Which Features are Best for Successor Features?

Yann Ollivier

Abstract

In reinforcement learning, universal successor features (SFs) are a way to provide zero-shot adaptation to new tasks at test time: they provide optimal policies for all downstream reward functions lying in the linear span of a set of base features. But it is unclear what constitutes a good set of base features, that could be useful for a wide set of downstream tasks beyond their linear span. Laplacian eigenfunctions (the eigenfunctions of $\Delta + \Delta^*$ with Δ the Laplacian operator of some reference policy and Δ^* that of the time-reversed dynamics) have been argued to play a role, and offer good empirical performance.

Here, for the first time, we identify the optimal base features based on an objective criterion of downstream performance, in a non-tautological way without assuming the downstream tasks are linear in the features. We do this for three generic classes of downstream tasks: reaching a random goal state, dense random Gaussian rewards, and random "scattered" sparse rewards. The features yielding optimal expected downstream performance turn out to be the *same* for these three task families. They do not coincide with Laplacian eigenfunctions in general, though they can be expressed from Δ: in the simplest case (deterministic environment and decay factor γ close to 1), they are the eigenfunctions of $\Delta^{-1} + (\Delta^{-1})^*$.

We obtain these results under an assumption of large behavior cloning regularization with respect to a reference policy, a setting often used for offline RL. Along the way, we get new insights into KL-regularized policy gradient, and into the lack of SF information in the norm of Bellman gaps.

1 Introduction and Related Work

The successor features (SFs) framework holds the promise of solving any new reinforcement learning task in a fixed environment in an almost zero-shot manner, without extensive learning or planning for each new task (see e.g., [BBQ+18, TRO23, TO21] and the references therein). At train time, the agent observes reward-free transitions in an environment, and learns some features and a parametric family of policies. At test time, the agent is faced with a new task specified via a reward function r. The function r is linearly projected onto the set of features, and then a suitable pre-trained policy is applied. At test time, the linear projection requires no learning or planning, and only uses a relatively small number of reward values (or knowledge of the reward function itself, e.g., for goal-reaching).

A good choice of features is crucial for successor features, and several approaches to select relevant features for SFs have been proposed. Given a finite number of features φ, SFs can produce policies within a family of tasks directly related to φ [BDM+17, BBQ+18, ZSBB17, GHB+19], but this often uses hand-crafted φ or features φ that linearize some known training rewards. VISR and APS [HDB+19, LA21] build φ automatically via diversity criteria [EGIL18, GRW16]. Forward-backward representations [TO21, TRO23] use successor measures [BTO21] to learn features φ that model the main variations of the distribution of visited states depending on the starting point and policy. In related contexts, further spectral variants have been proposed to provide a basis of features to express reward functions and Q-functions: the singular value decomposition of the inverse Laplacian, the singular value decomposition of the transition matrix P_{π_0} induced by an exploration policy π_0, the eigenfunctions of $P_{\pi_0} + P^*_{\pi_0}$, and more (see [GB20, RZL+22] and the references therein).

[TRO23] compared SFs with features built from a number of approaches: Laplacian eigenfunctions [WTN18, MM07], forward-backward representations, APS, auto-encoders, inverse dynamics models, spectral decompositions of the transition matrix, and more. Forward-backward representations and Laplacian eigenfunctions were found to perform best on average.

Yet all these choices of features are based on somewhat heuristic arguments and discussions. Here, for the first time, we fully characterize the best features for SFs mathematically, by directly optimizing expected downstream performance.

We define three agnostic models of downstream tasks: random Gaussian reward functions; reaching a random goal state; and "scattered" random rewards (a random number of sparse rewards placed at randomly located states, with random signs and magnitudes). We then ask which features provide the best expected performance on these tasks.

Surprisingly, the conclusions are identical for all three classes: *even though these three models cover very different tasks* (dense rewards, single-state rewards, multiple sparse rewards), *the best features are the same.* Thus, the conclusions are robust to the precise choice of a downstream task model. The existence of clear optimal features for these tasks is itself a nontrivial result, given the prior-free nature of these reward models.

Overview of results. All along, following previous work for offline reinforcement learning (see e.g. the survey [LKTF20]), we work with entropy-regularized policies that stay relatively close to a reference policy π_0. We estimate the *regularized return* G_r^π of a policy π for a reward r, which includes a Kullback–Leibler penalty for deviating from a π_0 (Definition 1). We assume that the regularization constant (temperature) T is relatively large, and derive optimal features up to an error $O(1/T^2)$.

Namely, we look for the features φ such that, when using the policy $\hat{\pi}$ estimated by successor features (Definition 4), the expected regularized return

$$\mathbb{E}_r[G_r^{\hat{\pi}}] \tag{1}$$

of $\hat{\pi}$ is maximal. The expectation is over rewards r in three simple models of downstream tasks (Section 3). Our main findings are as follows. All results are up to an error $O(1/T^2)$ on optimality.

