}(s, a, s', g)$ with HER. The bias of HER comes from the fact that the sampling of goals g with $\mu_{\text{HER}}(s, a, s', g)$ is not independant of s' knowing (s, a). On the contrary, in deterministic environments, the disribution of g knowing (s, a) is independant of s' because s' is uniquely determined by (s, a).

We study the sampling distribution of transitions (s, a, s', g) used in HER. Formally, we sample a transition (s, a, s', g) by sampling τ, U, K, L and defining $(s, a, s', g') := \Phi(\tau, U, K, L)$ as:

- If $U = 0$, $\Phi(\tau, U = 1, K, L) = (s_k, a_k, s_{k+1}, g)$
- If $U = 1$, $\Phi(\tau, U = 1, K, L) = (s_k, a_k, s_{k+1}, \varphi(s_{K+L}))$

Then, HER update can be equivalently defined as: sample (τ, U, K, L) as described above, define $(s, a, s', g) = \Phi(\tau, U, K, L)$, and:

$$\widehat{\delta\theta_{\text{HER}}}(s, a, s', g) := \partial_\theta \frac{1}{2}(Q_\theta(s, a, s', g) - R(s, g) - \gamma \sup_{a'} Q(s', a', g))^2 \tag{21}$$

Therefore:

$$\mathbb{E}\left[\widehat{\delta\theta_{\text{HER}}}\right] = \partial_\theta \mathbb{E}_{(s,a,s',g)\sim\mu_{\text{HER}}} \frac{1}{2}(Q_\theta(s, a, g) - R(s, g) - \gamma \sup_{a'} Q(s', a', g))^2 \tag{22}$$

where we define μ_{HER} to be the distribution of (s, a, s', g) given by the distribution of $\Phi_*(\rho \otimes p_U \otimes p_L \otimes p_K)$, where Φ_* is the *push-forward* operator on measures. We now compute μ_{HER}. Let $f : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathcal{G} \to \mathbb{R}$ be a test function, we have:

$$\mathbb{E}_{s,a,s',g\sim\mu_{\text{HER}}}\left[f(s,a,s',g)\right] = \mathbb{E}_{\tau,U,K,L}\left[f(\Phi(\tau,U,K,L))\right] \tag{23}$$

$$\begin{aligned} &= (1-\alpha)\mathbb{E}_{\tau,U,K,L}\left[f(\Phi(\tau,U,K,L))|U=0\right] \\ &\quad + \alpha\mathbb{E}_{\tau,U,K,L}\left[f(\Phi(\tau,U,K,L))|U=1\right] \end{aligned} \tag{24}$$

Moreover:

$$\mathbb{E}_{\tau,U,K,L}\left[f(\Phi(\tau,U,K,L))|U=0\right] = \sum_K p_K(k) \int_{g,s_0,a_0,\ldots} \rho(g,s_0,a_0,\ldots) f(s_k,a_k,s_{k+1},g) \tag{25}$$

$$= \sum_k p_K(k) \int_{g,s_0,a_0,\ldots} \rho_{\mathcal{G}}(g)\rho_0(s_0|g)(P^{\pi_{\text{exp}}})^k(s|s_0,g)\pi_{\text{expl}}(a|s,g)P(s'|s,a)f(s,a,s',g) \tag{26}$$

$$= \int_{s,a,s',g} f(s,a,s',g)\left(\rho_{\mathcal{G}}(g)\int_{s_0}\sum_k p_K(k)\rho_0(s_0|g)(P^{\pi_{\text{exp}}})^k(s|s_0,g)\pi_{\text{expl}}(a|s,g)P(s'|s,a)\right) \tag{27}$$

$$= \int_{s,a,s',g} f(s,a,s',g)\rho_{\mathcal{G}}(g)\nu(s|g)\pi_{\text{expl}}(a|s,g)P(s'|s,a) \tag{28}$$

with

$$\nu(s|g) := \rho_{\mathcal{G}}(g)\int_{s_0}\rho_0(s_0|g)\sum_k p_K(k)(P^{\pi_{\text{exp}}})^k(s|s_0,g) \tag{29}$$

which is the future distribution of states s when sampling a goal g and following the exploration policy $\pi_{\text{expl}}(.|.,g)$, with p_K as the distribution of future timesteps. If $p_K(k) = (1-\gamma)\gamma^k$, this definition of ν coincides with the definition of ν^π in the following sections. This is the reason why we use the same notation, even though ν is here slightly more general.

We now compute:

$$\mathbb{E}_{\tau,U,K,L}\left[f(\Phi(\tau,U,K,L))|U=1\right] = \sum_{k,l} p_K(k)p_L(l)\int_{g,s_0,a_0,\ldots}\rho(g,s_0,a_0,\ldots)f(s_k,a_k,s_{k+1},\varphi(s_{k+l})) \tag{30}$$

If $l = 0$, the re-sampled goal is $g' = \varphi(s)$. Else, the law of g' knowing s_k, a_k, s_{k+1} is the law of $\varphi(s_{k+l})$, which by using the Markov property is the law of $\varphi(\tilde{s})$ if $\tilde{s}$ is sampled as $(P^{\pi_{\text{expl}}})^{l-1}(.|s_{k+1},g)$. Therefore:

$$\mathbb{E}_{\tau,U,K,L}\left[f(\Phi(\tau,U,K,L))|U=1\right] = \sum_{k,l\geqslant 0} p_K(k)p_L(l)\int_{g,s_0,a_0,\ldots}\rho(g,s_0,a_0,\ldots)f(s_k,a_k,s_{k+1},\varphi(s_{k+l})) \tag{31}$$

$$\begin{aligned} &= \sum_k p_K(k)\int_{g,s_0,\ldots,s_{k+1}}\rho(g,s_0,\ldots,s_{k+1})\left(p_L(0)f(s_k,a_k,s_{k+1},\varphi(s_k))\right) + \\ &\quad + \sum_k p_K(k)\int_{g,s_0,\ldots,s_{k+1}}\rho(g,s_0,\ldots,s_{k+1})\left(\sum_{l\geqslant 1}p_L(l)\int_{\tilde{s}}(P^{\pi_{\text{expl}}})^{l-1}(\tilde{s}|s_{k+1},g)f(s_k,a_k,s_{k+1},\varphi(\tilde{s}))\right) \end{aligned} \tag{32}$$

We define $\mu_{\texttt{future}}(\mathrm{d}g'|s,s',g) := p_L(0)\delta_{\varphi(s)}(\mathrm{d}g') + \sum_{l\geqslant 1} p_L(l)\varphi_*(\pi_{\exp} * P)^{l-1}(g'|s',g)$, where φ_* is the *push-forward* on measures, and we have:

$$\mathbb{E}_{\tau,U,K,L}\left[f(\Phi(\tau,U,K,L))|U=1\right] = \tag{33}$$

$$= \sum_k p_K(k) \int_{g,s_0,a_0,\ldots,s_{K+1},\tilde{s}} \rho(g,s_0,a_0,\ldots,s_{k+1})\mu_{\texttt{future}}(g'|s_k,s_{k+1},g)f(s_k,a_k,s_{k+1},g') \tag{34}$$

$$= \int_{s,a,s',g'} \left(\int_g \rho_{\mathcal{G}}(g)\nu(s|g)\pi_{\mathrm{expl}}(a|s,g)\mu_{\texttt{future}}(g'|s,s',g)\right) P(s'|s,a)f(s,a,s',g'). \tag{35}$$

Therefore,

$$\mu_{\mathrm{HER}}(s,a,s',g) = (1-\alpha)\rho_{\mathcal{G}}(g)\nu(s|g)\pi_{\mathrm{expl}}(a|s,g)P(s'|s,a) + \alpha\left(\int_{\tilde{g}} \rho_{\mathcal{G}}(\tilde{g})\nu(s|\tilde{g})\pi_{\mathrm{expl}}(a|s,\tilde{g})\mu_{\texttt{future}}(g|s,s',\tilde{g})\right) P(s'|s,a) \tag{36}$$

We now use the deterministic hypothesis. We know that for every s,a, $P(\mathrm{d}s'|s,a) = \delta_{\psi(s,a)}(\mathrm{d}s')$. We have, for any s,a:

$$P(\mathrm{d}s'|s,a)\mu_{\texttt{future}}(g|s,s',\tilde{g}) = \delta_{\psi(s,a)}(\mathrm{d}s')\mu_{\texttt{future}}(g|s,s',\tilde{g}) \tag{37}$$

$$= \delta_{\psi(s,a)}(\mathrm{d}s')\mu_{\texttt{future}}(g|s,\psi(s,a),\tilde{g}) \tag{38}$$

Therefore:

$$\mu_{\mathrm{HER}}(s,a,s',g) = \tag{39}$$

$$= (1-\alpha)\rho_{\mathcal{G}}(g)\nu(s|g)\pi_{\mathrm{expl}}(a|s,g)P(s'|s,a) + \alpha\left(\int_{\tilde{g}} \rho_{\mathcal{G}}(\tilde{g})\nu(s|\tilde{g})\pi_{\mathrm{expl}}(a|s,\tilde{g})\mu_{\texttt{future}}(g|s,\psi(s,a),\tilde{g})\right) P(s'|s,a) \tag{40}$$

$$= \tilde{\mu}(s,a,g)P(s'|s,a) \tag{41}$$

where

$$\tilde{\mu}(s,a,g) := (1-\alpha)\rho_{\mathcal{G}}(g)\nu(s,g)\pi_{\mathrm{expl}}(a|s,g) + \alpha\left(\int_{\tilde{g}} \rho_{\mathcal{G}}(g)\nu(s|\tilde{g})\pi_{\mathrm{expl}}(a|s,\tilde{g})\mu_{\texttt{future}}(g|s,\psi(s,a),\tilde{g})\right)$$

Therefore:

$$\mathbb{E}\left[\widehat{\delta\theta_{\mathrm{HER}}}\right] = \partial_\theta \int_{s,a,s',g} \tilde{\mu}(s,a,g)P(s'|s,a)(Q(s,a,g) - R(s,g') - \gamma \sup_{a'} Q(s',a',g))^2 \tag{42}$$

$$= \partial_\theta \int_{s,a,s',g} \tilde{\mu}(s,a,g)\delta_{\psi(s,a)}(s')(Q(s,a,g) - R(s,g') - \gamma \sup_{a'} Q(s',a',g))^2 \tag{43}$$

$$= \partial_\theta \int_{s,a,s',g} \tilde{\mu}(s,a,g)(Q(s,a,g) - R(s,g) - \gamma \sup_{a'} Q(\psi(s,a),a',g))^2 \tag{44}$$

$$= \partial_\theta \int_{s,a,g} \tilde{\mu}(s,a,g)(Q(s,a,g) - R(s,g) - \gamma \mathbb{E}_{s'\sim P(\mathrm{d}s'|s,a)} \sup_{a'} Q(s',a',g))^2 \tag{45}$$

$$= \partial_\theta \int_{s,a,g} \tilde{\mu}(s,a,g)(Q(s,a,g) - T\cdot Q(s,a,g))^2 \tag{46}$$

We define $\|Q\|_{\tilde{\mu}}$ as:

$$\|Q\|_{\tilde{\mu}}^2 := \int_{s,a,g} \tilde{\mu}(s,a,g)Q(s,a,g)^2. \tag{47}$$

We now prove that $\|.\|_{\tilde{\mu}}$ is a norm for the space of continuous functions on $\mathcal{S}\times\mathcal{A}\times\mathcal{G}$. This is equivalent to showing that the support of the probability measure $\tilde{\mu}$, $\mathrm{supp}(\tilde{\mu})$ is equal to $\mathcal{S}\times\mathcal{A}\times\mathcal{G}$. Because

$\tilde{\mu}(s,a,g) \geqslant (1-\alpha)\rho_{\mathcal{G}}(g)\nu(s|g)\pi_{\text{expl}}(a|s,g)$, we know that $\text{supp}(\rho_{\mathcal{G}}(g)\nu(s|g)\pi_{\text{expl}}(a|s,g)) \subset \text{supp}(\tilde{\mu})$. Since for every s,g, $\text{supp}(\pi_{\text{expl}}(a|s,g)) = \mathcal{A}$, $\text{supp}(\rho_{\mathcal{G}}(g)\nu(s|g)\pi_{\text{expl}}(a|s,g)) = \text{supp}(\rho_{\mathcal{G}}(g)\nu(s|g)) \times \mathcal{A}$. Moreover, $\text{supp}\rho_{\mathcal{G}} = \mathcal{G}$. Therefore, we only need to prove that for every g, $\text{supp}(\nu(.|g)) = \mathcal{S}$.

Let $g \in \mathcal{G}$. Because of the definition of ν and because $p_K(k) > 0$ for every k, we have $\text{supp}(\nu(s|g)) = \bigcup_{k\geqslant 0, s_0\in\mathcal{S}} \text{supp}\left((P^{\pi_{\text{expl}}})^k(s|s_0,g)\right)$.

We define the function $\Psi : \mathcal{S}\times\left(\cup_{k\geqslant 1}\mathcal{A}^k\right) \to \mathcal{S}$, corresponding to the action of sequences of action , as follows: for every a, $\Psi(s,a) = \psi(s,a)$, and for every k, $(a_1,...,a_k) \in \mathcal{A}^k$, $\Psi(s,(a_1,...,a_{k+1})) := \psi(\Psi(s,(a_1,...,a_k)),a_{k+1})$. Ψ is continuous. Moreover, we assumed that for any pair of states (s,s'), there is $k \geqslant 0$ and a sequence of actions $(a_0,...,a_k)$ such that applying this sequence of actions from s leads to s'. This means that for every s, $\Psi(s,.)$ is a surjective continuous function.

Moreover, with

$$\text{supp}(P^{\pi_{\text{expl}}})^{k+1}(s|s_0,g) = \cup_{s\in\text{supp}(P^{\pi_{\text{expl}}})^k(s|s_0,g)}\text{supp}\left(\psi(s,\cdot)_*\pi_{\text{expl}}(.|s,g)\right) \tag{48}$$

$$\supseteq \cup_{s\in\text{supp}(P^{\pi_{\text{expl}}})^k(s|s_0,g)}\left(\psi(s,\text{supp}(\pi_{\text{expl}}(.|s,g)))\right) \tag{49}$$

by using the continuity of $\psi(s,.)$. Then:

$$\text{supp}(P^{\pi_{\text{expl}}})^{k+1}(s|s_0,g) \supseteq \cup_{s\in\text{supp}(P^{\pi_{\text{expl}}})^k(s|s_0,g)}\left(\psi(s,\mathcal{A})\right) \tag{50}$$

$$= \psi(\text{supp}(P^{\pi_{\text{expl}}})^k(s|s_0,g)\times\mathcal{A}) \tag{51}$$

By induction, we have: $\text{supp}(P^{\pi_{\text{expl}}})^k(s|s_0,g) \supseteq \Psi(s,\mathcal{A}^k)$. Therefore:

$$\text{supp}(\nu(s|g)) = \bigcup_{k\geqslant 0, s_0\in\mathcal{S}} \text{supp}(P^{\pi_{\text{expl}}})^k(s|s_0,g) \tag{52}$$

$$\supseteq \bigcup_{k\geqslant 0, s_0\in\mathcal{S}} \Psi(s_0,\mathcal{A}^k) \tag{53}$$

$$= \bigcup_{s_0\in\mathcal{S}} \Psi(s_0, \bigcup_{k\geqslant 0}\mathcal{A}^k) \tag{54}$$

$$= \mathcal{S} \tag{55}$$

This concludes the proof. The main property we use in the theorem is that $\mu_{\texttt{future}}(g'|s,s',g)$ is independant of s'. Therefore, a simple way to remove HER bias is to define $p_L(l) = \mathbb{1}_{l=0}$. Still, this would not remove the issue of vanishing rewards, since the fixed point of HER are the same than those of UVFA.

□

In the following, we will use again the results derived above. In particular, we know that:

$$\mathbb{E}\left[\widehat{\delta\theta}_{\text{HER}}\right] = \mathbb{E}_{(s,a,s',g)\sim\mu_{\text{HER}}}\left[\partial_\theta \frac{1}{2}(Q_\theta(s,a,g) - R(s,g) - \gamma \sup_{a'} Q(s',a',g))^2\right] \tag{56}$$

with

$$\mu_{\text{HER}}(s,a,s',g) = (1-\alpha)\rho_{\mathcal{G}}(g)\nu(s|g)\pi_{\text{expl}}(a|s,g)P(s'|s,a) + \alpha\left(\int_{\tilde{g}}\rho_{\mathcal{G}}(\tilde{g})\nu(s|\tilde{g})\pi_{\text{expl}}(a|s,\tilde{g})\mu_{\texttt{future}}(g|s',\tilde{g})\right)P(s'|s,a) \tag{57}$$

$$\mu_{\texttt{future}}(g'|s',g) = \sum_l p_L(l)\varphi_*(\pi_{\text{exp}} * P)^l(g'|s',g) \tag{58}$$

$$\nu(s|g) = \rho_{\mathcal{G}}(g)\int_{s_0}\rho_0(s_0|g)\sum_k p_K(k)(\pi_{\text{exp}} * P)^k(s|s_0,g) \tag{59}$$

B.2 Proof of HER bias

Let $\mathcal{M} = \langle \mathcal{S}, \mathcal{G}, \mathcal{A}, P, R, \gamma \rangle$ be a multi-goal finite Markov Decision Process, with $\mathcal{G} = \mathcal{S}$ and $R(s, g) = \mathbb{1}_{s=g}$. We define $S = |\mathcal{S}|$ the number of states.