- In deterministic environments, the optimal features for regularized successor features are the largest eigenfunctions of

$$\Delta^{-1} + (\Delta^{-1})^* - (1-\gamma^2)(\Delta^{-1})^*\Delta^{-1} \tag{2}$$

  (Theorem 12), where $\Delta := \mathrm{Id} - \gamma P_{\pi_0}$ is the Laplacian operator of the reference policy π_0, and $(\Delta^{-1})^*$ is the adjoint of Δ^{-1} acting on $L^2(\rho)$. [1] Here ρ is the stationary distribution of state-actions under the reference policy.

- For $\gamma \to 1$ this reduces to the largest eigenfunctions of $\Delta^{-1} + (\Delta^{-1})^*$ (Theorem 10), corresponding to long-range (low frequency) information on the reward function.

- For $\gamma \to 0$ this reduces to the smallest eigenfunctions of $P_{\pi_0}^* P_{\pi_0}$ (Proposition 11), corresponding to short-range (high-frequency) information on the reward function.

- For general, non-deterministic environments, we show that the optimal features are the reward functions r whose *advantage functions* have maximal norm for a given norm of r (Theorem 8 and Corollary 9).

  Explicitly, the optimal features are the largest eigenfunctions of the matrix

$$\mathrm{diag}(\rho)^{-1}(\Delta^{-1})^\top \left(\mathrm{diag}(\rho) - \pi_0^\top \mathrm{diag}(\rho_S)\pi_0\right) \Delta^{-1}. \tag{3}$$

  where again, ρ is the stationary distribution of state-actions under π_0, and ρ_S is its marginal on states. In deterministic environments, this expression simplifies to the results above.

- We show that, in successor features, the average norm of Bellman gaps for downstream tasks is uninformative as to which features perform best: it only depends on the number of linearly independent features (Proposition 6).

[1] This adjoint is $\mathrm{diag}(\rho)^{-1}(\Delta^{-1})^\top\mathrm{diag}(\rho)$, not $(\Delta^{-1})^\top$.

- As a key intermediate result, we show that for KL-penalized policy improvement, the optimality gap due to imperfect Q-function estimation is equal to the norm of the error on the advantage function (Theorem 2).

Comparison with Laplacian eigenfunctions and forward-backward representations. These results complement the empirical results in [TRO23], where forward-backward representations and Laplacian eigenfunctions performed best among the SF methods tested.

The Laplacian operator of a policy π_0 is $\Delta = \mathrm{Id} - \gamma P_{\pi_0}$, and Laplacian eigenfunctions are the smallest eigenfunctions of $\Delta + \Delta^*$. Forward-backward representations learn a finite-rank approximation of Δ^{-1} as $F^\top B$ and then use B as features. Here we find the optimal features to be the eigenfunctions of $\Delta^{-1} + (\Delta^{-1})^*$ (in the simplest case of a deterministic environment and $\gamma \to 1$). In general $\Delta^{-1} + (\Delta^{-1})^* \neq (\Delta + \Delta^*)^{-1}$ except in very specific environments (see discussion after Theorem 10). So in general, the optimal features derived here differ from Laplacian eigenfunctions. They also differ from B in the forward-backward representation: they would be closer to extracting the symmetric part of the finite-rank approximation $F^\top B$.

Still, our results may explain the good empirical performance of forward-backward representations and Laplacian eigenfunctions for SFs: they are the methods that come closest to the theoretically optimal features among the methods tested in [TRO23].

Limitations and perspectives. A first limitation of these results comes from the entropy regularization: all the statements about optimality hold up to an error $O(1/T^2)$. It is not clear how far this approximation extends in practice. Still, a regularized setup is natural for zero-shot reinforcement learning based on a fixed foundation model such as successor feature: the policies deployed at test time only depend on information from the trainset used to build the model, so it makes sense not to deviate too much from the states and behaviors explored in the trainset [LKTF20].

Second, in this work, we provide no algorithms to actually learn the optimal features (either the eigenfunctions of $\Delta^{-1} + (\Delta^{-1})^*$, or the reward functions r that maximize the norm of the advantage function). This is left to future work.

Finally, more fundamental limitations come from the setup of successor features itself. SFs rely on the linear projection of the reward onto a subspace of features. This method can only be exact in a linear subspace of reward functions, and the task encoding is linear. Recent works such as attempt to remove this limitation by defining *auto-regressive features* that let finer task encoding features depend on previously computed, coarse task encoding features, resulting in a fully nonlinear task encoding. SFs are also limited to tackling new tasks within a given environment and dynamics, not new environment or different dynamics in the same environment; some workarounds have been proposed to handle similar enough environments [ZSBB17, ALG+21].

Acknowledgments. The author would like to thank Andrea Tirinzoni and Ahmed Touati for comments on the text, suggestions for references, and helpful discussions around this topic.

2 Preliminaries: Notation, Universal Successor Features, Regularized Policies

2.1 Setup and Notation

Markov decision processes, matrix notation. Let $\mathcal{M} = (S, A, P, \gamma)$ be a reward-free Markov decision process (MDP) with state space S, action space A, transition probabilities $P(s'|s,a)$ from state s to s' given action a, and discount factor $0 < \gamma < 1$ [SB18]. We assume that S and A are finite. [2] A policy π is a function $\pi\colon S \to \mathrm{Prob}(A)$ mapping a state s to the probabilities of actions in A. Given $(s_0, a_0) \in S \times A$ and a policy π, we denote $\mathbb{E}[\cdot|s_0, a_0, \pi]$ the expectations under state-action sequences $(s_t, a_t)_{t\geqslant 0}$ starting at (s_0, a_0) and following policy π in the environment, defined by sampling $s_t \sim P(s_t|s_{t-1}, a_{t-1})$ and $a_t \sim \pi(a_t|s_t)$. We define $P_\pi(s', a'|s, a) := P(s'|s, a)\pi(a'|s')$, the state-action transition probabilities induced by π. Given a reward function $r\colon S \times A \to \mathbb{R}$, the Q-function of π for r is $Q_r^\pi(s_0, a_0) := \sum_{t\geqslant 0} \gamma^t \mathbb{E}[r(s_t, a_t)|s_0, a_0, \pi]$. The *value function* of π for r is $V_r^\pi(s) := \mathbb{E}_{a\sim\pi(s)} Q_r^\pi(s, a)$, and the *advantage function* is is $A_r^\pi(s, a) := Q_r^\pi(s, a) - V_r^\pi(s)$.

For the statements and proofs, it is convenient to view these objects as vectors and matrices. We treat rewards and Q-functions as column vectors of size $\#S \times \#A$. We view P as a matrix of size $(\#S \times \#A) \times \#S$ with entries $P_{(sa)s'} = P(s'|s, a)$. A policy π is seen as a matrix of size $\#S \times (\#S \times \#A)$ with entries $\pi_{s(s'a)} = \pi(a|s)\mathbb{1}_{s=s'}$. We denote $P_\pi := P\pi$. With these conventions, the Q-function satisfies the Bellman equation

$$Q_r^\pi = r + \gamma P_\pi Q_r^\pi. \tag{4}$$

Reference policy π_0, invariant distribution ρ. Let π_0 be some reference policy. We assume that π_0 is ergodic. Let ρ be the asymptotic distribution of state-actions induced by π_0 (the stationary distribution of P_{π_0}). Typically, π_0 is an exploration policy used to build some training set for RL algorithms;

[2] In principle, most of the ideas in this text extend to continuous spaces, but this would make the statements unnecessarily technical.

in this situation, the distribution of states and actions in the training set is approximately ρ. Let ρ_S be the marginal distribution of states under ρ. We abbreviate

$$\grave{\rho} := \operatorname{diag}(\rho), \qquad \grave{\rho}_S := \operatorname{diag}(\rho_S) \tag{5}$$

the diagonal matrices of size $\#S \times \#A$ with diagonal entries equal to $\grave{\rho}_{(sa)(sa)} = \rho(s,a)$, and of size $\#S$ with diagonal entries equal to ρ_S, respectively.

$L^2(\rho)$ norm, advantage norm. Two for functions f on $S \times A$ will play an important role: the $L^2(\rho)$ norm, defined as

$$\|f\|^2_{L^2(\rho)} := \mathbb{E}_{(s,a)\sim\rho}[f(s,a)^2] = f^\top \grave{\rho} f \tag{6}$$

and the *advantage norm*, defined as

$$\|f\|^2_A := \mathbb{E}_{(s,a)\sim\rho}\left[\left(f(s,a) - \mathbb{E}_{a'\sim\pi_0(s)}[f(s,a')]\right)^2\right]. \tag{7}$$

By construction,

$$\|Q^{\pi_0}_r\|_A = \|A^{\pi_0}_r\|_{L^2(\rho)} \tag{8}$$

though this does not hold for other policies $\pi \neq \pi_0$, because the expectation inside $\|f\|^2_A$ is with respect to π_0.

We denote $\langle\cdot,\cdot\rangle_{L^2(\rho)}$ and $\langle\cdot,\cdot\rangle_A$ the inner products associated with these two norms.

The *adjoint* of a linear operator M on $L^2(\rho)$ is the unique operator M^* such that $\langle x, My\rangle_{L^2(\rho)} = \langle M^*x, y\rangle_{L^2(\rho)}$ for any $x, y \in L^2(\rho)$. It is given by the matrix $M^* = \grave{\rho}^{-1} M^\top \grave{\rho}$.