Let $\tilde{\mathcal{M}}$ be the augmented MDP with a *freeze* action a^*, defined as:

- The augmented state space $\tilde{\mathcal{S}} = \mathcal{S} \times \{0, 1\}$, where $\tilde{s} = (s, x)$ is said to be frozen if $x = 1$.
- The augmented action space $\tilde{\mathcal{A}} = \mathcal{A} \cup \{a^*\}$, where a^* is the *freeze* action.
- The goal space does not change ($\tilde{\mathcal{G}} = \mathcal{G} = \mathcal{S}$). For an augmented state $\tilde{s} = (s, x)$, the reward is $\tilde{R}(\tilde{s}, g) = \tilde{R}((s, x), g) = R(s, g)$
- If $\tilde{s} = (s, x)$ and $\tilde{s}' = (s', x')$ are two augmented states, the transition operator $\tilde{P}(\tilde{s}'|\tilde{s}, a)$:
  - If the state is frozen ($x = 1$), the agent can't move: $\tilde{P}((s', y)|(s, x), a) = \mathbb{1}_{s'=s}\mathbb{1}_{y=1}$
  - If the state is not frozen ($x = 0$) and $a = a^*$, the agent is sent to a uniformly random frozen state: $\tilde{P}((s', y)|(s, 0), a) = \mathbb{1}_{y=1}\frac{1}{S}$
  - Else, the dynamic is the same than for $\mathcal{M}$: if $x = 0$ and $a \neq a^*$, then $P((s', y)|(s, 0), a) = \mathbb{1}_{y=0}P(s'|s, a)$.

We can now prove the existence of MDPs such that HER will be biased in these environments.

THEOREM 9 (FORMAL STATEMENT OF THEOREM 1). *Let $\mathcal{M}$ be a finite MDP, and $\tilde{\mathcal{M}}$ the augmented MDP with the freeze action a^* defined above. We assume that for every s, a, g the exploration policy satisfies $\pi_{\text{expl}}(a|s, g) > 0$, and that for every every s, g, $\nu(s|g) > 0$, where ν is defined in equation (29). This means that from the given distribution, every state s has a non-zero probability of being reached when following the exploration policy conditioned by g: $\pi_{\text{expl}}(a|s, g)$.*

Let Q_∞ be a fixed point of tabular HER, and Q^* the true optimal Q-function. Then, for every unfrozen state $(s, 0)$ and goal g, HER overstimates the value of action a^*:

$$Q_\infty((s, 0), a^*, g) > Q^*((s, 0), a^*, g) \tag{60}$$

Proof. The principle of the proof is the following. First, we prove that for *frozen* states $\tilde{s} = (s, 1)$, HER converge converge to the true value $Q_\infty(\tilde{s}, a, g) = Q^*(\tilde{s}, a, g)$. Then, we compute the action-value of action a^* for every *unfrozen* state for the true Q^* and for the fixed point Q_∞. HER samples transitions $((s, 0), a^*, (s', 1), g)$. Let us consider the law of s' knowing s, a^*, g: with probability $(1 - \alpha)$ the goal g was re-sampled from the `future` sampling strategy, therefore, because after a^* the position will be frozen, we know that $s' = g$, the goal is reached and the final return is $O(\frac{1}{1-\gamma})$. With probability α, the goal g the original goal, the law of s' is uniform, and the return is of order $O(\frac{1}{S(1-\gamma)})$. Therefore, when estimating the return after action a^* with HER, the computed value will be of order $O(\frac{(1-\alpha)}{1-\gamma})$, while the true value is of order $O(\frac{1}{S(1-\gamma)})$.

We now prove the theorem. We consider Q_∞, a fixed point of the algorithm, which means that starting from Q_∞, the stochastic update defined by HER has mean 0: $\mathbb{E}\left[\widehat{\delta Q}_{\text{HER}}\right] = 0$. We know that

$$\mathbb{E}\left[\widehat{\delta Q}_{\text{HER}}\right] = \mathbb{E}_{(s,a,s',g)\sim\mu_{\text{HER}}}\left[\partial_\theta \frac{1}{2}(Q_\theta(s, a, g) - R(s, g) - \gamma \sup_{a'} Q(s', a', g))^2\right] \tag{61}$$

$$= \mathbb{E}_{(s,a,s',g)\sim\mu_{\text{HER}}}\left[E_{s,a,g}(Q_\theta(s, a, g) - R(s, g) - \gamma \sup_{a'} Q(s', a', g))\right], \tag{62}$$

where $(E_{s,a,g})$ is the canonical basis of the tabular model. Therefore, for every (s, a, g) (because $\mu_{\text{HER}}(s, a, g) > 0$ for every (s, a, g)), we have:

$$Q_\infty(s, a, g) = R(s, g) + \gamma \mathbb{E}_{s'\sim\mu_{\text{HER}}(s'|s,a,g)}\left[\sup_{a'} Q_\infty(s', a', g)\right] \tag{63}$$

First, we prove that the values of frozen states $Q_\infty((s, 1), a, g)$ is equal to the true optimal Q-values. In that case, $\tilde{P}(\tilde{s}'|(s, 1), a) = \delta_{(s,1)}(\tilde{s}')$ we can check that $\mu_{\text{HER}}(\tilde{s}'|(s, 1), a, g) = \delta_{(s,1)}(\tilde{s}')$.

Therefore:

$$Q_\infty((s,1),a,g) = R(s,g) + \gamma \sup_{a'} Q_\infty((s,1),a',g) \tag{64}$$

Therefore for every s, a, g, $Q_\infty((s,1),a,g) = \frac{1}{1-\gamma}R(s,g)$.

Then, we compute the values of $Q_\infty((s,0),a^*,g)$, with the *freeze* action for an unfrozen state. We have:

$$Q_\infty((s,0),a^*,g) = R(s,g) + \gamma \mathbb{E}_{(s',y)\sim\mu_{\mathrm{HER}}((s',y)|(s,0),a^*,g)} \sup_{a'} Q_\infty((s',y),a',g) \tag{65}$$

$$= R(s,g) + \frac{\gamma}{1-\gamma}\mathbb{E}_{(s',y)\sim\mu_{\mathrm{HER}}((s',1)|s,a^*,g)}\left[\mathbb{1}_{s'=g}\right] \tag{66}$$

$$= R(s,g) + \frac{\gamma}{1-\gamma}\mu_{\mathrm{HER}}((s',y)=(g,1)|s,a,g) \tag{67}$$

because $\mu_{\mathrm{HER}}((s',y)|(s,0),a^*,g)$ is non zero only if $y = 1$, and $Q_\infty((s',1),a',g) = \frac{1}{1-\gamma}R(s',g) = \frac{1}{1-\gamma}\mathbb{1}_{s'=g}$. We now compute $\mu_{\mathrm{HER}}((s',y) = (g,1)|s,a,g)$. We use that $P((s',y)|(s,0),a^*) = \mathbb{1}_{y=1}/S$, and $\mu_{\texttt{future}}(g|(s',y)) = \mathbb{1}_{s'}$ if $y = 1$.

$$\mu_{\mathrm{HER}}((s,0),a^*,(s',1),g) = (1-\alpha)\mu_0((s,0),g)\pi_{\mathrm{expl}}(a^*|(s,0),g)\frac{1}{S} + \alpha\left(\int_{\tilde{g}}\mu_0((s,0),\tilde{g})\pi_{\mathrm{expl}}(a^*|(s,0),\tilde{g})\right)\frac{1}{S}\mathbb{1}_{g=s'} \tag{68}$$

Therefore, for every $s' \neq g$: $\mu_{\mathrm{HER}}((s,0),a^*,(s',1),g) < \mu_{\mathrm{HER}}((s,0),a^*,(g,1),g)$. So:

$$\sum_{s'}\mu_{\mathrm{HER}}((s,0),a^*,(s',1),g) < S\mu_{\mathrm{HER}}((s,0),a^*,(g,1),g) \tag{69}$$

and finally $\mu_{\mathrm{HER}}((s',y)=(g,1)|(s,0),a^*,g) > \frac{1}{S}$. Then we have:

$$Q_\infty((s,0),a^*,g) > R(s,g) + \frac{\gamma}{S(1-\gamma)} \tag{70}$$

On the contrary, we can easily check that for any policy π, $Q^\pi((s,0),a^*,g) = R(s,g) + \frac{\gamma}{S(1-\gamma)}$. In particular, by taking $\pi = \pi^*$, we have:

$$Q_\infty((s,0),a^*,g) > Q^*((s,0),a^*,g) \tag{71}$$

□

C Goal-dependent Q-functions in continuous spaces

C.1 Optimal Bellman Operator for action-value measures

With continuous states and goals, in a stochastic environment, the goal-dependent optimal Q-function Q^*_ε with reward $R_\varepsilon(s,g) = \mathbb{1}_{\|\varphi(s)-g\|\leqslant\varepsilon}$ vanishes when $\varepsilon \to 0$: the probability of exactly reaching a goal state is usually 0. Likewise, a direct application of TD would never learn anything because rewards would likely never be observed.

Instead, the goal-dependent Q-function is a *measure* over goals. Intuitively, for every infinitesimally small set of goals $\mathrm{d}g$, the quantity $Q^*(s,a,\mathrm{d}g)$ is the expected amount of time spent in $\mathrm{d}g$ by the policy that tries to maximize time spent in $\mathrm{d}g$, starting at (s,a).

Formally, for every state-action (s,a), $Q^*(s,a,\cdot)$ is a measure over goals, solution to the Bellman equation

$$Q^*(s,a,\mathrm{d}g) = \delta_{\varphi(s)}(\mathrm{d}g) + \gamma\,\mathbb{E}_{s'\sim P(\mathrm{d}s'|s,a)}\max_{a'} Q^*(s',a',\mathrm{d}g) \tag{72}$$

where, as above, $\varphi\colon S \to G$ is the function defining the target features, and where $\delta_{\varphi(s)}$ is the Dirac measure at $\varphi(s)$ in goal space. This is an equality between measures, and the supremum is a supremum of measures (Bogachev, 2007, Section 4.7).

Existence and uniqueness of solutions, and a formal derivation of a TD algorithm, are nontrivial in this setting. Uniqueness never holds without restrictions: the infinite measure always solves (72). But

it is not possible to restrict ourselves to finite-mass measures, because sometimes the solution we want has infinite mass. The need to deal with possibly infinite measures restricts the use of uniqueness proofs by γ-contractivity arguments in some norm.

Intuitively, the total mass $Q^*(s, a, \mathcal{G})$ of the goal state $\mathcal{G}$ describes how much different action sequences result in non-overlapping distributions of states. If the state space $\mathcal{A}$ is finite and $|\mathcal{A}| = A$, the total mass of the horizon-t part of the Q^*-function can be as much as $\gamma^t A^t$: this is realized when every possible sequence of t actions leads to a disjoint part of the state of goals. In Appendix C.3 we provide a simple continuous MDP in which every action sequence leads to a distinct state: as there are an infinite number of action sequences when $t \to \infty$, the total mass $Q^*(s, a, \mathcal{G})$ is infinite.

We still prove the existence of a canonical solution, equal both to the *smallest* solution and to the limit of the horizon-t solution when $t \to \infty$.

THEOREM 10 (FORMAL STATEMENT OF THEOREM 3). *Let $\mathcal{Q}$ be the set of functions from $\mathcal{S} \times \mathcal{A}$ into positive measures over $\mathcal{G}$. Assume the set of actions $\mathcal{A}$ is countable. Let T be the Bellman operator mapping $Q \in \mathcal{Q}$ to $T \cdot Q$ with*

$$T \cdot Q(s, a, \cdot) := \delta_{\varphi(s)}(\cdot) + \gamma \, \mathbb{E}_{s' \sim P(\mathrm{d}s'|s,a)} \sup_{a'} Q(s', a', \cdot) \tag{73}$$

where the supremum is a supremum of measures and $\delta_{\varphi(s)}$ is the Dirac measure at $\varphi(s) \in \mathcal{G}$.

Let $\mathbf{0} \in \mathcal{Q}$ be the measure 0.

Let $Q_t := T^t \mathbf{0}$. (By expanding the definition of T, this is the solution of the expectimax problem at time horizon t.) Then when $t \to \infty$, for every state-action (s, a) and for every measurable set $G \subset \mathcal{G}$, $Q_t(s, a, G)$ converges to a finite or infinite limit $Q^*(s, a, G)$. This limit Q^* is an element of $\mathcal{Q}$ and solves the Bellman equation $TQ^* = Q^*$. It is the smallest such solution. In finite state spaces, it is the only solution with finite mass. Moreover, for any goal-dependent policy π, its Bellman operator T^π and Q-value $Q^\pi := \lim_{t\to\infty} (T^\pi)^t \mathbf{0}$ can be defined similarly (see equation (78)) and satisfy $Q^\pi \leqslant Q^*$ as measures.

Proof. Assume the action space $\mathcal{A}$ is countable. Let $\mathcal{Q}$ be the set of measurable functions from $\mathcal{S} \times \mathcal{A}$ to the set of measures on $\mathcal{G}$.

For Q_1 and Q_2 in $\mathcal{Q}$, we write $Q_1 \leqslant Q_2$ if $Q_1(s, a, X) \leqslant Q_2(s, a, X)$ for any state-action (s, a) and measurable set $X \subset \mathcal{G}$. The Bellman operator of Definition 73 acts on $\mathcal{Q}$ and is obviously monotonous: if $Q_1 \leqslant Q_2$ then $TQ_1 \leqslant TQ_2$.

Since the zero measure $\mathbf{0} \in \mathcal{Q}$ is the smallest measure, we have $T\mathbf{0} \geqslant \mathbf{0}$. Since T is monotonous, by induction we have $T^{t+1}\mathbf{0} \geqslant T^t \mathbf{0}$ for any $t \geqslant 0$. Thus, the $(T^t \mathbf{0})_{t \geqslant 0}$ form an increasing sequence of measures. Therefore, for every state-action (s, a) and measurable set X, the sequence $(T^t \mathbf{0})(s, a, X)$ is increasing, and thus converges to a limit. We denote this limit by $Q^*(s, a, X)$. We have to prove that $Q^* \in \mathcal{Q}$, namely, that for each (s, a), $Q^*(s, a, \cdot)$ is a measure. The only non-trivial point is σ-additivity.

Denote $Q_t := T^t \mathbf{0}$. If (X_i) is a countable collection of disjoint measurable sets, we have

$$\begin{aligned} Q^*(s, a, \cup_i X_i) = \lim_{t\to\infty} Q_t(s, a, \cup_i X_i) &= \lim_{t\to\infty} \sum_i Q_t(s, a, X_i) \\ &= \sum_i \lim_{t\to\infty} Q_t(s, a, X_i) = \sum_i Q^*(s, a, X_i) \end{aligned} \tag{74}$$

where the limit commutes with the sum thanks to the monotone convergence theorem, using that Q_t is non-decreasing. Therefore, Q^* is a measure.

Let us prove that $TQ^* = Q^*$. We have

$$TQ^*(s, a, \cdot) = \delta_{\varphi(s)} + \gamma \mathbb{E}_{s' \sim P(s'|s,a)} \sup_{a'} Q^*(s', a', \cdot) \tag{75}$$

by definition. For any s', denote $\tilde{Q}_t(s', \cdot) := \sup_{a'} Q_t(s', a', \cdot)$ where the supremum is as measures over $\mathcal{G}$. Since Q_t is non-decreasing, so is $\tilde{Q}_t$.

For any state s', we have

$$\sup_{a'} Q^*(s', a', \cdot) = \sup_{a'} \sup_t Q_t(s', a', \cdot) = \sup_t \sup_{a'} Q_t(s', a', \cdot) = \sup_t \tilde{Q}_t(s', \cdot) \tag{76}$$

since supremums commute. Now, since $\tilde{Q}_t$ is non-decreasing, thanks to the monotone convergence theorem, the supremum commutes with integration over $s' \sim P(s'|s,a)$ (which does not depend on t), namely,

$$\mathbb{E}_{s'\sim P(s'|s,a)} \sup_{a'} Q^*(s',a',\cdot) = \mathbb{E}_{s'\sim P(s'|s,a)} \sup_t \tilde{Q}_t(s',\cdot)$$
$$= \sup_t \mathbb{E}_{s'\sim P(s'|s,a)} \tilde{Q}_t(s',\cdot) = \sup_t \mathbb{E}_{s'\sim P(s'|s,a)} \sup_{a'} Q_t(s',a',\cdot) \quad (77)$$

and so $TQ^* = \sup_t TQ_t$. Now, since $Q^t = T^t\mathbf{0}$, we have $TQ^t = T^{t+1}\mathbf{0}$, so that $\sup_{t\geqslant 0} TQ^t = \sup_{t\geqslant 1} T^t\mathbf{0} = Q^*$. So Q^* is a fixed point of T.

Let us prove that Q^* is the smallest such fixed point. Let Q' such that $TQ' = Q'$. Since $\mathbf{0} \leqslant Q'$ and T is monotonous, we have $T\mathbf{0} \leqslant TQ' = Q'$. By induction, $T^t\mathbf{0} \leqslant Q'$ for any $t \geqslant 0$. Therefore, $\sup_t T^t\mathbf{0} \leqslant Q'$, i.e., $Q^* \leqslant Q'$.

The statement for finite state spaces reduces to the classical uniqueness property of the usual Q function, separately for each goal state.

Optimality of the policy is proved by following classical arguments. Let $\pi(a|s,g)$ be any goal-dependent policy and let $Q \in \mathcal{Q}$. Define the Bellman operator associated to π by

$$(T^\pi Q)(s,a,\cdot) := \delta_s + \gamma \mathbb{E}_{s'\sim P(s'|s,a)} \sum_{a'} (\pi * Q)(s',a',\cdot) \quad (78)$$

where for each action a, the measure $(\pi * Q) \in \mathcal{Q}$ is defined via $(\pi * Q)(s',a',X) := \int_{g\in X} \pi(a'|s',g) Q(s',a',\mathrm{d}g)$, so that the sum of $(\pi * Q)$ over all actions a' represents the expected value of $Q(s',a',\cdot)$ under the goal-dependent policy π; this formulation allows the policy to depend on the goal.