Working in $L^2(\rho)$ rather than the Euclidean metric on rewards and Q-functions is mathematically the most natural way to have results that still make sense for general state spaces beyond the finite case. It also links all metrics to the data: since the distribution of samples in the training set is approximately ρ, the norm $\|f\|^2_{L^2(\rho)} = \mathbb{E}_{(s,a)\sim\rho}[f(s,a)^2]$ can be estimated empirically. In contrast, the Euclidean norm $\|f\|^2$ would be estimated by sampling random states uniformly distributed in the full space (or in a given domain of $\mathbb{R}^n$ for continuous states): such states might be irrelevant or even non-realistic. (See also the discussion of reward models in Section 3.)

Laplacian operator Δ. For $\gamma \leqslant 1$, we define the *Laplacian operator* of π_0 as

$$\Delta := \operatorname{Id} - \gamma P_{\pi_0}. \tag{9}$$

The Bellman equation rewrites $\Delta Q^{\pi_0}_r = r$.

For $\gamma < 1$ we have $\Delta^{-1} = \sum_{t\geqslant 0} \gamma^t P^t_{\pi_0}$. For $\gamma = 1$ (the most standard definition of the Laplacian), $\Delta = \operatorname{Id} - P_{\pi_0}$ is not invertible, as 1 is an eigenvalue of P_{π_0} associated with the constant eigenvector $\mathbb{1}$. However, since π_0 is ergodic, the multiplicity of this eigenvalue is 1: by the Perron–Frobenius theorem, Δ is invertible on the orthogonal of the constant functions.

We denote by $L^2_0(\rho)$ the subset of functions f on $S \times A$ whose average under ρ is 0. Thus, for $\gamma = 1$, Δ is invertible on $L^2_0(\rho)$.

2.2 Regularized Policies, Regularized Return

We are only going to consider policies that do not deviate too much from the reference policy π_0. This is often considered in the offline reinforcement learning setup or to enforce safety [LKTF20].

This can be done by adding a Kullback–Leibler regularization term to the reward function, as described, for instance, in [LKTF20, §4.3] (policy penalty methods), which we follow here.

Definition 1 (Regularized return). *Let $T \geqslant 0$ be a temperature parameter. We define the* regularized reward function *for a policy π as*

$$\bar{r}(s,a) := r(s,a) - T\,\mathrm{KL}(\pi(s)\,\|\,\pi_0(s)) \tag{10}$$

where $\mathrm{KL}(\pi(s)\,\|\,\pi_0(s))$ is the Kullback–Leibler divergence between the policies π_0 and π at s.

We define the regularized return *of policy π for reward r as its expected return for the regularized reward, namely,*

$$G^\pi_r := \mathbb{E}_{s_0\sim\rho}\left[\sum_{t\geqslant 0}\gamma^t \bar{r}(s_t,a_t) \mid s_0, \pi\right] \tag{11}$$

and we say that a policy is regularized-optimal *for r if it maximizes G^π_r.*

Maximizing G^π_r is equivalent to maximizing the expected return $\mathbb{E}_{s\sim\rho}V^{\pi_r}(s)$ plus a behavior cloning term that keeps π close to π_0 at each state.

For large T, the maximizer π is $O(1/T)$-close to π_0, as the penalty term dominates. This justifies that, in the following, we consider policies that are parameterized as $\mathrm{Bolt}_{\pi_0}(f)$ for some f, at temperature T.

There are several variants of this definition: e.g., we could have considered the KL divergence in the opposite direction, or we could have estimated the KL term at states $s \sim \rho$ visited by π_0, instead of states s visited by π as in (11). In the regime we consider (large T), these variants all lead to the same conclusions, because the differences are $O(1/T^2)$. For instance, using a behavior cloning penalty

$$\mathbb{E}_{s\sim\rho, a\sim\pi_0(s)} \ln \pi(a|s) \tag{12}$$

would provide the same conclusions. For large T, maximizing the BC-regularized return is also equivalent to performing one step of natural policy gradient starting at π_0, with learning rate $1/T$. Thus, for large T, the regularized return has many different interpretations.

One can check that Boltzmann policies given by the Q-functions of π_0 are approximately regularized-optimal (Corollary 3). But we will need a finer result, which describes the optimality gap depending on the policy, as follows.

Theorem 2 (Regularized return of Boltzmann policies). *Let r be any reward function, and let $Q_r^{\pi_0}$ be the Q-function of the reference policy for reward r. Let $\hat{Q}$ be any function on $S \times A$, and consider the policy $\pi = \mathrm{Bolt}_{\pi_0}(\hat{Q})$.*

When $T \to \infty$, the regularized return of policy π satisfies

$$G_r^\pi = G_r^{\pi_0} + \frac{1}{2T(1-\gamma)} \left(\|Q_r^{\pi_0}\|_A^2 - \left\|\hat{Q} - Q_r^{\pi_0}\right\|_A^2 \right) + O(1/T^2). \tag{13}$$

In particular, at first order, this regularized return is maximal when $\hat{Q} = Q_r^{\pi_0}$, corresponding to the Boltzmann policy $\pi = \mathrm{Bolt}_{\pi_0}(Q_r^{\pi_0})$. (Note that Boltzmann policies have been defined with respect to π_0, so $\hat{Q} = 0$ corresponds to $\pi = \pi_0$.)