Since π is a probability distribution, we have

$$\sum_{a'} (\pi * Q)(s',a',X) \leqslant \max_{a'} Q(s',a',X) \quad (79)$$

where the right-hand-side is a maximum of measures (thus selecting the best a' for each goal): this is clear from decomposing X into the components where each action a' is optimal.

Therefore, for any $Q \in \mathcal{Q}$, we have the inequality of measures

$$T^\pi Q \leqslant TQ \quad (80)$$

where T is the optimal Bellman operator from above. Since the latter is monotonous over $Q \in \mathcal{Q}$, for any $Q, Q' \in \mathcal{Q}$ with $Q \leqslant Q'$, we have $T^\pi Q \leqslant TQ'$.

Consequently, by induction, $(T^\pi)^t\mathbf{0} \leqslant T^t\mathbf{0}$ for any horizon $t \geqslant 0$. The monotonous limit $Q^\pi := \lim_{t\to\infty} (T^\pi)^t\mathbf{0}$ exists for the same reasons as $T^t\mathbf{0}$, representing the Q-function (measure) of policy π. Therefore, $Q^\pi = \lim_{t\to\infty} (T^\pi)^t\mathbf{0} \leqslant \lim_{t\to\infty} T^t\mathbf{0} = Q^*$. This proves that the policy π has returns no greater than Q^*.

□

C.2 Parametric goal-dependent Q-learning.

In this section, we formally derive the δ-DQN update introduced in Section 3.2. Let us consider parametric models for Q:

$$Q_\theta(s,a,\mathrm{d}g) := q_\theta(s,a,g)\rho_{\mathcal{G}}(\mathrm{d}g) \quad (81)$$

and we will learn q_θ. [1]

The resulting parametric update is off-policy: we assume access to a sampling distribution $(s,a,s') \sim \rho_{\mathrm{SA}}(\mathrm{d}s,\mathrm{d}a)P(\mathrm{d}s'|s,a)$ in a Markov decision. Typically, this can correspond to transitions (s_k,a_k,s_{k+1}) from exploration trajectory with $g \sim \rho_{\mathcal{G}}$, $s_0 \sim \rho_0(.|g)$, then $a_t \sim \pi_{\mathrm{expl}}(.|s_t,g)$

[1] The factor $\rho_{\mathcal{G}}$, or some other measure, is needed to get a well-defined object in continuous state spaces. In discrete spaces, it results in an g-dependent scaling of the Q function, which still has the same optimal policy for each g.

and $s_{t+1} \sim P(.|s_t, a_t)$. Here, our statement with a distribution ρ_{SA} is more general. Given a measure-valued function of (s, a), such as $Q(s, a, \mathrm{d}g)$, we define its norm as

$$\|Q\|^2_{\rho_{\mathrm{SA}},\rho_{\mathcal{G}}} := \mathbb{E}_{(s,a)\sim\rho_{\mathrm{SA}},\, g\sim\rho_{\mathcal{G}}}[q(s, a, g)^2] \tag{82}$$

where $q(s, a, g) := Q(s, a, \mathrm{d}g)/\rho_{\mathcal{G}}(\mathrm{d}g)$ is the density of Q with respect to $\rho_{\mathcal{G}}$, if it exists (otherwise the norm is infinite).

Let $Q_\theta = q_\theta(s, a, g)\rho(\mathrm{d}g)$ be our current estimate of Q, and $Q_{\mathrm{tar}}(s, a, \mathrm{d}g) = q_{\mathrm{tar}}(s, a, g)\rho_{\mathcal{G}}(\mathrm{d}g)$ a target measure, we define the loss:

$$J_Q(\theta) := \|Q_\theta - T \cdot Q_{\mathrm{tar}}\|^2_{\rho_{\mathrm{SA}},\rho_{\mathcal{G}}} \tag{83}$$

where T is the optimal Bellman operator, and our goal is to obtain an unbiased estimate of $\partial J_Q(\theta)$. In the statement of Theorem 4, there is a hidden mathematical subtlety with continuous states regarding the norm $\|Q_\theta - T \cdot Q_{\mathrm{tar}}\|^2_{\rho_{\mathrm{SA}},\rho_{\mathcal{G}}}$. Indeed, $Q_\theta(s, a, \mathrm{d}g) = q_\theta(s, a, g)\rho_{\mathcal{G}}(\mathrm{d}g)$ is absolutely continuous with respect to $\rho_{\mathcal{G}}$, while $T \cdot Q_{\mathrm{tar}}$ is not, due to the Dirac term $\delta_{\varphi(s)}(\mathrm{d}g)$. This makes the norm $\|Q_\theta - T \cdot Q_{\mathrm{tar}}\|^2_{\rho_{\mathrm{SA}},\rho_{\mathcal{G}}}$ infinite (see its definition in (82)). However, the *gradient* of this norm is actually still well-defined. There are at least two ways to handle this rigorously, which lead to the same result. It is possible to do the computation in the finite state space case and observe that the resulting gradient still makes sense in the continuous case (which can be obtained by a limiting argument). The other way we will use here, is to observe that the loss $J_Q(\theta)$ is equal to

$$J_Q(\theta) = \frac{1}{2}\|Q_\theta\|^2_{\rho_{\mathrm{SA}},\rho_{\mathcal{G}}} - \langle Q_\theta, TQ_{\mathrm{tar}}\rangle_{\rho_{\mathrm{SA}},\rho_{\mathcal{G}}} + \frac{1}{2}\|T \cdot Q_{\mathrm{tar}}\|^2_{\rho_{\mathrm{SA}},\rho_{\mathcal{G}}} \tag{84}$$

where

$$\langle Q_1, Q_2\rangle_{\rho_{\mathrm{SA}},\rho_{\mathcal{G}}} := \int_{s,a} Q_1(s, a, \mathrm{d}g)Q_2(s, a, \mathrm{d}g)\frac{1}{\rho_{\mathcal{G}}(\mathrm{d}g)}. \tag{85}$$

Even though $\|Q_1 - Q_2\|_{\rho_{\mathrm{SA}},\rho_{\mathcal{G}}}$ is finite only if Q_1 and Q_2 are both absolutely continuous with respect to $\rho_{\mathcal{G}}(\mathrm{d}g)$, the dot product $\langle Q_1, Q_2\rangle_{\rho_{\mathrm{SA}},\rho_{\mathcal{G}}}$ is still defined if only one of Q_1 or Q_2 is absolutely continuous. Therefore, we can define:

$$J'_Q(\theta) = \frac{1}{2}\|Q_\theta\|^2_{\rho_{\mathrm{SA}},\rho_{\mathcal{G}}} - \langle Q_\theta, T \cdot Q_{\mathrm{tar}}\rangle_{\rho_{\mathrm{SA}},\rho_{\mathcal{G}}} \tag{86}$$

For a given Q_{tar}, $J'_Q(\theta)$ and $J_Q(\theta)$ have the same minima and gradients, but $J'_Q(\theta')$ is always well defined and finite. Namely, J_Q and J'_Q differ by a constant in the finite case, and by an "infinite constant" in the continuous case. We will work with the loss J'_Q, which is finite even in the continuous case.

THEOREM 11 (FORMAL STATEMENT OF THEOREM 4). *Let $Q_\theta(s, a, \mathrm{d}g) = q_\theta(s, a, g)\rho_{\mathcal{G}}(\mathrm{d}g)$ be a current estimate of $Q^*(s, a, \mathrm{d}g)$. Let likewise $Q_{\mathrm{tar}}(s, a, \mathrm{d}g) = q_{\mathrm{tar}}(s, a, g)\rho_{\mathcal{G}}(\mathrm{d}g)$ be a target Q-function. We consider the loss function $J'_Q(\theta)$ defined in equation* (86).

We consider the following update to bring Q_θ closer to TQ_{tar} with T the optimal Bellman operator: Let $(s, a, s') \sim \rho_{SA}(\mathrm{d}s, \mathrm{d}a)P(s'|s, a)$ be samples of the environment and $g \sim \rho_{\mathcal{G}}$ sampled independently. Let $\widehat{\delta\theta}_{\delta\text{-DQN}}(s, a, s', g)$ be

$$\widehat{\delta\theta}_{\delta\text{-DQN}}(s, a, s', g) := \partial_\theta q_\theta(s, a, \varphi(s)) + \partial_\theta q_\theta(s, a, g)\left(\gamma \max_{a'} q_{\mathrm{tar}}(s', a', g) - q_\theta(s, a, g)\right) \tag{87}$$

Then $\widehat{\delta\theta}_{\delta\text{-DQN}}$ is an unbiased estimate of $\partial_\theta J'_Q(\theta)$: $\mathbb{E}\left[\widehat{\delta\theta}_{\delta\text{-DQN}}\right] = -\partial_\theta J'_Q(\theta)$.

In particular, the true optimal state-action measure Q^* is a fixed point of this update: if $Q_\theta = Q_{\mathrm{tar}} = Q^*$ then $\mathbb{E}\left[\widehat{\delta\theta}_{\delta\text{-DQN}}\right] = 0$.

Here we have presented the update using a fixed "target network" with parameter θ_0 (typically a previous value of θ), a common practice for parametric Q-learning.

For this theorem, we sample goals g independently of (s, a, s'). In practice, this could be a source of variance, as sampling goals far from the current state should produce close-to-0 Q-values. If we instead sample goals from a distribution $\mu(g|s, a)$, this introduces an implicit *scaling* factor $\alpha(s, g)$ to the reward. This is discussed in details and the end of Appendix E.2 in the case of the V-function.

Proof. By definition of the optimal Bellman operator T and the target Q_{tar}, we have:

$$TQ_{\mathrm{tar}}(s,a,\mathrm{d}g) = \delta_{\varphi(s)}(\mathrm{d}g) + \gamma\, \mathbb{E}_{s'\sim P(s'|s,a)} \left[\sup_{a'} q_{\mathrm{tar}}(s',a',g)\right] \rho_{\mathcal{G}}(\mathrm{d}g) \tag{88}$$

By definition of $J'_Q(\theta)$ and of the norm $\|\cdot\|_{\rho_{\mathrm{SA}},\rho_{\mathcal{G}}}$, we have

$$\begin{aligned} J'_Q(\theta) &= \frac{1}{2}\|Q_\theta\|^2_{\rho_{\mathrm{SA}},\rho} - \langle Q_\theta, TQ_{\mathrm{tar}}\rangle_{\rho_{\mathrm{SA}},\rho} && (89)\\ &= \frac{1}{2}\int_{s,a,g} q_\theta^2(s,a,g)\rho_{\mathrm{SA}}(\mathrm{d}s,\mathrm{d}a)\rho(\mathrm{d}g) - \int_{s,a,g} q_\theta(s,a,g)(T\cdot Q_{\mathrm{tar}})(s,a,\mathrm{d}g)\rho_{\mathrm{SA}}(\mathrm{d}s,\mathrm{d}a) && (90)\end{aligned}$$

Consequently,

$$\begin{aligned} \partial_\theta J'(\theta) &= \int_{s,a,g} \partial_\theta q_\theta(s,a,g) q_\theta(s,a,g)\rho_{\mathrm{SA}}(\mathrm{d}s,\mathrm{d}a)\rho_{\mathcal{G}}(\mathrm{d}g) - \int_{s,a,g} \partial_\theta q_\theta(s,a,g) TQ_{\mathrm{tar}}(s,a,\mathrm{d}g)\rho_{\mathrm{SA}}(\mathrm{d}s,\mathrm{d}a) && (91)\\ &= \int_{s,a,g} \rho_{\mathrm{SA}}(\mathrm{d}s,\mathrm{d}a)\partial_\theta q_\theta(s,a,g)\left(Q_\theta(s,a,\mathrm{d}g) - TQ_{\mathrm{tar}}(s,a,\mathrm{d}g)\right) && (92)\end{aligned}$$

assuming q_θ is smooth enough so that the derivative makes sense and commutes with the integral. Moreover, we have:

$$TQ_{\mathrm{tar}}(s,a,\mathrm{d}g) - Q_\theta(s,a,\mathrm{d}g) = \delta_{\varphi(s)}(\mathrm{d}g) + \gamma\, \mathbb{E}_{s'\sim P(s'|s,a)}[\sup_{a'} q_{\mathrm{tar}}(s',a',g) - q_\theta(s,a,g)]\rho_{\mathcal{G}}(\mathrm{d}g) \tag{93}$$

Therefore,

$$\begin{aligned} -\partial_\theta J'(\theta) &= \int_{s,a} \rho_{\mathrm{SA}}(\mathrm{d}s,\mathrm{d}a)\partial_\theta q_\theta(s,a,g)\delta_{\varphi(s)}(\mathrm{d}g) \\ &\quad + \int_{s,a,g} \rho_{\mathrm{SA}}(\mathrm{d}s,\mathrm{d}a)\rho_{\mathcal{G}}(\mathrm{d}g)\left(\gamma\, \mathbb{E}_{s'\sim P(s'|s,a)}[\sup_{a'} q_{\theta_0}(s',a',g) - q_\theta(s,a,g)]\right) && (94)\\ &= \int_{s,a} \rho_{\mathrm{SA}}(\mathrm{d}s,\mathrm{d}a)\partial_\theta q_\theta(s,a,\varphi(s)) \\ &\quad + \int_{s,a,g} \rho_{\mathrm{SA}}(\mathrm{d}s,\mathrm{d}a)\rho_{\mathcal{G}}(\mathrm{d}g)\left(\gamma\, \mathbb{E}_{s'\sim P(s'|s,a)}[\sup_{a'} q_{\theta_0}(s',a',g) - q_\theta(s,a,g)]\right) && (95)\end{aligned}$$

By definition of $\widehat{\delta\theta_Q}$, we have:

$$\begin{aligned} \mathbb{E}_{s,a\sim\rho_{\mathrm{SA}}}\left[\widehat{\delta\theta_Q}\right] &= \mathbb{E}_{s,a\sim\rho_{\mathrm{SA}},g\sim\rho_{\mathcal{G}}(\mathrm{d}g)}\left[\partial_\theta q_\theta(s,a,\varphi(s))\right] \\ &\quad + \mathbb{E}_{s,a\sim\rho_{\mathrm{SA}},g\sim\rho_{\mathcal{G}}(\mathrm{d}g)}\left[\partial_\theta q_\theta(s,a,g)\left(\gamma \sup_{a'} q_{\theta_0}(s',a',g) - q_\theta(s,a,g)\right)\right] && (96)\\ &= -\partial_\theta J'(\theta) && (97)\end{aligned}$$

Finally, if $Q_{\mathrm{tar}} = Q_\theta = Q^*$, then $TQ_{\mathrm{tar}} = Q^*$ and:

$$\begin{aligned} \partial J'_Q(\theta) &= \int_{s,a,g} \rho_{\mathrm{SA}}(\mathrm{d}s,\mathrm{d}a)\partial_\theta q_\theta(s,a,g)\left(Q_\theta(s,a,\mathrm{d}g) - TQ_{\mathrm{tar}}(s,a,\mathrm{d}g)\right) && (98)\\ &= 0 && (99)\end{aligned}$$

□

C.3 Examples of MDPs with Infinite Mass for Q^*

Here are two simple examples of MDPs with finite action space, for which the mass of the goal-dependent Q-measure $Q^*(s, a, \mathrm{d}g)$ is infinite. The first has discrete states, the second, continuous ones.

Take for $\mathcal{S}$ an infinite rooted dyadic tree, namely, $\mathcal{S} = \{\varnothing, 0, 1, 00, 01, \ldots\}$ the set of binary strings of finite length $k \geqslant 0$, and $\mathcal{G} = \mathcal{S}$. Consider the two actions "add a 0 at the end" and "add a 1 at the end". Then, for every state s, $Q^*(s, a, \cdot)$ is a measure that gives mass γ^k to all states g that are extensions of s by a length-k string that starts with a. Thus, its mass is $1 + \sum_{k\geqslant 1} \gamma^k 2^{k-1}$. This is infinite as soon as $\gamma \geqslant 1/2$. This extends to any number of actions by considering higher-degree trees.

A similar example with continuous states is obtained as follows. Let $\mathcal{S} = [0; 1) \times [0; 1)$. Let $C = \{\varnothing, 0, 1, 00, 01, \ldots\}$ the dyadic tree above. For each string $w \in X$, consider the set $B_w \subset \mathcal{S}$ defined as follows: B_w is made of those points $(x, y) \in \mathcal{S}$ such that the binary expansion of x starts with w, and $y \in [1 - 1/2^k; 1 - 1/2^{k+1})$ where k is the length of w. Graphically, this creates a tree-like partition of the square $\mathcal{S}$, where the empty string corresponds to the bottom half, the strings $w = 0$ and $w = 1$ correspond to two sets on the left and right above the bottom hald, etc. Define the following MDP with two actions 0 and 1: with action 0, every state $s \in B_w$ goes to a uniform random state in B_{w0}, and with action 1, every state $s \in B_w$ goes to a uniform random state in B_{w1}. The goal-dependent Q-function Q^* is similar to the dyadic tree above, but is continuous. Its mass is infinite for the same reasons.

D The successor goal measure $M(s, g, \mathrm{d}g')$

D.1 Definition and existence of the successor goal measure

THEOREM 12. *The* successor *measure*

$$\nu^\pi(\mathrm{d}s|s_0, g) = (1-\gamma)\sum_{k\geqslant 0} \gamma^k (P^\pi)^k(\mathrm{d}s|s_0, g) \tag{100}$$

is a well defined probability measure over $\mathcal{S}$ for every s_0, g. It satisfies the fixed-point equation:

$$\nu^\pi(\mathrm{d}s|s_0, g) = (1-\gamma)\delta_{s_0}(\mathrm{d}s) + \mathbb{E}_{a\sim\pi(\mathrm{d}a|s_0,g), s_1\sim P(\mathrm{d}s_1|s_0,a)}\left[\nu^\pi(\mathrm{d}s|s_0, g)\right] \tag{101}$$

The successor-goal measure *is defined as:*

$$M^\pi(s, g, .) := \frac{1}{1-\gamma}\varphi_*\nu^\pi(.|s, g) \tag{102}$$

where φ_* is the push-forward *operator on measures. We define the Bellman operator mapping $M(s, g_1, \mathrm{d}g2)$ to $T_\pi M$ with*

$$(T_\pi \cdot M)(s, g_1, \mathrm{d}g_2) = \delta_\varphi(s)(\mathrm{d}g_2) + \gamma\mathbb{E}_{a\sim\pi(a|s,g_1), s'\sim P(\mathrm{d}s'|s,a)}\left[M(s', g_1, \mathrm{d}g_2)\right], \tag{103}$$

Then, M^π is a fixed point of T^π.

Proof. For every k, $(P^\pi)^k(\mathrm{d}s|s_0, g)$ is a probability measure over $\mathcal{S}$. Therefore, for any measurable set $S \subset \mathcal{S}$, the sum $(1-\gamma)\sum_{k\geqslant 0}\gamma^k(P^\pi)^k(S|s_0, g) \leqslant 1$, and the sum converges. $\nu^\pi(\mathrm{d}s|s_0, g)$ is a positive measure as a convergent sum of positive measure. Its total mass is $(1-\gamma)\sum_{k\geqslant 0}\gamma^k(P^\pi)^k(\mathcal{S}|s_0, g) = (1-\gamma)\sum_{k\geqslant 0}\gamma^k = 1$. Therefore, $\nu^\pi(\mathrm{d}s|s_0, g)$ is a well-defined probability measure.

We now prove the fixed point equation. We have:

$$\nu^\pi(\mathrm{d}s|s_0, g) = (1-\gamma)(P^\pi)^0(\mathrm{d}s|s_0, g) + (1-\gamma)\sum_{k\geqslant 1}\gamma^k(P^\pi)^k(\mathrm{d}s|s_0, g) \tag{104}$$

$$= (1-\gamma)\delta_{\varphi(s_0)}(\mathrm{d}s) + \left(P^\pi\left((1-\gamma)\sum_{k\geqslant 1}\gamma^k(P^\pi)^{k-1}\right)\right)(\mathrm{d}s|s_0, g) \tag{105}$$

$$= (1-\gamma)\delta_{\varphi(s_0)}(\mathrm{d}s) + \gamma\left(P^\pi * \nu^\pi\right)(\mathrm{d}s|s_0, g) \tag{106}$$

$$= (1-\gamma)\delta_{s_0}(\mathrm{d}s) + \mathbb{E}_{a\sim\pi(\mathrm{d}a|s_0,g), s_1\sim P(\mathrm{d}s_1|s_0,a)}\left[\nu^\pi(\mathrm{d}s|s_0, g)\right] \tag{107}$$

where $(P^\pi * \nu^\pi)(\mathrm{d}s|s_0, g) := \int_{s_1} P^\pi(\mathrm{d}s_1|s_0, g)\nu^\pi(\mathrm{d}s|s_1, g)$. We now show the Bellman fixed point equation of M^π. We now that

$$\nu^\pi(\mathrm{d}s|s_0, g) = (1-\gamma)\delta_{s_0}(\mathrm{d}s) + \mathbb{E}_{a\sim\pi(\mathrm{d}a|s_0,g), s_1\sim P(\mathrm{d}s_1|s_0,a)}\left[\nu^\pi(\mathrm{d}s|s_0, g)\right] \tag{108}$$

By applying the push-forward operator φ_* we have:

$$(1-\gamma)M^\pi(s_0, g, \mathrm{d}g') = (1-\gamma)\delta_{s_0}(\mathrm{d}s) + \varphi_*\left(\int_{a,s_1} \pi(a|s_0, g)P(\mathrm{d}s_1|s_0, a)\nu^\pi(\mathrm{d}s|s_1, g)\right) \tag{109}$$

$$= (1-\gamma)\delta_{s_0}(\mathrm{d}s) + \left(\int_{a,s_1} \pi(a|s_0, g)P(\mathrm{d}s_1|s_0, a)\varphi_*\nu^\pi(\mathrm{d}s|s_1, g)\right) \tag{110}$$

$$= (1-\gamma)\delta_{s_0}(\mathrm{d}s) + \int_{a,s_1} \pi(a|s_0, g)P(\mathrm{d}s_1|s_0, a)M^\pi(s_1, g, \mathrm{d}g') \tag{111}$$

□

D.2 The Policy Evaluation Update

In this section, we prove Theorem 5 for learning M^π via temporal differences algorithm, TD and $TD^{(n)}$. This theorem very similar to Theorem 11. We directly prove the result for $\mathrm{TD}^{(n)}$, as the standard TD update stated in Theorem 5 corresponds to the $\mathrm{TD}^{(n)}$ update for $n = 1$.

The resulting parametric update is on-policy: Let π be a policy, we assume access to a sampling distribution $(g, s_0, ..., , s_n) \sim \rho_{\mathrm{SG}}(\mathrm{d}g, \mathrm{d}s_0)P^\pi(\mathrm{d}s_1|s_0, g)...P^\pi(\mathrm{d}s_n|s_{n-1}, g)$, where ρ_{SG} is any distribution on $\mathcal{S}\times\mathcal{G}$. Typically, this can correpond to couples (g, s_k) from trajectory with $g \sim \rho_{\mathcal{G}}$, $s_0 \sim \rho_0(.|g)$, $s_{t+1} \sim P^\pi(.|s_t, g)$. Here, our statement with a distribution ρ_{SG} is more general.

Given a measure-valued function of (s, s), such as $M(s, g, \mathrm{d}g')$, we define its norm as

$$\|M\|^2_{\rho_{\mathrm{SG}},\rho_{\mathcal{G}}} := \mathbb{E}_{(s,g)\sim\rho_{\mathrm{SG}},\, g'\sim\rho_{\mathcal{G}}}[m(s, g, g')^2] \tag{112}$$

where $m(s, g, g') := M(s, g, \mathrm{d}g')/\rho_{\mathcal{G}}(\mathrm{d}g')$ is the density of $M^\pi(s, g, .)$ with respect to $\rho_{\mathcal{G}}$, if it exists (otherwise the norm is infinite).

THEOREM 13 (FORMAL STATEMENT OF THEOREM 5). *Let $M_\theta(s, g, \mathrm{d}g') = m_\theta(s, g, g')\rho_{\mathcal{G}}(\mathrm{d}g')$ be a current estimate of $M^\pi(s, g, \mathrm{d}g')$. Let likewise $M_{\mathrm{tar}}(s, g, \mathrm{d}g') = m_{\mathrm{tar}}(s, g, g')\rho_{\mathcal{G}}(\mathrm{d}g')$ be a target M, and consider the following update to bring M_θ closer to $(T^\pi)^n M_{\mathrm{tar}}$ with T^π the Bellman operator.*

Let $\tau = (g, s_0, ..., s_n) \sim \rho_{\mathrm{SG}}(\mathrm{d}g, \mathrm{d}s_0)P^\pi(\mathrm{d}s_1|s_0, g)...P^\pi(\mathrm{d}s_n|s_{n-1}, g)$ be a sample of the environment and $g' \sim \rho_{\mathcal{G}}$ is a goal sampled independently. Let $\widehat{\delta\theta}_{\delta\text{-TD}(n)}$ be

$$\widehat{\delta\theta}_{\delta\text{-TD}(n)}(\tau, g') := \sum_{l=0}^{n-1} \gamma^l \partial_\theta m_\theta(s_0, g, \varphi(s_l)) + \partial_\theta m_\theta(s_0, g, g')\left(\gamma^n m_\theta(s_n, g, g') - m_\theta(s_n, g, g')\right) \tag{113}$$

Then $\widehat{\delta\theta}_{\delta\text{-TD}(n)}$ is an unbiased estimate of the Bellman error: $\mathbb{E}_{\tau,g'}\left[\widehat{\delta\theta}_{\delta\text{-TD}}(\tau, g')\right] = \frac{1}{2}\partial_\theta\|M_\theta - (T^\pi)^n M_{\mathrm{tar}}\|^2_{\rho_{\mathrm{SG}},\rho_{\mathcal{G}}}$.

In particular, the true M^π is a fixed point of this udpate: if $M_\theta = M_{\mathrm{tar}} = M^\pi$, then

$$\mathbb{E}\left[\widehat{\delta\theta}_{\delta\text{-TD}(n)}\right] = 0 \tag{114}$$

For this theorem, we sample goals g' independently of τ. In practice, this could be a source of variance, as sampling goals far from the current state should produce close-to-0 V-values. If we instead sample goals from a distribution $\mu(g|s, a)$, this introduces an implicit *scaling* factor $\alpha(s, g)$ to the reward. This is discussed in details and the end of Appendix E.2 in the case of the V-function.

As in Appendix C.2, there is a hidden mathematical subtlety with continuous states regarding the norm $\|M_\theta - (T^\pi)^n M_{\mathrm{tar}}\|$, which is infinite because $(T^\pi)^n M_{\mathrm{tar}}$ is not absolutely continuous with

respect to $\rho_{\mathcal{G}}$. However, as in Appendix C.2, the *gradient* of $\|M_\theta - (T^\pi)^n M_{\mathrm{tar}}\|$ is finite. Because the rigorous way to handle it is exactly the same technique as in Appendix C.2, we will not derive it in this section.

Proof. The proof is very similar to the proof of Theorem 11. Similarly to the derivation of (92), we have:

$$-\partial_\theta \frac{1}{2}\partial_\theta \|M_\theta - (T^\pi)^n M_{\mathrm{tar}}\|^2_{\rho_{\mathrm{SG}},\rho_{\mathcal{G}}} = \int_{s_0,g,g'} \rho_{\mathrm{SG}}(\mathrm{d}s_0,\mathrm{d}g)\partial_\theta m(s_0,g,g')((T^\pi)^n M_{\mathrm{tar}}(s,g,\mathrm{d}g') - M_\theta(s,g,\mathrm{d}g')) \tag{115}$$

Moreover:

$$\begin{aligned}(T^\pi)^n M^{\mathrm{tar}}(s,g,\mathrm{d}g') - M_\theta(s,g,\mathrm{d}g') = &\sum_{k=0}^{n-1} \gamma^k \mathbb{E}_{s_1,\dots,s_k|s_0,g}\left[\delta_{\varphi(s_k)}(\mathrm{d}g)\right] \\ &+ \left(\gamma^n \mathbb{E}_{s_n|s_0,g}\left[m_{\mathrm{tar}}(s_n,g,g')\right] - m_\theta(s_0,g,g')\right)\rho_{\mathcal{G}}(\mathrm{d}g')\end{aligned} \tag{116}$$

Therefore:

$$\begin{aligned}-\partial_\theta J(\theta) = &\int_{s_0,g,g'} \rho_{\mathrm{SG}}(\mathrm{d}s_0,\mathrm{d}g)\partial_\theta m(s_0,g,g') \sum_{k=0}^{n-1} \gamma^k \mathbb{E}_{s_1,\dots,s_k|s_0,g}\left[\delta_{\varphi(s_k)}(\mathrm{d}g)\right] \\ &+ \int_{s_0,g,g'} \rho_{\mathrm{SG}}(\mathrm{d}s_0,\mathrm{d}g)\partial_\theta m(s_0,g,g')\left(\gamma^n \mathbb{E}_{s_n|s_0,g}\left[m_\theta(s_n,g,g')\right] - m_\theta(s_0,g,g')\right)\rho_{\mathcal{G}}(\mathrm{d}g')\end{aligned} \tag{117}$$

$$\begin{aligned}-\partial_\theta J(\theta) = &\int_{s_0,g,g'} \rho_{\mathrm{SG}}(\mathrm{d}s_0,\mathrm{d}g)\mathbb{E}_{s_1,\dots,s_n|s_0,g}\left[\sum_{k=0}^{n-1} \gamma^k \partial_\theta m(s_0,g,\varphi(s_k))\right] \\ &+ \int_{s_0,g,g'} \rho_{\mathrm{SG}}(\mathrm{d}s_0,\mathrm{d}g)\partial_\theta m(s_0,g,g')\left(\gamma^n \mathbb{E}_{s_n|s_0,g}\left[m_\theta(s_n,g,g')\right] - m_\theta(s_0,g,g')\right)\rho_{\mathcal{G}}(\mathrm{d}g')\end{aligned} \tag{118}$$

Therefore, $\mathbb{E}_{\tau,g'}\left[\widehat{\delta\theta}_{\delta\text{-TD}}(\tau,g')\right] = \frac{1}{2}\partial_\theta \|M_\theta - (T^\pi)^n M_{\mathrm{tar}}\|^2_{\rho_{\mathrm{SG}},\rho_{\mathcal{G}}}$.

Finally, if $M_\theta = M_{\mathrm{tar}} = M^\pi$, we have:

$$-\partial_\theta \frac{1}{2}\partial_\theta \|M_\theta - (T^\pi)^n M_{\mathrm{tar}}\|^2_{\rho_{\mathrm{SG}},\rho_{\mathcal{G}}} = \int_{s_0,g,g'} \rho_{\mathrm{SG}}(\mathrm{d}s_0,\mathrm{d}g)\partial_\theta m(s_0,g,g')((T^\pi)^n M_{\mathrm{tar}}(s,g,\mathrm{d}g') - M_\theta(s,g,\mathrm{d}g')) \quad = 0 \tag{119}$$

This concludes the proof. □

E The continuous density setting

E.1 The continuous density assumption

Here, we introduce the continuity assumption, which will be used in this section, to formalize the relation between the multi-goal formulation with infinitely sparse Dirac rewards with the standard formulation with reward located in a neighborhood of size ε around the goal, and to derive a policy gradient theorem.

ASSUMPTION 1. *We assume that $\mathcal{S}$ and $\mathcal{G}$ are finite dimensional vector spaces, and that $\mathcal{A}$ is a compact subset of a finite dimensional vector space. Moreover, $\rho_{\mathcal{G}}(\mathrm{d}g)$ is absolutely continuous with respect to the Lebesgue measure on $\mathcal{G}$, and we write $p_{\mathcal{G}}$ its density: $p_{\mathcal{G}}(g)\lambda(\mathrm{d}g)$, where $p_{\mathcal{G}}$ is a continuous function. Similarly, $\rho(\mathrm{d}s_0|g)$ the distribution of initial states given a goal is supposed to be absolutely continuous with respect the Lebesgue measure: $\rho(\mathrm{d}s_0|g) = p_0(s_0|g)\lambda(\mathrm{d}g)$, with p_0 continuous. The transition probability measure $P(\mathrm{d}s'|s,a)$ is absolutely continuous with respect to the Lebesgue measure on $\mathcal{S}$, and we write $p(s'|s,a)$ its density, which is continuous.*

We assume that $\mathrm{supp}\rho_{\mathcal{G}}$ *is compact and that there is a compact subset* $K_{\mathcal{S}} \subset \mathcal{S}$ *such that for every* $s, a \in \mathcal{S}, \mathcal{G}$, $\mathrm{supp}P(\mathrm{d}s'|s,a) \subset K_{\mathcal{S}}$.

We consider only policies in Π, *the set of policies* π *such that* $\pi(a|s,g)$ *is a continuous function of* a, s, g.

We assume $\dim \mathcal{G} \leqslant \mathcal{S}$ *and* φ *is a surjective linear function, and* $\varphi(\mathcal{S}) = \mathcal{G}$.

Let us comment Assumption 1. First, we require P, $\rho_{\mathcal{G}}$, and ρ_0 to be absolutely continuous with respect to Lebesgue measure. This is typically true in environments such that, at every step, the environment adds a noise absolutely continuous with respect to Lebesgue measure (for instance Gaussian) to the position. On the contrary, in environments such that the agent lies in a submanifold of dimension lower than $\dim \mathcal{S}$, the assumption is not satisfied. In the `Torus(n)` environment, with the state representation $s \in [0,1)^n$, the environment satisfies this assumption. But with the representation used in the experiments $\tilde{s} = (\cos(2\pi s_1), \sin(2\pi s_1), ..., \cos(2\pi s_n), \sin(2\pi s_n)) \in [-1,1]^{2n}$, this assumption is not satisfied. The assumption that φ is linear is often satisfied in practice, when the achieved goal of a state corresponds to a some coordinates of s. For instance, in FetchReach, the state s contains information on the position and velocity of the robotic arm, while the achieved goal is the position of the extremity of the robotic arm. This assumption could be generalized to φ a submersion (a differentiable function such that its $\mathrm{d}\varphi_s$ is surjective for every s), but we used the linear assumption for the simplicity of the proof.

Under this assumption, we have the following lemma on the probability distribution ν^π introduced in Appendix D.1 and M^π:

LEMMA 14. *Under Assumption 1, there is a function* $q^\pi(s|s_0,g)$ *such that for any* (s_0,g)*:*

$$\nu^\pi(\mathrm{d}s|s_0,g) = (1-\gamma)\delta_{s_0}(\mathrm{d}s) + q^\pi(s|s_0,g)\lambda(\mathrm{d}s) \tag{120}$$

and $q^\pi(s|s_0,g)$ *is a continuous function of* s, s_0, g.

Moreover, $M^\pi(s,g,\mathrm{d}g') = \varphi_*\big(\nu^\pi(\mathrm{d}s'|s,g)\big)(\mathrm{d}g')$ *(where* φ_* *is the push-forward operator) and there is a function* $\tilde{m}^\pi(s,g,g')$ *such that for any* s, g*:*

$$M^\pi(s,g,\mathrm{d}g') = \delta_{\varphi(s)} + \tilde{m}^\pi(s,g,g')\lambda(\mathrm{d}g'). \tag{121}$$

and $\tilde{m}^\pi(s,g,g')$ *is a continuous function of* (s,g,g').