Corollary 3. *For any reward r, the policy $\pi = \mathrm{Bolt}_{\pi_0}(Q_r^{\pi_0})$ maximizes G_r^π up to an $O(1/T^2)$ error when $T \to \infty$.*

Policy gradient algorithms such as PPO and TRPO also learn a new policy π_{t+1} by maximizing expected return subject to a constraint on the KL divergence between π_{t+1} and the current policy π_t [SWD+17, SLA+15]. The KL constraint is implemented in a different way, not directly as a penalty as in (10), but the intuition is similar, with Boltzmann policies $\pi_0 \exp(\hat{Q}/T)$ corresponding to updating the log-probabilities of the policy π_t, with learning rate $1/T$.

Therefore, qualitatively, Theorem 2 stresses that the optimality gap in KL-regularized policy gradient updates is directly given by the L^2 error on the advantage function.

2.3 Successor Features for Zero-Shot RL, Regularized Successor Features

Successor features pre-compute the Q-functions of a number of basic reward functions $\varphi_1, \ldots, \varphi_d$ or their linear combinations. At test time, the reward function for the test task is projected onto this basis, and a precomputed policy is applied [BDM+17, BBQ+18]. This results in a *zero-shot* approach to new RL tasks in a given environment, since no learning or planning is needed at test time. The only computation at test time is the coefficients z of the linear projection of the reward r onto the features φ,

$$z = (\mathbb{E}_{(s,a)\sim\rho}[\varphi(s,a)\varphi(s,a)^\top])^{-1}\mathbb{E}_{(s,a)\sim\rho}[r(s,a)\varphi(s,a)] \tag{14}$$

which only requires to estimate the correlation $\mathbb{E}_{(s,a)\sim\rho}[r(s,a)\varphi(s,a)]$ between the reward and the features. (The covariance matrix can be precomputed.) This correlation can be estimated empirically given some reward samples.

In this text, since we are concerned with *regularized* policies and return, we will only need the simplest version of successor features, in which we only compute successor features with respect to the reference policy π_0. This plays out as follows.

Definition 4 (Regularized successor features). *Regularized successor features (RSFs) is the following procedure for regularized zero-shot RL.*

Let $\varphi\colon S \times A \to \mathbb{R}^d$ be a fixed d-dimensional feature map. At train time, compute a successor feature map $\psi\colon S \times A \to \mathbb{R}^d$ satisfying

$$\psi(s_0, a_0) = \mathbb{E}\left[\sum_{t\geqslant 0} \gamma^t \varphi(s_t, a_t) \mid s_0, a_0, \pi_0\right], \tag{15}$$

namely, ψ solves the vector-valued Bellman equation $\psi = \varphi + \gamma P_{\pi_0}\varphi$. Also denote $C := \varphi^\top \hat{\rho}\varphi = \mathbb{E}_{(s,a)\sim\rho}[\varphi(s,a)\varphi(s,a)^\top]$.

Then, at test time, given any reward function r, estimate

$$z = C^{-1}\mathbb{E}_{(s,a)\sim\rho}[r(s,a)\varphi(s,a)], \tag{16}$$

estimate the Q-function of π_0 for r by

$$\hat{Q}(s,a) := z^\top \psi(s,a) \tag{17}$$

and apply the Boltzmann policy $\hat{\pi} = \mathrm{Bolt}_{\pi_0}(\hat{Q})$.

The following proposition is an immediate consequence of Corollary 3.

Proposition 5. *If r lies in the linear span of the features φ, then the policy $\hat{\pi}$ is regularized-optimal for r, up to an error $O(1/T^2)$.*

The Q-functions and policies recovered by SFs only depend on the linear span of the features φ. Thus, without loss of generality, we can assume that the features are linearly independent and apply a change of basis $\varphi \leftarrow C^{-1/2}\varphi$ in feature space, after which $C = \mathrm{Id}$. Thus, in the following, we always assume $C = \mathrm{Id}$.

Although derived in a different way, the forward-backward (FB) setup from [TO21, TRO23] can also be seen as projecting the rewards onto trained features $\varphi(s,a) = (\mathrm{Cov}_\rho B)^{-1}B(s,a)$ at test time, where B are the features learned by the FB model [TRO23]. Therefore, our conclusions also cover this case.

3 Three Reward Models for Downstream Tasks

We will compute the average performance of successor features for three families of rewards: random Gaussian rewards (with a white noise continuous limit), random goal-reaching (Dirac rewards), and "scattered random rewards". We define each of those in turn.