The function $\tilde{m}^\pi$ *satisfies for every* $(s,g,g') \in K_{\mathcal{S}} \times \mathcal{G} \times \mathcal{G}$ *the fixed point equation:*

$$\tilde{m}(s,g,g') = \gamma \int_a \lambda(\mathrm{d}a)\pi(a|s,g)\left(\tilde{p}(g'|s,a) + \int_{s'} \lambda(\mathrm{d}s')p(s'|s,a)\tilde{m}^\pi(s',g,g')\right) \tag{122}$$

Proof. We have:

$$\nu^\pi(\mathrm{d}s|s_0,g) = (1-\gamma)\sum_{k\geqslant 0}\gamma^k(P^\pi)^k(\mathrm{d}s|s_0,g) \tag{123}$$

We know that

$$(P^\pi)(\mathrm{d}s'|s,g) = \lambda(\mathrm{d}s')\int_a \lambda(\mathrm{d}a)\pi(a|s,g)p(s'|s,a),$$

and by induction, for $k \geqslant 1$,

$$(P^\pi)^k(\mathrm{d}s|s_0,g) = \lambda(\mathrm{d}s)\int_{a_0,\dots,s_{k-1},a_{k-1}} \pi(a_0|s_0,g)\left(\prod_{i=1}^{k-1} p(s_i|s_{i-1},a_{i-1})\pi(a_i|s_i,g)\right)p(s|s_{k-1},a_{k-1}).$$

We define:

$$q^\pi(s|g,s_0) := (1-\gamma)\sum_{k\geqslant 1}\gamma^k \int_{a_0,\dots,s_{k-1},a_{k-1}} \pi(a_0|s_0,g)\left(\prod_{i=1}^{k-1} p(s_i|s_{i-1},a_{i-1})\pi(a_i|s_i,g)\right)p(s|s_{k-1},a_{k-1}) \tag{124}$$

We now check that q^π is well-defined and continuous. For every $k \geqslant 1$, the function

$$(g, s_0, a_0, ..., s_{k-1}, a_{k-1}, s) \mapsto \pi(a_0|s_0, g) \left(\prod_{i=1}^{k-1} p(s_i|s_{i-1}, a_{i-1}) \pi(a_i|s_i, g) \right) p(s|s_{k-1}, a_{k-1})$$

is continuous and the supports of π are p compact sets. Therefore, for every $k \geqslant 0$, the function

$$(g, s_0, s) \mapsto \int_{a_0, s_1, ..., s_{k-1}, a_{k-1}} \pi(a_0|s_0, g) \left(\prod_{i=1}^{k-1} p(s_i|s_{i-1}, a_{i-1}) \pi(a_i|s_i, g) \right) p(s|s_{k-1}, a_{k-1})$$

is well defined and continuous.

Moreover, for every $k \geqslant 0$, and (s, g):

$$\left| \gamma^k \int_{a_0, ..., s_{k-1}, a_{k-1}} \pi(a_0|s_0, g) \left(\prod_{i=1}^{k-1} p(s_i|s_{i-1}, a_{i-1}) \pi(a_i|s_i, g) \right) p(s|s_{k-1}, a_{k-1}) \right| \leqslant \tag{125}$$

$$\leqslant \gamma^k \int_{a_0, ..., s_{k-1}, a_{k-1}} \pi(a_0|s_0, g) \left(\prod_{i=1}^{k-1} p(s_i|s_{i-1}, a_{i-1}) \pi(a_i|s_i, g) \right) \|p\|_\infty \tag{126}$$

$$= \gamma^k \|p\|_\infty \tag{127}$$

and $\sum_{k \geqslant 0} \gamma^k \|p\|_\infty \leqslant \infty$. Therefore, $q^\pi(s|g, s_0)$ is a continuous function and we have:

$$\nu^\pi(\mathrm{d}s|s_0, g) = (1 - \gamma)\delta_{s_0}(\mathrm{d}s) + q^\pi(s|s_0, g)\lambda(\mathrm{d}s). \tag{128}$$

Moreover, the support of ν^π is compact and for every $s_0 \in K_{\mathcal{S}}$, we have $\mathrm{supp}\,(\nu^\pi(.|s_0, g)) \subset K_{\mathcal{S}}$.

We now show the existence of $\tilde{m}^\pi$. We have:

$$M^\pi(s, g, \mathrm{d}g') = \frac{1}{1-\gamma} \big(\varphi_* \nu^\pi(\mathrm{d}s'|s, g)\big)(\mathrm{d}g') = \varphi_* \left((\delta_s(\mathrm{d}s')) \, (\mathrm{d}g') + \frac{1}{1-\gamma} \varphi_* \left(q^\pi(s'|s, g)\lambda(\mathrm{d}s') \right) (\mathrm{d}g') \right.$$

First, $\varphi_*(\delta_s) = \delta_{\varphi(s)}$. Then, we study the second part $\varphi_* \left(q^\pi(s'|s, g)\lambda(\mathrm{d}s') \right) (\mathrm{d}g')$, and show that there is a continuous function $\tilde{m}(s, g, g')$ such that

$$\frac{1}{1-\gamma} \varphi_* \left(q^\pi(s'|s, g)\lambda(\mathrm{d}s') \right) (\mathrm{d}g') = \tilde{m}(s, g, g')\lambda(\mathrm{d}g') \tag{129}$$

Let $f(g)$ be a continuous test function. We have:

$$\int_{g' \in \mathcal{G}} f(g') \varphi_* \left(q^\pi(s'|s, g)\lambda(\mathrm{d}s) \right) (\mathrm{d}g') = \int_{s'} f(\varphi(s')) q^\pi(s'|s, g)\lambda(\mathrm{d}s') \tag{130}$$

We use the change of variable $s' = e + k$ with $k \in \mathrm{Ker}\,\varphi$ and $e \in \mathrm{Ker}\,\varphi^\perp$ and use that $\varphi(s') = \varphi(e)$, and $\varphi_{\mathrm{Ker}\,\varphi^\perp}$ the restriction of φ to $\mathrm{Ker}\,\varphi^\perp$ is invertible. In order to use continuity theorems on integrals, we want to restrict the integral domains to compact sets. We define the orthogonal projections of $K_{\mathcal{S}}$ on $\mathrm{Ker}\,\varphi$ and $\mathrm{Ker}\,\varphi^\perp$: $K = \mathrm{proj}_{\mathrm{Ker}\,\varphi}(\mathrm{K}_{\mathcal{S}})$ and $E = \mathrm{proj}_{\mathrm{Ker}\,\varphi^\perp}(\mathrm{K}_{\mathcal{S}})$. K and E are compact sets and $\mathrm{supp}\,(q^\pi(s'|s, g)) \subset \{e + k \ , \ (k, e) \in K \times E\}$ for every $s \in K_{\mathcal{S}}$. We have:

$$\int_{g' \in \mathcal{G}} f(g') \varphi_* \left(q^\pi(s'|s, g)\lambda(\mathrm{d}s') \right) (\mathrm{d}g') = \int_{e \in \mathrm{Ker}\,\varphi^\perp, k \in \mathrm{Ker}\,\varphi} f(\varphi(e + k)) q^\pi(e + k|s, g)\lambda(\mathrm{d}e, \mathrm{d}k) \tag{131}$$

$$= \int_{e \in E, k \in K} f(\varphi(e + k)) q^\pi(e + k|s, g)\lambda(\mathrm{d}e, \mathrm{d}k) \tag{132}$$

$$= \int_{e \in E} f(\varphi(e))\lambda(\mathrm{d}e) \int_{k \in K} q^\pi(e + k|s, g)\lambda(\mathrm{d}k) \tag{133}$$

where we can switch integrals because the sets are compact and the functions continuous. We use the change of variable: $g' = (\varphi_{|\,\mathrm{Ker}\,\varphi^\perp})^{-1}(e)$. For simplicity, we use the notation $\varphi^{-1} = (\varphi_{|\,\mathrm{Ker}\,\varphi^\perp})^{-1}$.

$$\int_{g' \in \mathcal{G}} f(g') \varphi_* \left(q^\pi(s'|s, g)\lambda(\mathrm{d}s') \right) (\mathrm{d}g') = \int_{g' \in \mathcal{G}} f(g')\lambda(\mathrm{d}g') \left(\det(\varphi^{-1}) \int_{k \in K} \mathbb{1}_E(\varphi^{-1}(g') q^\pi(\varphi^{-1}(g') + k|s, g)\lambda(\mathrm{d}k) \right) \tag{134}$$

$$= \int_{g' \in \mathcal{G}} f(g')\lambda(\mathrm{d}g') \left(\det(\varphi^{-1}) \int_{k \in K} q^\pi(\varphi^{-1}(g') + k|s, g)\lambda(\mathrm{d}k) \right) \tag{135}$$

where the last line is obtained by using that $\mathbb{1}_E(s')q^\pi(s'|.,.) = q^\pi(s'|.,.)$ because $s' \notin E \Rightarrow q^\pi(s'|.,.) = 0$. We define $\tilde{m}^\pi(s,g,g') = \frac{1}{1-\gamma}\det(\varphi)^{-1}\int_{k\in K} q^\pi(\varphi^{-1}(g')+k|s,g)\lambda(\mathrm{d}k)$. The function $(s,g,k,g') \to q^\pi(\varphi^{-1}(g')+k|s,g)$ is continuous and K is compact. Therefore, $\tilde{m}^\pi$ is continuous and bounded, and:

$$\frac{1}{1-\gamma}\varphi_*\left(q^\pi(s'|s,g)\lambda(\mathrm{d}s')\right)(\mathrm{d}g') = \tilde{m}^\pi(s,g,g')\lambda(\mathrm{d}g')$$

Moreover, the support of $\tilde{m}(s,g,g')\lambda(\mathrm{d}g')$ is compact and $\mathrm{supp}\left(\tilde{m}^\pi(s,g,g')\lambda(\mathrm{d}g')\right) \subset \varphi(K_{\mathcal{S}})$.

We now prove the fixed point equation on $\tilde{m}^\pi$. We consider the Bellman equation on $M^\pi(s,g,\mathrm{d}g')$. We have:

$$M^\pi(s,g,\mathrm{d}g') = \delta_{\varphi(s)}(\mathrm{d}g') + \gamma\int_{s',a}\lambda(\mathrm{d}s',\mathrm{d}a)\pi(a|s,g)p(s'|s,a)M^\pi(s',g,\mathrm{d}g') \tag{136}$$

By using $M^\pi(s,g,\mathrm{d}g') = \delta_{\varphi(s)}(\mathrm{d}g') + \tilde{m}^\pi(s,g,g')\lambda(\mathrm{d}g')$, we have:

$$\tilde{m}^\pi(s,g,g')\lambda(\mathrm{d}g') = \gamma\int_{s',a}\lambda(\mathrm{d}s',\mathrm{d}a)\pi(a|s,g)p(s'|s,a)\left(\delta_{\varphi(s')}(\mathrm{d}g') + \tilde{m}(s,g,g')\lambda(\mathrm{d}g')\right) \tag{137}$$

Let $f(g')$ be a continuous test function, we have:

$$\int_{g'} f(g')\tilde{m}^\pi(s,g,g')\lambda(\mathrm{d}g') = \tag{138}$$

$$= \gamma\int_{s',a,g'}\lambda(\mathrm{d}s',\mathrm{d}a)f(g')\pi(a|s,g)p(s'|s,a)\left(\delta_{\varphi(s')}(\mathrm{d}g') + \tilde{m}(s,g,g')\lambda(\mathrm{d}g')\right) \tag{139}$$

$$= \gamma\int_{s',a}\lambda(\mathrm{d}s',\mathrm{d}a)\pi(a|s,g)p(s'|s,a)\left(f(\varphi(s')) + \int_{g'}\lambda(\mathrm{d}g')f(g')\pi(a|s,g)p(s'|s,a)\tilde{m}(s',g,g')\right) \tag{140}$$

$$= \gamma\int_{a,g'} f(g')\pi(a|s,g)\tilde{p}(g'|s,a) + \gamma\int_{a,s',g'}\lambda(\mathrm{d}a,\mathrm{d}s',\mathrm{d}g')f(g')\pi(a|s,g)p(s'|s,a)\tilde{m}(s',g,g')f(g') \tag{141}$$

where $\tilde{p}(g|s,a)$ is the density with respect to Lebesgue measure $\lambda(\mathrm{d}g)$ of $\varphi_*P(\mathrm{d}s'|s,a)$. Formally, the existence proof of $\tilde{p}$ is the same than for $\tilde{m}$ in equation (129), and is using the fact that P is continuous with respect to $\lambda(\mathrm{d}s)$ and φ is a surjective linear operator. Therefore, we have, for λ-almost s,g,g':

$$\tilde{m}(s,g,g') = \gamma\int_a\lambda(\mathrm{d}a)\pi(a|s,g)\left(\tilde{p}(g'|s,a) + \gamma\int_{s'}\lambda(\mathrm{d}s')p(s'|s,a)\tilde{m}^\pi(s',g,g')\right) \tag{142}$$

Because $\tilde{m}^\pi$ is continuous, this relation is true for every s,g,g', in particular if $g = g'$. □

E.2 The Value Measure Under the Continuous Density Assumption

Under Assumption 1, we can rigorously define the value measure $V^\pi(s,\mathrm{d}g)$ as follows. Then, we briefly show why learning directly $V^\pi(s,\mathrm{d}g)$ without bias poses technical issues as states in Section 3.3, which is the reason why we learn M^π.

THEOREM 15. *Under Assumption 1, we can define the value-*measure $V^\pi(s,\mathrm{d}g)$ *as the measure on* $\mathcal{G}\times\mathcal{G}$*:*

$$V^\pi(s,\mathrm{d}g) = \delta_{\varphi(s)}(\mathrm{d}g) + \tilde{m}(s,g,g)\lambda(\mathrm{d}g) \tag{143}$$

The value measure V^π *satisfies the fixed point equation:*

$$V^\pi(s,\mathrm{d}g) = \delta_{\varphi(s)}(\mathrm{d}g) + \gamma\mathbb{E}_{s'\sim P(\mathrm{d}s'|s,g)}\left[V^\pi(s',\mathrm{d}g)\right] \tag{144}$$

Finally, the value-measure is consistent with the value function $V^\varepsilon(s,g)$ *when* $\varepsilon\to 0$*. Formally, the measure on* $K_{\mathcal{S}}\times\mathcal{G}$*:* $\lambda(\mathrm{d}s)\frac{1}{\lambda(\varepsilon)}V^\pi_\varepsilon(s,g)\lambda(\mathrm{d}g)$ *converges weakly to* $\lambda(\mathrm{d}s)V^\pi(s,\mathrm{d}g)$ *when* $\varepsilon\to 0$*.*

Proof. Let $f(g)$ be a continuous test function. We have:

$$\int_g V^\pi(s, \mathrm{d}g) f(g) = f(\varphi(s)) + \int_g \tilde{m}^\pi(s, g, g) f(g) \lambda(\mathrm{d}g) \tag{145}$$

Moreover, we know from Lemma 14 that

$$\tilde{m}^\pi(s, g, g) = \gamma \int_a \lambda(\mathrm{d}a) \pi(a|s, g) \left(\tilde{p}(g|s, a) + \int_{s'} \lambda(\mathrm{d}s') p(s'|s, a) \tilde{m}^\pi(s', g, g) \right) \tag{146}$$

Therefore:

$$\int_g V^\pi(s, \mathrm{d}g) f(g) = f(\varphi(s)) + \gamma \int_{g,a} \lambda(\mathrm{d}a, \mathrm{d}g) f(g) \pi(a|s, g) \left(\tilde{p}(g|s, a) + \int_{s'} \lambda(\mathrm{d}s') p(s'|s, a) \tilde{m}^\pi(s', g, g) \right) \tag{147}$$

On the other side, we have:

$$\int_g f(g) \mathbb{E}_{a \sim \pi(.|s,g), s' \sim P(\mathrm{d}s'|s,a)} \left[V^\pi(s', \mathrm{d}g) \right] = \tag{148}$$

$$= \int_{g,a,s'} \lambda(\mathrm{d}a, \mathrm{d}s') f(g) \pi(a|s, g) p(s'|s, a) \left(\delta_{\varphi(s')}(\mathrm{d}g) + \tilde{m}^\pi(s', g, g) \lambda(\mathrm{d}g) \right) \tag{149}$$

$$= \int_{a,s'} \lambda(\mathrm{d}a, \mathrm{d}s') \pi(a|s, g) p(s'|s, a) f(\varphi(s')) + \int_{a,s',g} \lambda(\mathrm{d}a, \mathrm{d}s', \mathrm{d}g) f(g) \pi(a|s, g) p(s'|s, a) \tilde{m}^\pi(s', g, g) \tag{150}$$

For the first part, we use the change of variable $g = \varphi(s')$, and we have:

$$\int_g f(g) \mathbb{E}_{a \sim \pi(.|s,g), s' \sim P(\mathrm{d}s'|s,a)} \left[V^\pi(s', \mathrm{d}g) \right] = \tag{151}$$

$$= \int_{a,g} \lambda(\mathrm{d}a, \mathrm{d}g) \pi(a|s, g) \tilde{p}(g|s, a) f(g) + \int_{a,s',g} \lambda(\mathrm{d}a, \mathrm{d}s', \mathrm{d}g) f(g) \pi(a|s, g) p(s'|s, a) \tilde{m}^\pi(s', g, g) \tag{152}$$

where $\tilde{p}(.|s, a)$ is the density of $\varphi_* P(.|s, a)$ (where φ_* is the push-forward operator) with respect to Lebesgue measure. Therefore, we have:

$$\int_g V(s, \mathrm{d}g) f(g) = \int_g f(g) \left(\delta_{\varphi(s)}(\mathrm{d}g) + \gamma \mathbb{E}_{s' \sim P^\pi(\mathrm{d}s'|s,g)} \left[V^\pi(s', \mathrm{d}g) \right] \right) \tag{153}$$

and we can conclude:

$$V^\pi(s, \mathrm{d}g) f(g) = \delta_{\varphi(s)}(\mathrm{d}g) + \gamma \mathbb{E}_{s' \sim P^\pi(\mathrm{d}s'|s,g)} \left[V^\pi(s', \mathrm{d}g) \right] \tag{154}$$

We now show that know that the measure on $K_{\mathcal{S}} \times \mathcal{G}$: $\lambda(\mathrm{d}s) \frac{1}{\lambda(\varepsilon)} V_\varepsilon^\pi(s, g) \lambda(\mathrm{d}g)$ converges weakly to $\lambda(\mathrm{d}s) V^\pi(s, \mathrm{d}g)$ when $\varepsilon \to 0$. We know that:

$$V_\varepsilon^\pi(s, g) = \mathbb{E} \left[\sum_{k \geqslant 0} \gamma^k R_\varepsilon(s_k, g) | s_0 = s \right] \tag{155}$$

$$= \frac{1}{1 - \gamma} \int_{s' \in \mathcal{S}} \nu^\pi(\mathrm{d}s'|s, g) R_\varepsilon(s', g) \tag{156}$$

$$= \int_{s' \in \mathcal{S}} \nu^\pi(\mathrm{d}s'|s, g) \mathbb{1}_{\|\varphi(s') - g\| \leqslant \varepsilon} \tag{157}$$

We use the change of variable $g' = \varphi(s')$, and we have (with φ_* the push-forward operator):

$$V_\varepsilon^\pi(s, g) = \int_{g' \in \mathcal{S}} (\varphi_* \nu^\pi)(\mathrm{d}g'|s, g) \mathbb{1}_{\|g' - g\| \leqslant \varepsilon} \tag{158}$$

$$= \int_{g' \in \mathcal{S}} M^\pi(s, g, \mathrm{d}g') \mathbb{1}_{\|g' - g\| \leqslant \varepsilon} \tag{159}$$

$$= M^\pi(s, g, B(g, \varepsilon)) \tag{160}$$

Let $F := \{g \in \mathcal{G}, \inf_{s\in K_\mathcal{S}} \|g - \varphi(s)\| < 1\}$. Therefore, for every $\varepsilon < 1$, the support of $\lambda(\mathrm{d}s)\frac{1}{\lambda(\varepsilon)}V^\pi_\varepsilon(s,g)\lambda(\mathrm{d}g)$ is compact and is a subset of F. Let $f(s,g)$ be a continuous bounded test function and $0 < \varepsilon < 1$. We have:

$$\int_{s\in K_\mathcal{S}, g\in\mathcal{G}} f(s,g)\frac{1}{\lambda(\varepsilon)}V^\pi_\varepsilon(s,g)\lambda(\mathrm{d}s,\mathrm{d}g) = \int_{s\in K_\mathcal{S}, g\in F} f(s,g)\frac{1}{\lambda(\varepsilon)}M^\pi(s,g,B(g,\varepsilon))\lambda(\mathrm{d}s,\mathrm{d}g) \tag{161}$$

We know that $M^\pi(s,g,\mathrm{d}g') = \delta_{\varphi(s)}(\mathrm{d}g') + \tilde{m}^\pi(s,g,g')\lambda(\mathrm{d}g')$. Therefore, $M^\pi(s,g,B(g,\varepsilon)) = \mathbb{1}_{\|g-\varphi(s)\|\leqslant\varepsilon} + \int_{g'}\frac{\mathbb{1}_{\|g-g'\|\leqslant\varepsilon}}{\lambda(\varepsilon)}\tilde{m}^\pi(s,g,g')$, and we have:

$$\int_{s\in K_\mathcal{S}, g\in\mathcal{G}} f(s,g)\frac{1}{\lambda(\varepsilon)}V^\pi_\varepsilon(s,g)\lambda(\mathrm{d}s,\mathrm{d}g) = \tag{162}$$

$$= \int_{s\in K_\mathcal{S}, g\in F}\frac{\mathbb{1}_{\|g-\varphi(s)\|\leqslant\varepsilon}}{\lambda(\varepsilon)}f(s,g)\lambda(\mathrm{d}s,\mathrm{d}g) + \int_{s\in K_\mathcal{S}, g\in F, g'\in\mathcal{G}} f(s,g)\frac{\mathbb{1}_{\|g-g'\|\leqslant\varepsilon}}{\lambda(\varepsilon)}\tilde{m}^\pi(s,g,g')\lambda(\mathrm{d}s,\mathrm{d}g,\mathrm{d}g') \tag{163}$$

$$= \int_u\left(\int_{s\in K_\mathcal{S}}\lambda(\mathrm{d}s)f(s,\varphi(s)+u) + \int_{s\in K_\mathcal{S}, g\in\mathcal{G}} f(s,g)\tilde{m}^\pi(s,g,g+u)\lambda(\mathrm{d}s,\mathrm{d}g)\right)U_\varepsilon(\mathrm{d}u) \tag{164}$$

where $U_\varepsilon(\mathrm{d}u)$ is the uniform measure on $B(0,\varepsilon)$ the ball of size ε around 0: $U_\varepsilon(\mathrm{d}u) := \frac{\mathbb{1}_{\|u\|\leqslant\varepsilon}}{\lambda(\varepsilon)}\lambda(\mathrm{d}u)$. We can switch the order of integration because $f, \tilde{m}^\pi$ are continuous, bounded, and the integral is computed on compact sets. The function $u \to \int_{s\in K_\mathcal{S}}\lambda(\mathrm{d}s)f(s,\varphi(s)+u) + \int_{s\in K_\mathcal{S}, g\in F} f(s,g)\tilde{m}(s,g,g+u)\lambda(\mathrm{d}s,\mathrm{d}g)$ is bounded and continuous. Since $U_\varepsilon(\mathrm{d}u)$ converges weakly to $\delta_0(\mathrm{d}u)$, we have:

$$\lim_{\varepsilon\to 0}\int_{s\in K_\mathcal{S}, g\in\mathcal{G}} f(s,g)\frac{1}{\lambda(\varepsilon)}V^\pi_\varepsilon(s,g)\lambda(\mathrm{d}s,\mathrm{d}g) = \tag{165}$$

$$= \lim_{\varepsilon\to 0}\int_u\left(\int_{s\in K_\mathcal{S}}\lambda(\mathrm{d}s)f(s,\varphi(s)+u) + \int_{s\in K_\mathcal{S}, g\in F} f(s,g)\tilde{m}^\pi(s,g,g+u)\lambda(\mathrm{d}s,\mathrm{d}g)\right)U_\varepsilon(\mathrm{d}u) \tag{166}$$

$$= \int_u\left(\int_{s\in K_\mathcal{S}}\lambda(\mathrm{d}s)f(s,\varphi(s)+u) + \int_{s\in K_\mathcal{S}, g\in\mathcal{G}} f(s,g)\tilde{m}^\pi(s,g,g+u)\lambda(\mathrm{d}s\,\mathrm{d}g)\right)\delta_0(\mathrm{d}u) \tag{167}$$

$$= \int_{s\in K_\mathcal{S}}\lambda(\mathrm{d}s)f(s,\varphi(s)) + \int_{s\in K_\mathcal{S}, g} f(s,g)\tilde{m}^\pi(s,g,g)\lambda(\mathrm{d}s,\mathrm{d}g) \tag{168}$$

$$= \int_{s\in K_\mathcal{S}, g} f(s,g)V^\pi(s,\mathrm{d}g)\lambda(\mathrm{d}s) \tag{169}$$

This concludes the proof. □

Obstacles for learning V^π directly. We briefly show why learning V^π directly without bias poses technical issues, stemming from the necessity to work on-policy for V and the resulting correlation between visited states and goals along trajectories in the training set. As a result, the "obvious" analogue of δ-DQN for V introduces uncontrolled bias and implicit preferences among all possible states s that achieve the same goal g. This problem disappears only if the correspondence between s and g is one-to-one (e.g., $\varphi = \mathrm{Id}$). This is why we learn the more complicated object M^π instead of V^π in Section 3.3.

Assume similarly to Theorem 13 that we can sample state-goal pairs from a distribution $\rho_{\mathrm{SG}}(\mathrm{d}s,\mathrm{d}g)$ over $\mathcal{S}\times\mathcal{G}$, and define the norm $\|\cdot\|_{\rho_{\mathrm{SG}}}$ as

$$\|V\|_{\rho_{\mathrm{SG}}} = \int_{s,g}\rho_{\mathrm{SG}}(\mathrm{d}s,\mathrm{d}g)\left(\frac{V(s,\mathrm{d}g)}{\rho_\mathcal{G}(\mathrm{d}g)}\right)^2 \tag{170}$$

where $\frac{V(s,\mathrm{d}g)}{\rho_{\mathcal{G}}(\mathrm{d}g)}$ is the density of $V(s,\mathrm{d}g)$ with respect to $\rho_{\mathcal{G}}(\mathrm{d}g)$ (if it does not exist, the norm is infinite). We assume we have a model $V_\theta(s,\mathrm{d}g) = v_\theta(s,g)\rho_{\mathcal{G}}(\mathrm{d}g)$, a target $V_{\mathrm{tar}}(s,\mathrm{d}g) = v_{\mathrm{tar}}(s,g)\rho_{\mathcal{G}}(\mathrm{d}g)$, and want to estimate:

$$\frac{1}{2}\partial_\theta \|V_\theta - T^\pi V_{\mathrm{tar}}\|^2_{\rho_{\mathrm{SG}}} \tag{171}$$

where $T^\pi V(s,\mathrm{d}g) = \delta_{\varphi(s)}(\mathrm{d}g) + \gamma \mathbb{E}_{s'\sim P^\pi(.|s,g)} V(s',\mathrm{d}g)$. Then, informally, we have:

$$\frac{1}{2}\partial_\theta \|V_\theta - T^\pi V_{\mathrm{tar}}\|^2_{\rho_{\mathrm{SG}}} = \frac{1}{2}\partial_\theta \int_{s,g} \rho_{\mathrm{SG}}(\mathrm{d}s,\mathrm{d}g)\left(\frac{V_\theta(s,\mathrm{d}g)}{\rho_{\mathcal{G}}(\mathrm{d}g)} - \frac{TV_{\mathrm{tar}}(s,\mathrm{d}g)}{\rho_{\mathcal{G}}(\mathrm{d}g)}\right)^2 \tag{172}$$

$$= \frac{1}{2}\partial_\theta \int_{s,g} \rho_{\mathrm{SG}}(\mathrm{d}s,\mathrm{d}g)\left(v_\theta(s,g) - \frac{TV_{\mathrm{tar}}(s,\mathrm{d}g)}{\rho_{\mathcal{G}}(\mathrm{d}g)}\right)^2 \tag{173}$$

$$= \int_{s,g} \rho_{\mathrm{SG}}(\mathrm{d}s,\mathrm{d}g)\partial_\theta v_\theta(s,g)\left(v_\theta(s,g) - \frac{TV_{\mathrm{tar}}(s,\mathrm{d}g)}{\rho_{\mathcal{G}}(\mathrm{d}g)}\right) \tag{174}$$

$$= \int_{s,g} \rho_{\mathrm{SG}}(\mathrm{d}s,\mathrm{d}g)\partial_\theta v_\theta(s,g)\left(v_\theta(s,g) - \gamma\mathbb{E}_{s'\sim P^\pi(.|s,g)}\left[v_{\mathrm{tar}}(s',g)\right]\right) + \tag{175}$$

$$+ \int_{s,g} \rho_{\mathrm{SG}}(\mathrm{d}s,\mathrm{d}g)\partial_\theta v_\theta(s,g)\frac{\delta_{\varphi(s)}(\mathrm{d}g)}{\rho_{\mathcal{G}}(\mathrm{d}g)} \tag{176}$$

If we assume that $\rho_{\mathrm{SG}}(\mathrm{d}s,\mathrm{d}g)$ has a density $\alpha(s,g)$ with respect to $\rho_{\mathrm{SG}}(\mathrm{d}s)\otimes\rho_{\mathcal{G}}(\mathrm{d}g)$, namely, $\rho_{\mathrm{SG}}(\mathrm{d}s,\mathrm{d}g) = \alpha(s,g)\rho_{\mathrm{SG}}(\mathrm{d}s)\rho_{\mathcal{G}}(\mathrm{d}g)$, then the second part, corresponding to the Dirac reward, is equal to:

$$\int_{s,g} \rho_{\mathrm{SG}}(\mathrm{d}s,\mathrm{d}g)\partial_\theta v_\theta(s,g)\frac{\delta_{\varphi(s)}(\mathrm{d}g)}{\rho_{\mathcal{G}}(\mathrm{d}g)} = \int_{s,g} \rho_{\mathrm{SG}}(\mathrm{d}s)\alpha(s,g)\partial_\theta v_\theta(s,g)\delta_{\varphi(s)}(\mathrm{d}g) \tag{177}$$

$$= \int_s \rho_{\mathrm{SG}}(\mathrm{d}s)\alpha(s,\varphi(s))\partial_\theta v_\theta(s,\varphi(s)) \tag{178}$$

If $\alpha(s,g)$ is always equal to 1, the integral $\int_s \rho_{\mathrm{SG}}(\mathrm{d}s)\partial_\theta v_\theta(s,\varphi(s))$ can be estimated without bias by sampling $s\sim\rho_{\mathrm{SG}}(\mathrm{d}s)$ and estimating $v_\theta(s,\varphi(s))$.

However, the case $\alpha(s,g) = 1$ for every s,g corresponds to s and g independent in ρ_{SG}. This is difficult to realize in practice. Learning V requires actions to be selected on-policy (term $\mathbb{E}_{s'\sim P^\pi(.|s,g)}$ above). If we set a goal g and an initial state s_0, and generate an exploration trajectory by following the policy $\pi(.|.,g)$ for that goal, obviously the states s visited by the trajectory are going to be correlated to g, by an unknown factor α. Independence could be ensured by re-sampling a new target goal at each step, independently from the current state, and selecting the next action from the policy for this goal. But such an exploration strategy would be essentially random and would not be efficient.

Assume we just ignore this problem and sample exploration trajectories $(g, s_0, s_1, ...)$ as with other methods, namely, with $g\sim\rho_{\mathcal{G}}$, $s_0\sim\rho_0(\mathrm{d}s_0|g)$ and $s_{t+1}\sim P^\pi(.|s_t,g)$, and define the estimate

$$\widehat{\delta\theta}_V(s,s',g) = \partial_\theta v_\theta(s,\varphi(s)) + \partial_\theta v_\theta(s,g)\left(\gamma v_{\mathrm{tar}}(s',g)v_\theta(s,g)\right) \tag{179}$$

similarly to updates of δ-DQN or δ-TD. In that case, we have:

$$\mathbb{E}_{s,g\sim\rho_{\mathrm{SG}},s'\sim P^\pi(.|s,g)}\left[\widehat{\delta\theta}_V(s,s',g)\right] = \|V_\theta - T^\pi_\alpha V_{\mathrm{tar}}\|_{\rho_{\mathrm{SG}}} \tag{180}$$

where:

$$T^\pi_\alpha V = \alpha(s,g)\delta_{\varphi(s)} + \mathbb{E}_{s'\sim P^\pi(.|s,g)}\left[V(s',\mathrm{d}g)\right]. \tag{181}$$

This is an unbiased estimate of the TD error with the *rescaled reward* $\alpha(s,g)\delta_{\varphi(s)}(\mathrm{d}g)$ instead of $\delta_{\varphi(s)}(\mathrm{d}g)$.

If $\mathcal{S} = \mathcal{G}$ and $\varphi = \mathrm{Id}$, such a reward rescaling is not an issue. Indeed, in that case, $\alpha(s,g)\delta_s(\mathrm{d}g) = \alpha(g,g)\delta_s(\mathrm{d}g)$ as the Dirac measure is nonzero only for $s = g$. This means that for every goal g, the value function for that goal is rescaled by a constant $\alpha(g,g)$, and we learn $\alpha(g,g)V(s,\mathrm{d}g)$ instead of

$V(s, \mathrm{d}g)$. This does not change the ranking of state values for each goal g, nor the direction of policy improvement for each goal (but it changes the relative importance of learning different goals g).

On the contrary, if $\mathcal{S} \neq \mathcal{G}$, for a fixed goal g, this implicit reward rescaling can favor some states s over others among the set of states s achieving this goal ($\varphi(s) = g$). For instance, assume the the agent starts at s_0 and wants to reach g, and that there are two states s_1, s_2 such that $\varphi(s_1) = \varphi(s_2) = g$. Even if s_1 is easier to reach than s_2 from s_0, the policy π might *prefer* to reach s_2 because its implicitly rescaled reward is higher. Therefore, the algorithm could converge to non-optimal policies and is not unbiased. It would still learn to reach g, but not necessarily in an optimal way.