All the models depend on the distribution ρ of states in the training data (the stationary distribution of the exploration policy π_0): either via the norm $L^2(\rho)$, or via putting rewards at random states sampled from ρ. In practice, both can be estimated by sampling random states from the training data. [3]

These are some of the most agnostic models we can find on an arbitrary state equipped with an arbitrary probability distribution. These models do not favor spatial smoothness a priori: white noise is non-smooth and scale-free, while the goal-reaching and scattered rewards are sparse. In these models, the Fourier transform of the reward is uniformly spread over all frequencies. It is interesting that we can reach meaningful conclusions about optimal features even with such uninformative priors.

All models are built to have well-defined continuous-space limits, and still make sense in an abstract state space equipped with a measure ρ. To avoid excessive technicality, we restrict ourselves to the finite case in this text.

Gaussian rewards (white noise). For this model, we simply sample a random Gaussian reward vector r of size $S \times A$, with density $\propto \exp(-\|r\|^2_{L^2(\rho)}/2)$. Including a variance σ^2 just rescales the rewards, so we take $\sigma = 1$ for simplicity.

The corresponding continuous-space limit is a *white noise* random reward: a random distribution $r(s,a)$ such that for any subset $X \subset S \times A$, the integral $\int_X r(s,a)\rho(\mathrm{d}s, \mathrm{d}a)$ is a centered Gaussian with variance $\rho(X)$, and the integrals on two disjoint subsets X and X' are independent.

This model naturally places more variance (uncertainty) on the reward at places less covered by the training set, since the weight from $\|r\|^2_{L^2(\rho)}$ will be smaller where ρ is small.

Goal-reaching (Dirac rewards). In this model, we first select a random state-action $(s^\star, a^\star) \sim \rho$ in $S \times A$. Then we put a reward $1/\rho(s^\star, a^\star)$ at

[3] We could also have used uniform measures on finite state spaces, or the Lebesgue measure on continuous states. But, first, this does not extend to an abstract state space equipped with a policy π_0, for which the $L^2(\rho)$ norm is mathematically the most natural and the only norm available. Second, in practice, it is much more natural to sample states from the training data than to sample random states from the Lebesgue measure in a large domain, which might produce irrelevant or irrealistic states, while ρ will be supported on realistic states.

$(s^\star, a^\star)$, and 0 everywhere else:

$$r(s,a) = \frac{1}{\rho(s^\star, a^\star)} \mathbb{1}_{(s,a)=(s^\star,a^\star)}. \tag{18}$$

The $1/\rho$ factor maintains $\int r \, \mathrm{d}\rho = 1$. Without this scaling, all Q-functions degenerate to 0 in continuous spaces, as discussed in [BO21]. Indeed, if we omit this factor, and just set the reward to be 1 at a given goal state $s^\star \in S$ in a continuous space S, the probability of exactly reaching that state with a stochastic policy is usually 0, and all Q-functions are 0. Thanks to the $1/\rho$ factor, the continuous limit is a *Dirac function* reward, infinitely sparse, corresponding to the limit of putting a reward 1 in a small ball $B(s^\star, \varepsilon)$ of radius $\varepsilon \to 0$ around $s^\star$, and rescaling by $1/\rho(B(s^\star, \varepsilon))$ to keep $\int r \, \mathrm{d}\rho = 1$. This produces meaningful, nonzero Q-functions in the continuous limit [BO21].

This model combines well with successor features or the FB framework: indeed, the task representation vector z in (16) can be computed via the expectation

$$\mathbb{E}_{(s,a)\sim\rho}[r(s,a)\varphi(s,a)] = \varphi(s^\star, a^\star) \tag{19}$$

(both in finite spaces and in the continuous-space limit).

Scattered random rewards. This model amounts to putting Dirac rewards at several states instead of one, each with a different random magnitude and sign. This describes general mixtures of sparse rewards at several random locations.

This model depends on an intensity parameter $\kappa > 0$ which controls the number of states that have a nonzero reward. We also fix some arbitrary distribution p_w over $\mathbb{R}$ with some mean μ and variance $\sigma^2 \geqslant 0$, e.g., a Gaussian.

To build a scattered random reward, we sample a random N-tuple of state-actions $((s_1, a_1), \ldots, (s_N, a_N))$ by a general Poisson point process on $S \times A$ with intensity $\kappa\rho$: namely, we first sample an integer $N \sim \mathrm{Poisson}(\kappa)$, then sample N random state-actions $(s_i, a_i) \sim \rho$, $i = 1, \ldots, N$, independently. [4] At each (s_i, a_i), we place a Dirac reward as above, but multiply it by some random weight $w_i \sim p_w$, sampled independently from everything else. The reward is 0 on the rest of the space. Explicitly,

$$r(s,a) = \sum_{i=1}^{N} \frac{w_i}{\rho(s_i, a_i)} \mathbb{1}_{(s,a)=(s_i,a_i)}. \tag{20}$$

[4] The Poisson law ensures that the number of states selected in some part of $S \times A$ is independent from the number of states selected in any other part. In this process, the number of state-actions selected in any part $X \subset S \times A$ follows a Poisson law with parameter $\kappa\rho(X)$, with distinct subsets being independent. Once more, this ensures a meaningful continuous-time limit, as the general Poisson process is well-defined in a continuous space with measure ρ.

As for Dirac rewards above, the factor $1/\rho$ ensures a meaningful limit for continuous spaces (where r becomes a random distribution, a random sum of Dirac functions wrt ρ). In this model, the SF task representation vector z in (16) is given by

$$z = C^{-1} \sum_{i=1}^{N} w_i \, \varphi(s_i, a_i) \tag{21}$$

both in finite spaces and in the continuous-space limit.

4 The Bellman Gap Norm is Uninformative for Successor Features

We start with a negative result: for the families of rewards above, the size of expected Bellman gaps of the Q-functions estimated by SFs is *independent* of the choice of features: it only depends on the number of linearly independent features.

This *does not mean that all choices of features perform equally well*: as we will see, regularized returns do depend on the features. It just means that the squared Bellman gap error is a poor proxy.

Proposition 6 (Average Bellman gaps do not depend on the features). *Given a reward r, let $\hat{Q}(s,a) := z^\top \psi(s,a)$ be the Q-function (17) estimated by regularized successor features, with z given by (16). Assume the features are linearly independent.*

Then, for either the random Gaussian reward or the random goal-reaching reward of Section 3, on average for r in the model, the norm of the Bellman gaps of $\hat{Q}$ only depends on the number of features d. More precisely,

$$\mathbb{E}_r \left\| \hat{Q} - r - \gamma P_{\pi_0} \hat{Q} \right\|^2_{L^2(\rho)} = \#S \times \#A - d \tag{22}$$

where the expectation is over a random reward r from the model.

So, even if we had access to downstream reward functions r, we could not learn good features by minimizing the expected Bellman gaps on those rewards. No meaningful analysis of the performance of a set of features is possible based on the norm of Bellman errors.

We have expressed this theorem using Bellman gaps with respect to the reference policy π_0: this is what is needed for regularized SFs, thanks to Theorem 2 and Corollary 3. However, a similar result holds for *universal* successor features [BBQ+18], with Bellman gaps taken for the estimated optimal policy $\hat{\pi}$ for each reward r: the proof in Section 6 directly covers this case as well.

It might be surprising that the expected norm of Bellman gaps does not align well with the optimal return, since in practice, Q-functions are routinely learned by the TD algorithm based on reducing Bellman gaps. But TD does not actually minimize the expected norm of Bellman gaps, due to the "double sampling problem" and updating only the Q in the left-hand-side of the Bellman equation. In general, TD does not minimize any norm and can diverge [TVR97], but in certain cases it is known to minimize the *Dirichlet norm* [5] of the error on Q [Oll18]. For γ close to 1, the Dirichlet norm is the sum of the Bellman gap norm and the advantage seminorm (Proposition 16 with $\gamma = 1$). The advantage seminorm is the one that matters for KL-regularized policy gradient (Theorem 2). So in the end, the discrepancy between TD and minimizing expected Bellman gaps might be a blessing.

5 The Optimal Features for Regularized Successor Features

We first introduce the *advantage kernel*, a symmetric matrix that describes the norm of the advantage function for a given reward function.

Definition 7. *Let π be a fixed policy with invariant distribution ρ, and consider the map that to any reward function r associates the norm $\|A_r^\pi\|^2_{L^2(\rho)}$ of its advantage function. Since the Q-function is linear in r for a given π, this is a quadratic function of r for fixed π.*

Therefore, there exists a symmetric positive semi-definite matrix[6] $\mathcal{A}_\pi$ *of size $(\#S \times \#A) \times (\#S \times \#A)$ such that*

$$\|A_r^\pi\|^2_{L^2(\rho)} = r^\top \mathcal{A}_\pi r \tag{23}$$

for any r. We call this matrix the advantage kernel *of policy π.*

Since the Q-function of reward r for policy π_0 is $Q_r^{\pi_0} = \Delta^{-1} r$, one can compute $\mathcal{A}_{\pi_0}$ in terms of Δ^{-1}. Expressing the norm of the advantage function as the difference between the norms of the Q-function and value function, $\|Q_r^{\pi_0}\|_A^2 = \|Q_r^{\pi_0}\|^2_{L^2(\rho)} - \|\pi_0 Q_r^{\pi_0}\|^2_{L^2(\rho_S)} = (Q_r^{\pi_0})^\top (\grave{\rho} - \pi_0^\top \grave{\rho}_S \pi_0) Q_r^{\pi_0}$, we obtain

$$\mathcal{A}_{\pi_0} = (\Delta^{-1})^\top (\grave{\rho} - \pi_0^\top \grave{\rho}_S \pi_0) \Delta^{-1}. \tag{24}$$

The following theorem expresses the expected regularized return for regularized successor features depending on the features, for each of the three reward models defined in Section 3, and up to an $O(1/T^2)$ error. The optimal choice of features corresponds to the features that maximize $x^\top \mathcal{A}_{\pi_0} x$ for a given norm $\|x\|^2_{L^2(\rho)}$.