E.3 Equivalence Between $\varepsilon \to 0$ and the Dirac Setting

DEFINITION 16. *We say that π_2 is* better *than π_1 with infinitely sparse rewards if the two measures $\lambda(\mathrm{d}s)V^{\pi_1}(s, \mathrm{d}g)$ and $\lambda(\mathrm{d}s)V^{\pi_2}(s, \mathrm{d}g)$ on $K_{\mathcal{S}} \times \mathcal{G}$ satisfy: $\lambda(\mathrm{d}s)V^{\pi_1}(s, \mathrm{d}g) \leqslant \lambda(\mathrm{d}s)V^{\pi_2}(s, .)$.*

We say that π_2 is asymptotically better *than π_1 when $\varepsilon \to 0$ if for all s, g,*

$$\liminf_{\varepsilon \to 0} \frac{V_\varepsilon^{\pi_2}(s, g)}{V_\varepsilon^{\pi_1}(s, g)} \geqslant 1.$$

THEOREM 17. *We assume Assumption 1 and take $\pi_1, \pi_2 \in \Pi$.*

Then, π_2 is better than π_1 with infinitely sparse rewards if and only if π_2 is asymptotically better *than π_1 when $\varepsilon \to 0$. In particular, a policy π^* is an optimal policy with infinitely sparse rewards if and only if it is an optimal policy when $\varepsilon \to 0$.*

Proof. We know that $V^\pi(s, \mathrm{d}g) = \delta_{\varphi(s)}(\mathrm{d}g) + \tilde{m}(s, g, g)^\pi \lambda(\mathrm{d}g)$. Moreover:

$$V_\varepsilon^\pi(s_0, g) = M(s_0, g, B(g, \varepsilon)) \tag{182}$$
$$= \mathbb{1}_{\varphi(s_0)=g} + \lambda(\varepsilon)\tilde{m}^\pi(s_0, g, g) + o(\lambda(\varepsilon)) \tag{183}$$

Therefore, for any policies $\pi_1, \pi_2 \in \Pi$:

$$\frac{V_\varepsilon^{\pi_2}(s, g)}{V_\varepsilon^{\pi_1}(s, g)} = \frac{\mathbb{1}_{\varphi(s)=g} + \tilde{m}^{\pi_2}(s, g, g)\lambda(\varepsilon) + o(\lambda(\varepsilon))}{\mathbb{1}_{\varphi(s)=g} + \tilde{m}^{\pi_1}(s, g, g)\lambda(\varepsilon) + o(\lambda(\varepsilon))} \tag{184}$$
$$= \mathbb{1}_{\varphi(s)=g} + \mathbb{1}_{\varphi(s)\neq g}\frac{\tilde{m}^{\pi_2}(s, g, g)}{\tilde{m}^{\pi_1}(s, g, g)} + o_{\varepsilon\to 0}(1) \tag{185}$$

Therefore, by definition, π_2 is asymptotically better than π_1 when $\varepsilon \to 0$ if and only if, for all $(s, g) \in \mathcal{S} \times \mathcal{G}$:

$$\mathbb{1}_{\varphi(s)=g} + \mathbb{1}_{\varphi(s)\neq g}\frac{\tilde{m}^{\pi_2}(s, g, g)}{\tilde{m}^{\pi_1}(s, g, g)} \geqslant 1 \tag{186}$$

If $\varphi(s) \neq g$, this inequality is equivalent to $\tilde{m}^{\pi_2}(s, g, g) \geqslant m^{\pi_1}(s, g, g)$. Because $\tilde{m}^{\pi_1}$ and $\tilde{m}^{\pi_2}$ are continuous, $\tilde{m}^{\pi_2}(s, g, g) \geqslant m^{\pi_1}(s, g, g)$ for all $\varphi(s) \neq g$ is equivalent to $\tilde{m}^{\pi_2}(s, g, g) \geqslant m^{\pi_1}(s, g, g)$ for every (s, g). Therefore, π_2 is asymptotically better than π_1 when $\varepsilon \to 0$ if and only if, for all (s, g), $m^{\pi_2}(s, g, g) \geqslant m^{\pi_1}(s, g, g)$.

On the other side π_2 is better than π_1 with infinitely sparse rewards if and only if:

$$\lambda(\mathrm{d}s)V^{\pi_1}(s, \mathrm{d}g) \preceq \lambda(\mathrm{d}s)V^{\pi_2}(s, \mathrm{d}g) \tag{187}$$
$$\Leftrightarrow \quad \lambda(\mathrm{d}s)\delta_{\varphi(s)}(\mathrm{d}g) + m_{\pi_1}(s, g, g)\lambda(\mathrm{d}s, \mathrm{d}g) \preceq \lambda(\mathrm{d}s)\delta_{\varphi(s)}(\mathrm{d}g) + m_{\pi_2}(s, g, g)\lambda(\mathrm{d}s, \mathrm{d}g) \tag{188}$$
$$\Leftrightarrow \quad m_{\pi_1}(s, g, g)\lambda(\mathrm{d}s, \mathrm{d}g) \leqslant m_{\pi_2}(s, g, g)\lambda(\mathrm{d}s, \mathrm{d}g) \tag{189}$$

Therefore, for λ-almost every (s, g), $m_{\pi_1}(s, g, g) \leqslant m_{\pi_2}(s, g, g)$. Therefore: π_2 is better than π_1 with infinitely sparse rewards if and only if $\tilde{m}^{\pi_2}(s, g, g) \geqslant m^{\pi_1}(s, g, g)$ for λ-almost every s, g. This concludes the proof.

□

In the following statement, we introduce 3 definitions of expected return: the return $J(\pi)$ with infinitely sparse reward, the return $J_\varepsilon(\pi)$ with sparse reward R_ε, and the estimated return $J_n(\pi)$ with the value measure approximator v_n. Then, we show that these three definitions are consistent.

THEOREM 18. *We define $J(\pi)$, the expected return with infinitely sparse rewards for the goal density $p_{\mathcal{G}}$, as:*

$$J(\pi) := \int_{s_0,g} \lambda(\mathrm{d}s_0) p_{\mathcal{G}}(g) p_0(s_0|g) V^\pi(s_0, \mathrm{d}g). \tag{190}$$

We consider the expected return for the reward R_ε and the goal distribution $\rho(\mathrm{d}g)$.

$$J_\varepsilon(\pi) = \mathbb{E}_{g\sim\rho(\mathrm{d}g), s_0, a_0, \ldots}\left[\sum_{t\geqslant 0} \gamma^t R_\varepsilon(s_t, g)\right] = \int_{g,s_0} p_{\mathcal{G}}(g)\lambda(\mathrm{d}s_0, \mathrm{d}g) p_0(s_0|s) V_\varepsilon(s_0, g) \tag{191}$$

Let $(\hat{v}_n(s,g))_{n\geqslant 0}$ be any sequence of densities on $\mathcal{S}\times\mathcal{G}$ such that the measure on $\mathcal{S}\times\mathcal{G}$: $\lambda(\mathrm{d}s)\hat{v}_n(s,g)\rho_{\mathcal{G}}(\mathrm{d}g)$ converges weakly to $\lambda(\mathrm{d}s)V^\pi(s,\mathrm{d}g)$. We define $\tilde{\rho}(\mathrm{d}g) := \frac{1}{c}p_{\mathcal{G}}^2(g)\lambda(\mathrm{d}g)$ with $c := \int_g p_{\mathcal{G}}^2(g)\lambda(\mathrm{d}g)$, and $J_n(\pi)$ the estimator of the average return for the goal distribution $\tilde{\rho}$ with estimator $\hat{v}_n$:

$$J_n(\pi) := \mathbb{E}_{g\sim\tilde{\rho}(\mathrm{d}g), s_0\sim p(s_0|g)}\left[\hat{v}_n(s_0, g)\right] \tag{192}$$

Then the two estimators J_n and J_ε converge to J:

$$\frac{1}{\lambda(\varepsilon)} J_\varepsilon(\pi) \to_{\varepsilon\to 0} J(\pi) \tag{193}$$

$$cJ_n(\pi) \to_{n\to\infty} J(\pi) \tag{194}$$

Proof. We have:

$$J_\varepsilon(\pi) = \int_{s_0,g} V_\varepsilon(s_0, g) p_{\mathcal{G}}(g) p_0(s_0|g)\lambda(\mathrm{d}s_0, \mathrm{d}g) \tag{195}$$

$$J_n(\pi) = \int_{s_0,g} \hat{v}_n(s_0, g)\frac{1}{c} p_{\mathcal{G}}^2(g) p_0(s_0|g)\lambda(\mathrm{d}s_0, \mathrm{d}g) \tag{196}$$

and whe know from Theorem 15 that $\frac{V_\varepsilon(s,g)}{\lambda(\varepsilon)}\lambda(\mathrm{d}s, \mathrm{d}g)$ and $\hat{v}_n(s,g)p_{\mathcal{G}}(\mathrm{d}g)\lambda(\mathrm{d}s, \mathrm{d}g)$ converge weakly to $\lambda(\mathrm{d}s)V^\pi(s, \mathrm{d}g)$ on $K_{\mathcal{S}}\times\mathcal{G}$ when $\varepsilon\to 0$ and $n\to\infty$. Therefore, because p_0 and $p_{\mathcal{G}}$ are continuous bounded functions,

$$\lim_{\varepsilon\to 0}\frac{1}{\lambda(\varepsilon)} J_\varepsilon(\pi) = \int_{s_0,g} V^\pi(s_0, \mathrm{d}g) p_{\mathcal{G}}(g) p_0(s_0|g)\lambda(\mathrm{d}s_0) = J(\pi) \tag{197}$$

Similarly:

$$\lim_{n\to\infty} J_n(\pi) = \int_{s_0,g} V^\pi(s, \mathrm{d}g)\frac{1}{c} p_{\mathcal{G}}(g) p_0(s_0|g)\lambda(\mathrm{d}s_0) \tag{198}$$

$$= \frac{1}{c} J(\pi) \tag{199}$$

□

PROPOSITION 19. *We assume Assumption 1. Moreover, we assume that $p_{\mathcal{G}}(g) > 0$ for every $g\in\varphi(K_{\mathcal{S}})$, and $p_0(s_0|g) > 0$ for every $(s_0, g)\in K_{\mathcal{S}}\times\mathcal{G}$.*

We consider the partial order $\prec$ defined as: $\pi_1 \prec \pi_2$ if π_2 is strictly better than π_1 with infinitely sparse rewards: $\lambda(\mathrm{d}s)V^{\pi_1}(s,\mathrm{d}g) \prec \lambda(\mathrm{d}s)V^{\pi_2}(s,\mathrm{d}g)$ on $K_{\mathcal{S}}\times\mathcal{G}$.

Then $\pi\mapsto J(\pi)$ is strictly increasing for $\prec$.

Proof. The function $J(\pi)$ is clearly non-decreasing, and we have to check that we cannot have $\pi_1\prec\pi_2$ with $J(\pi_1) = J(\pi_2)$. Let $\pi_1, \pi_2\in\Pi$ such that $\pi_1\prec\pi_2$. Therefore, there is $U\subset K_{\mathcal{S}}\times\mathcal{G}$ such that $(\lambda\otimes V^{\pi_2}(.,.))(U) > (\lambda\otimes V^{\pi_1}(.,.))(U)$. Moreover, because $\operatorname{supp}(\lambda(\mathrm{d}s)V^\pi(s,\mathrm{d}g))\subset K_{\mathcal{S}}\times\varphi(K_{\mathcal{S}})$, therefore we can suppose $U\subset K_{\mathcal{S}}\times\varphi(K_{\mathcal{S}})$, and we have:

$$\int_{(s,g)\in U}\lambda(\mathrm{d}s, \mathrm{d}g)(\tilde{m}^{\pi_2}(s,g,g) - \tilde{m}^{\pi_1}(s,g,g)) > 0 \tag{200}$$

We already know that $\tilde{m}^{\pi_2}(s,g,g) - \tilde{m}^{\pi_1}(s,g,g) \geqslant 0$ for almost every s, g (see proof of Theorem 17). Therefore, there is $\varepsilon' > 0$ and $V \subset U$ with $\mathrm{d}\lambda(V) > 0$ such that for every $s, g \in V$, $\tilde{m}^{\pi_2}(s,g,g) - \tilde{m}^{\pi_1}(s,g,g) > \varepsilon'$.

We have:

$$J(\pi_2) - J(\pi_1) = \int_{s \in K_{\mathcal{S}}, g \in \mathcal{G}} p_0(s|g) p_{\mathcal{G}}(g) (V^{\pi_2}(s, \mathrm{d}g) - V^{\pi_1}(s, \mathrm{d}g) \tag{201}$$

$$\geqslant \int_{(s,g) \in V} p_0(s|g) p_{\mathcal{G}}(g) \left(\tilde{m}^{\pi_2}(s,g,g) - \tilde{m}^{\pi_1}(s,g,g)\right) \tag{202}$$

$$\geqslant \varepsilon' \int_{(s,g) \in V} p_0(s|g) p_{\mathcal{G}}(g) \tag{203}$$

$$> 0 \tag{204}$$

because $p_{\mathcal{G}}(g) > 0$ for λ-almost every g in $\varphi(K_{\mathcal{S}})$, and $p_0(s|g) > 0$ for λ-almost every s, g in $K_{\mathcal{S}} \times \mathcal{G}$. This concludes the proof.

□

E.4 Policy Gradient

THEOREM 20 (FORMAL STATEMENT OF INFORMAL THEOREM 7). *Let $\pi_\theta(a|s,g)$ be a parametrized goal-dependent policy, defined for every $\theta \in \Theta$. We assume that for every $\theta \in \Theta, s \in \mathcal{S}, g \in \mathcal{G}, a \in \mathcal{A}$, $\pi_\theta(a|s,g) > 0$. Moreover, we assume $\pi_\theta(a|s,g)$ is a continuous function of a, s, g, θ, and continuously differentiable with respect to θ.*

We define $\tilde{\rho}(\mathrm{d}g) := \frac{1}{c} p_{\mathcal{G}}^2(g) \lambda(\mathrm{d}g)$ with $c := \int_g p_{\mathcal{G}}^2(g) \lambda(\mathrm{d}g)$. We assume access to samples $g \sim \tilde{\rho}(\mathrm{d}g)$, $s_0 \sim \rho_0(\mathrm{d}s|g) = p_0(s_0|g)\lambda(\mathrm{d}s_0)$, $s \sim \nu^{\pi_\theta}(\ s|s_0, g)$, $a \sim \pi(a|s,g)$ and $s' \sim P(\mathrm{d}s'|s,a)$. Let $(\hat{v}_n(s,g))_{n \geqslant 0}$ be a sequence of densities, such that $\lambda(\mathrm{d}s)\hat{v}_n(s,g)\rho(\mathrm{d}g)$ converges weakly to $\lambda(\mathrm{d}s)V^{\pi_\theta}(s, \mathrm{d}g)$. We define the stochastic actor critic $\widehat{\delta\theta}^{(n)}_{\delta\text{-AC}}$ for estimate n as:

$$\widehat{\delta\theta}^{(n)}_{\delta\text{-AC}}(s,a,s',g) := \partial_\theta \log \pi_\theta(a|s,g) \left(\gamma \hat{v}_n(s',g) - \hat{v}_n(s,g)\right) \tag{205}$$

Then, we have:

$$\lim_{n \to \infty} \mathbb{E}_{g \sim \tilde{\rho}, s \sim \nu^\pi(.|g), a \sim \pi_\theta(.|s,g), s' \sim P(.|s,a)} \left[\widehat{\delta\theta}^{(n)}_{\delta\text{-AC}}(s,a,s',g)\right] = \frac{1-\gamma}{c} \partial_\theta J(\pi_\theta) \tag{206}$$

Moreover, we have:

$$\lim_{\varepsilon \to 0} \frac{1}{\lambda(\varepsilon)} \partial_\theta J_\varepsilon(\pi_\theta) = \partial_\theta J(\pi_\theta) \tag{207}$$

Proof. We first compute $\partial_\theta J(\pi_\theta)$. We have:

$$J(\pi_\theta) = \int_{s_0, g} V^{\pi_\theta}(s_0, \mathrm{d}g) p_{\mathcal{G}}(g) p_0(s_0|g) \lambda(\mathrm{d}s_0) \tag{208}$$

We know that $V^\pi(s, \mathrm{d}g) = \delta_{\varphi(s)}(\mathrm{d}g) + \tilde{m}^\pi(s,g,g)\lambda(\mathrm{d}g)$. We define for simplicity $v^\pi(s,g) = \tilde{m}^\pi(s,g,g)$. Moreover, we know, by taking $g' = g$ in Equation (122) in Lemma 14 that for every (s,g), we have:

$$v^{\pi_\theta}(s,g) = \gamma \int_a \lambda(\mathrm{d}a) \pi(a|s,g) \left(\tilde{p}(g|s,a) + \int_{s'} \lambda(\mathrm{d}s') p(s'|s,a) v^{\pi_\theta}(s',g)\right) \tag{209}$$

We define $F(s,g,\theta) = \gamma \int_a \pi_\theta(a|s,g) \tilde{p}(g|s,a)$. The function F_θ is continuous in s and g and continuously differentiable in θ, because $\tilde{p}$ is and π_θ are continuous, π_θ is continuously differentiable, and $\mathcal{A}$ is compact. From the proof of Equation (122) in Lemma 14, we know that $F(s,g,\theta)$ is the density of $\gamma \int_{a,s'} \pi(a|s,g) p(s'|s,a) \delta_{\varphi(s')}(\mathrm{d}g)$ with respect to the Lebesgue measure $\lambda(\mathrm{d}g)$. This remark will be used later in the computation. We now have:

$$v^{\pi_\theta}(s,g) = F(s,g,\theta) + \gamma \int_{a,s'} \pi_\theta(a|s,g) p(s'|s,a) v^{\pi_\theta}(s',g) \lambda(\mathrm{d}a, \mathrm{d}s') \tag{210}$$

Therefore:

$$v^{\pi_\theta}(s,g) = F(s,g,\theta) + \sum_{k\geqslant 1}\gamma^k \int_{a_0,s_1,\dots} \lambda(\mathrm{d}a_0,\mathrm{d}s_1,...,\mathrm{d}s_k)\left(\prod_{i=0}^{k-1}\pi_\theta(a_i|s_i,g)p(s_{i+1}|s_i,a_i)\right)F(s_k,g,\theta) \tag{211}$$

because it is a fixed point of v^π equation, and is the only fixed point which is continuous and bounded, because the space is compact, and π_θ, p are continuous an bounded.

Equation (211) can also be written:

$$v^{\pi_\theta}(s,g) = \frac{1}{1-\gamma}\int_{s'} \nu^{\pi_\theta}(\mathrm{d}s'|s,g)F(s',g,\theta) \tag{212}$$

Because $F(s',g,\theta)$ is continuously differentiable in θ and the support of ν^π is compact, $v^{\pi_\theta}(s,g)$ is differentiable. We will now now derive a fixed point equation on $\partial_\theta v^{\pi_\theta}$. We differentiate equation (210) and we get:

$$\partial_\theta v^{\pi_\theta}(s,g) = \partial_\theta F(s,g,\theta) + \gamma\int_{a,s}\lambda(\mathrm{d}a,\mathrm{d}s)\partial_\theta\pi_\theta(a|s,g)p(s'|s,a)v^{\pi_\theta}(s',g) + \tag{213}$$

$$+\gamma\int_{a,s}\lambda(\mathrm{d}a,\mathrm{d}s)\pi_\theta(a|s,g)p(s'|s,a)\partial_\theta v^{\pi_\theta}(s',g) \tag{214}$$

We define $G(s,g,\theta) := \partial_\theta F(s,g,\theta) + \gamma\int_{a,s'}\lambda(\mathrm{d}a,\mathrm{d}s')\partial_\theta\pi_\theta(a|s,g)p(s'|s,a)v^{\pi_\theta}(s',g)$. We have:

$$\partial_\theta v^{\pi_\theta}(s,g) = G(s,g,\theta) + \gamma\int_{a,s'}\lambda(\mathrm{d}a,\mathrm{d}s')\pi_\theta(a|s,g)p(s'|s,a)\partial_\theta v^{\pi_\theta}(s',g) \tag{215}$$

Similarly to the derivation of v^π from its fixed point equation (from (210) to (211)):

$$\partial_\theta v^{\pi_\theta}(s,g) = G(s,g,\theta) + \sum_{k\geqslant 1}\gamma^k\int_{a_0,s_1,\dots}\lambda(\mathrm{d}a_0,\mathrm{d}s_1,...,\mathrm{d}s_k)\left(\prod_{i=0}^{k-1}\pi_\theta(a_i|s_i,g)p(s_{i+1}|s_i,a_i)\right)G(s_k,g,\theta) \tag{216}$$

$$= \frac{1}{1-\gamma}\int_{s'}\nu^{\pi_\theta}(\mathrm{d}s'|s,g)G(s',g,\theta) \tag{217}$$

We now compute $\partial_\theta J(\pi_\theta)$. We have:

$$\partial_\theta J(\theta) = \partial_\theta\left(\int_{s_0,g}\lambda(\mathrm{d}s_0)p_{\mathcal{G}}(g)p_0(s_0|g)V^{\pi_\theta}(s_0,\mathrm{d}g)\right) \tag{218}$$

$$= \partial_\theta\left(\int_{s_0,g}\lambda(\mathrm{d}s_0)p_{\mathcal{G}}(g)p_0(s_0|g)\left(\delta_{\varphi(s_0)}(\mathrm{d}g) + v^{\pi_\theta}(s_0,g)\lambda(\mathrm{d}g)\right)\right) \tag{219}$$

$$= \partial_\theta\left(\int_{s_0,g}\lambda(\mathrm{d}s_0,\mathrm{d}g)p_{\mathcal{G}}(g)p_0(s_0|g)v^{\pi_\theta}(s_0,g)\right) \tag{220}$$

$$= \int_{s_0,g}\lambda(\mathrm{d}s_0,\mathrm{d}g)p_{\mathcal{G}}(g)p_0(s_0|g)\partial_\theta v^{\pi_\theta}(s_0,g) \tag{221}$$

$$= \frac{1}{1-\gamma}\int_{s_0,s,g}\lambda(\mathrm{d}s_0,\mathrm{d}g)p_{\mathcal{G}}(g)p_0(s_0|g)\nu^{\pi_\theta}(\mathrm{d}s|s_0,g)G(s,g,\theta) \tag{222}$$

We now show that:

$$G(s,g,\theta)\lambda(\mathrm{d}g) = \gamma\int_{s',a}V(s',\mathrm{d}g)\partial_\theta\pi_\theta(a|s,g)p(s'|s,a) \tag{223}$$

While this result might seem to come out of nowhere, remember that $F(s,g,\theta)$ was derived above as the measure density of $\gamma\int_{s',a}\pi(a|s,g)p(s'|s,g)\delta_{\varphi(s')}(\mathrm{d}g)$ with respect to Lebesgue measure. With

the following informal computation, we have:

$$G(s,g,\theta)\lambda(\mathrm{d}g) = \lambda(\mathrm{d}g)\partial_\theta \frac{1}{\lambda(\mathrm{d}g)} \int_{s',a} \lambda(\mathrm{d}s',\mathrm{d}a)\gamma\pi_\theta(a|s,g)p(s'|s,a)\delta_{\varphi(s')}(\mathrm{d}g) + \gamma \int_{s',a} \lambda(\mathrm{d}s',\mathrm{d}a)v^{\pi_\theta}(s',g)\partial_\theta\pi_\theta(a|s,g) \tag{224}$$

$$= \int_{s',a} \lambda(\mathrm{d}s',\mathrm{d}a)\gamma\partial_\theta\pi_\theta(a|s,g)p(s'|s,a)\left(\delta_{\varphi(s)}(\mathrm{d}g) + v^{\pi_\theta}(s',g)\lambda(\mathrm{d}g)\right) \tag{225}$$

$$= \int_{s',a} \lambda(\mathrm{d}s',\mathrm{d}a)\gamma V^{\pi_\theta}(s',\mathrm{d}g)\partial_\theta\pi_\theta(a|s,g)p(s'|s,a) \tag{226}$$

This derivation is informal because we differentiated through a density: we use $\lambda(\mathrm{d}g)\partial_\theta \frac{1}{\lambda(\mathrm{d}g)} = \partial_\theta$. We now derive the result rigorously. Let $f(g)$ be a continuous test function. We have:

$$\int_g f(g)G(s,g,\theta)\lambda(\mathrm{d}g) = \tag{227}$$

$$= \int_g \lambda(\mathrm{d}g)f(g)\left(\gamma\int_a \lambda(\mathrm{d}a)\partial_\theta\pi_\theta(a|s,g)\tilde{p}(g|s,a) + \gamma\int_{a,s'} \lambda(\mathrm{d}s',\mathrm{d}a)\partial_\theta\pi_\theta(a|s,g)p(s'|s,a)v^{\pi_\theta}(s',g)\right) \tag{228}$$

We consider the first part. The following is the reversed derivation of $\tilde{p}$ in Equations (138)-(141).We have:

$$\int_g \lambda(\mathrm{d}g)f(g)\left(\gamma\int_a \partial_\theta\pi_\theta(a|s,g)\tilde{p}(g|s,a)\right) = \gamma\int_{g,a} \lambda(\mathrm{d}g,\mathrm{d}a)f(g)\partial_\theta\pi_\theta(a|s,g)\tilde{p}(g|s,a) \tag{229}$$

$$= \gamma\int_{g,a,s'} \lambda(\mathrm{d}a,\mathrm{d}s')f(g)\partial_\theta\pi_\theta(a|s,g)p(s'|s,a)\delta_{\varphi(s')}(\mathrm{d}g) \tag{230}$$

Therefore:

$$\int_g f(g)G(s,g,\theta)\lambda(\mathrm{d}g) = \gamma\int_{g,a,s'} \lambda(\mathrm{d}a,\mathrm{d}s')f(g)\partial_\theta\pi_\theta(a|s,g)p(s'|s,a)\left(\delta_{\varphi(s')}(\mathrm{d}g) + v^{\pi_\theta}(s',g)\lambda(\mathrm{d}g)\right) \tag{231}$$

$$= \gamma\int_{g,a,s'} \lambda(\mathrm{d}a,\mathrm{d}s')f(g)\partial_\theta\pi_\theta(a|s,g)p(s'|s,a)V^{\pi_\theta}(s',\mathrm{d}g) \tag{232}$$

This establishes equation (223). Finally, from (222) and (223), we have:

$$\partial_\theta J(\pi_\theta) = \frac{1}{1-\gamma}\int_{g,s_0,s,a,s'} \lambda(\mathrm{d}s_0)\gamma p_{\mathcal{G}}(g)p_0(s_0|g)\nu^{\pi_\theta}(\mathrm{d}s|s_0,g)\partial_\theta\pi_\theta(a|s,g)p(s'|s,a)V^{\pi_\theta}(s',\mathrm{d}g) \tag{233}$$

We now show that Then, we have: $\lim_{n\to\infty}\mathbb{E}\left[\widehat{\delta\theta}^{(n)}_{\delta\text{-AC}}(s,a,s',g)\right] = \frac{1-\gamma}{c}\partial_\theta J(\pi_\theta)$ and $\lim_{\varepsilon\to 0}\frac{1}{\lambda(\varepsilon)}\partial_\theta J_\varepsilon(\pi_\theta) = \partial_\theta J(\pi_\theta)$.

We first compute $\partial_\theta J_\varepsilon(\pi_\theta)$. We apply the policy gradient theorem (Sutton & Barto, 2018) to the augmented state augmented (non-multi goal) environment $\tilde{\mathcal{S}} = \mathcal{S}\times\mathcal{G}$, and we have, for any *baseline* function $b(\tilde{s})$ with $\tilde{s}\in\tilde{\mathcal{S}}$:

$$\partial_\theta J_\varepsilon(\pi_\theta) = \frac{1}{1-\gamma}\int_{\tilde{s}_0,\tilde{s},a,\tilde{s}'} \lambda(\mathrm{d}a)\rho_0(\tilde{s}_0)\nu^{\pi_\theta}(\mathrm{d}\tilde{s}|\tilde{s}_0)\tilde{P}(\mathrm{d}\tilde{s}'|\tilde{s},a)\partial_\theta\pi_\theta(a|\tilde{s})\left(R_\varepsilon(\tilde{s}) + \gamma V^\pi_\varepsilon(\tilde{s}') - b(\tilde{s})\right) \tag{234}$$

$$= \frac{1}{1-\gamma}\int_{g,s_0,s,a,s'} \lambda(\mathrm{d}a)\rho_{\mathcal{G}}(\mathrm{d}g)\rho_0(\mathrm{d}s_0|g)\nu^{\pi_\theta}(\mathrm{d}s|s_0,g)\tilde{P}(\mathrm{d}s'|s,a)\partial_\theta\pi_\theta(a|s,g)\left(R_\varepsilon(s,g) + \gamma V^\pi_\varepsilon(s',g) - b(s,g)\right) \tag{235}$$

with the change of variable $\tilde{s} = (s, g)$, $\tilde{s}' = (s', g)$, $\tilde{s}_0 = (s_0, g)$. We use the baseline $b(s, g) = R_\varepsilon(s, g)$, and we have:

$$\frac{1}{\lambda(\varepsilon)}\partial_\theta J_\varepsilon(\pi_\theta) = \frac{1}{1-\gamma}\int_{s_0,s,a,s',g} \lambda(\mathrm{d}s_0, \mathrm{d}a, \mathrm{d}g) p_{\mathcal{G}}(g) p(s_0|g) \nu^\pi(\mathrm{d}s|s_0, g) \partial_\theta \pi_\theta(a|s, g) \left(\frac{\gamma V_\varepsilon(s', g)}{\lambda(\varepsilon)}\right) \tag{236}$$

We now compute $\mathbb{E}\left[\widehat{\delta\theta}^{(n)}_{\delta\text{-AC}}(s, a, s', g)\right]$. We have:

$$\mathbb{E}\left[\widehat{\delta\theta}^{(n)}_{\delta\text{-AC}}(s, a, s', g)\right] \tag{237}$$

$$= \int_{s_0,s,a,s',g} \lambda(\mathrm{d}g, \mathrm{d}s_0, \mathrm{d}s', \mathrm{d}a) \frac{1}{c} p_{\mathcal{G}}(g)^2 p(s_0|g) \nu^\pi(\mathrm{d}s|s_0, g) \pi_\theta(a|s, g) \partial_\theta \log \pi_\theta(a|s, g) (\gamma v_n(s', g) - v_n(s, g)) \tag{238}$$

$$= \int_{s_0,s,a,s',g} \lambda(\mathrm{d}g, \mathrm{d}s_0, \mathrm{d}s', \mathrm{d}a) \frac{1}{c} p_{\mathcal{G}}(g)^2 p(s_0|g) \nu^\pi(s|s_0, g) \partial_\theta \pi_\theta(a|s, g) (\gamma v_n(s', g) - v_n(s, g)) \tag{239}$$

We know that for every *baseline* function $b(s, g)$:

$$\int_a \partial_\theta \pi_\theta(a|s, g) b(s, g) = b(s, g) \partial_\theta \int_a \pi_\theta(a|s, g) = 0 \tag{240}$$

We define $b(s, g) = v_n(s, g)$, and we have:

$$\mathbb{E}\left[\widehat{\delta\theta}^{(n)}_{\delta\text{-AC}}(s, a, s', g)\right] = \gamma \int_{s_0,s,a,s',g} \lambda(\mathrm{d}g, \mathrm{d}s_0, \mathrm{d}s', \mathrm{d}a) \frac{1}{c} p_{\mathcal{G}}(g)^2 p_0(s_0|g) \nu^\pi(\mathrm{d}s|s_0, g) \partial_\theta \pi_\theta(a|s, g) p(s'|s, a) v_n(s', g) \tag{241}$$

We know from Lemma 14 that $\nu^\pi(\mathrm{d}s|s_0, g) = (1-\gamma)\delta_{s_0}(\mathrm{d}s) + q^\pi(s|s_0, g)\lambda(\mathrm{d}s)$ where q^π is continuous, bounded, and with compact support as a density. Therefore, for any goal g, if we take the expectation with respect to $s_0 \sim p_0(s_0|g)$:

$$\int_{s_0} p_0(s_0|g) \nu^\pi(\mathrm{d}s|s_0, g) = \int_{s_0} (1-\gamma) p_0(s_0|g) \delta_{s_0}(\mathrm{d}s) + q^\pi(s|s_0, g)\lambda(\mathrm{d}s) \tag{242}$$

$$= \left((1-\gamma)p_0(s|g) + \int_{s_0} p_0(s_0|g) q^\pi(s|s_0, g)\right)\lambda(\mathrm{d}s) \tag{243}$$

$$= \tilde{q}^\pi(s|g)\lambda(\mathrm{d}s) \tag{244}$$

where $\tilde{q}^\pi$ is continuous, bounded and with compact support as a density. Moreover, $p_{\mathcal{G}}$ and $p_0(s_0|g)$ are continuous bounded functions. Therefore:

$$\mathbb{E}\left[\widehat{\delta\theta}^{(n)}_{\delta\text{-AC}}(s, a, s', g)\right] = \gamma \int_{s',g} \lambda(\mathrm{d}s', \mathrm{d}g) v_n(s', g) \frac{1}{c} p_{\mathcal{G}}(g)^2 \int_{s,a} \lambda(\mathrm{d}s, \mathrm{d}a) \tilde{q}(s, g) p(s'|s, a) \partial_\theta \pi_\theta(a|s, g) \tag{245}$$

and similarly:

$$\frac{1}{\lambda(\varepsilon)}\partial_\theta J_\varepsilon(\pi_\theta) = \frac{\gamma}{1-\gamma}\int_{s,a,s',g} \lambda(\mathrm{d}s, \mathrm{d}a, \mathrm{d}s', \mathrm{d}g) p_{\mathcal{G}}(g) \tilde{q}(s|g) \partial_\theta \pi(a|s, g) p(s'|s, a) \frac{V_\varepsilon(s', g)}{\lambda(\varepsilon)} \tag{246}$$

$$= \frac{\gamma}{1-\gamma}\int_{s',g} \lambda(\mathrm{d}s', \mathrm{d}g) \frac{V_\varepsilon(s', g)}{\lambda(\varepsilon)} p_{\mathcal{G}}(g) \int_{s,a} \lambda(\mathrm{d}s, \mathrm{d}a) \tilde{q}(s, g) p(s'|s, a) \partial_\theta \pi_\theta(a|s, g) \tag{247}$$

We know that the two measures on $K_{\mathcal{S}} \times \mathcal{G}$ defined as $\lambda(\mathrm{d}s, \mathrm{d}g)\frac{V_\varepsilon(s,g)}{\lambda(\varepsilon)}$ and $\lambda(\mathrm{d}s)v_n(s, g)\rho(\mathrm{d}g) = \lambda(\mathrm{d}s, \mathrm{d}g)v_n(s, g)p_{\mathcal{G}}(g)$ converges weakly to $\lambda(\mathrm{d}s)V^{\pi_\theta}(s, \mathrm{d}g)$. Moreover, $(s', g) \to$

$\int_{s,a} \tilde{q}(s,g)p(s'|s,a)\partial_\theta \pi_\theta(a|s,g)$ is continuous and bounded because $\tilde{q}$, p and $\partial_\theta \pi_\theta$ are continuous, bounded, and the supports are compact. Therefore, from equation (245):

$$\mathbb{E}\left[\widehat{\delta\theta}^{(n)}_{\delta\text{-AC}}(s,a,s',g)\right] \to_{n\to\infty} \frac{\gamma}{c}\int_{s,a,s',g} \lambda(\mathrm{d}s,\mathrm{d}s',\mathrm{d}a)p_{\mathcal{G}}(g)V^{\pi_\theta}(s',\mathrm{d}g)\tilde{q}(s,g)p(s'|s,a)\partial_\theta\pi_\theta(a|s,g) \tag{248}$$

$$= \frac{\gamma}{c}\int_{s_0,s,a,s',g} \lambda(\mathrm{d}s_0,\mathrm{d}s,\mathrm{d}a,\mathrm{d}s')p_{\mathcal{G}}(g)p_0(s_0|g)\nu^{\pi}(\mathrm{d}s|s_0,g)V^{\pi_\theta}(s',\mathrm{d}g)\gamma p(s'|s,a)\partial_\theta\pi_\theta(a|s,g) \tag{249}$$

$$= \frac{1-\gamma}{c}\partial_\theta J(\pi_\theta) \tag{250}$$

and from equation (247)

$$\frac{1}{\lambda(\varepsilon)}\partial_\theta J_\varepsilon(\pi_\theta) \to_{\varepsilon\to 0} \partial_\theta J(\pi_\theta) \tag{251}$$

This concludes the proof. □