[5] The Dirichlet (semi)norm of a function f is $\langle f, (\mathrm{Id} - \gamma P_\pi) f \rangle_{L^2(\rho)} \geqslant 0$, also equal to $\frac{1}{2}\mathbb{E}(f(s_{t+1}, a_{t+1}) - f(s_t, a_t))^2)$. We refer to [Oll18]. This is also the objective minimized in Laplacian eigenfunctions.

[6] or positive semi-definite kernel for infinite state spaces S

Theorem 8 (Expected regularized return depending on the features). *Let φ be a set of basic features for successor features. Without loss of generality, we assume that these features are $L^2(\rho)$-orthonormal.*

Given a reward r, let $\hat{Q}(s,a) := z^\top \psi(s,a)$ be the Q-function (17) estimated by regularized successor features, with z given by (16). Let $\hat{\pi} = \mathrm{Bolt}_{\pi_0}(\hat{Q})$ be the estimated regularized-optimal policy.

Then, for the reward models of Section 3, on average for r in the model, the regularized return $G_r^{\hat{\pi}}$ of the estimated policy satisfies:

- *For the random Gaussian or random goal-reaching reward models,*

$$\mathbb{E}_r[G_r^{\hat{\pi}}] = \mathbb{E}_r[G_r^{\pi_0}] + \frac{1}{T(1-\gamma)} \operatorname{Tr}(\varphi^\top \mathcal{A}_{\pi_0} \varphi) + O(1/T^2). \tag{25}$$

- *For the scattered random reward model with intensity κ, mean μ and variance σ^2,*

$$\mathbb{E}_r[G_r^{\hat{\pi}}] = \mathbb{E}_r[G_r^{\pi_0}] + \frac{1}{T(1-\gamma)} \left(\kappa(\mu^2+\sigma^2) \operatorname{Tr}(\varphi^\top \mathcal{A}_{\pi_0} \varphi) - (\kappa\mu)^2 \varphi_{\mathrm{cst}}^\top \mathcal{A}_{\pi_0} \varphi_{\mathrm{cst}} \right) + O(1/T^2) \tag{26}$$

  where φ_{cst} is the $L^2(\rho)$-orthogonal projection of the constant reward $r = \mathbb{1}$ onto the span of the features.

This result allows us to work out which features optimize the gain: those that maximise $\operatorname{Tr}(\varphi^\top \mathcal{A}_{\pi_0} \varphi)$ under the constraint that φ is $L^2(\rho)$-orthonormal, leading to the following characterization. Scattered random rewards require a slightly longer proof to handle the extra term in (26), but the conclusion is the same.

Corollary 9 (Optimal features for regularized successor features). *For any of the three reward models of Section 3, the features φ that bring maximal regularized return up to $O(1/T^2)$ are the largest d eigenvectors of $\grave{\rho}^{-1} \mathcal{A}_{\pi_0}$, or equivalently the largest d extremal directions of $x^\top \mathcal{A}_{\pi_0} x / \|x\|^2_{L^2(\rho)}$. In that case we have*

$$\operatorname{Tr}(\varphi^\top \mathcal{A}_{\pi_0} \varphi) = \sum_{i=1}^{d} \lambda_i \tag{27}$$

where $\lambda_1, \ldots, \lambda_d$ are the largest d eigenvalues of $\grave{\rho}^{-1} \mathcal{A}_{\pi_0}$. [7]

We now turn to a more explicit characterization of these eigenvectors in deterministic environments. The results depend on the decay factor γ. We first consider the two extreme cases $\gamma = 1$ and $\gamma = 0$, as they lead to the simplest expressions.

[7] The operator $\grave{\rho}^{-1} \mathcal{A}_{\pi_0}$ is self-adjoint in $L^2(\rho)$, so it is diagonalizable. Its eigenvalues are the same as those of the symmetric matrix $\grave{\rho}^{-1/2} \mathcal{A}_{\pi_0} \grave{\rho}^{-1/2}$.

Remember that $\mathcal{A}_{\pi_0}$ vanishes for a constant reward (the advantage function is 0), so we only have to compute it for rewards orthogonal to the constants, $r \in L^2_0(\rho)$. For such rewards, Q-functions and advantage functions are defined even for $\gamma = 1$, and the Laplacian operator is invertible up to $\gamma = 1$.

Theorem 10 (Optimal features for $\gamma = 1$ in a deterministic environment). *Assume the environment is deterministic. For $\gamma = 1$ and $r \in L^2_0(\rho)$, the advantage kernel is given by*

$$r^\top \mathcal{A}_{\pi_0} r = \langle r, (\Delta^{-1} + (\Delta^{-1})^* - \mathrm{Id}) r\rangle_{L^2(\rho)} \tag{28}$$

where $\Delta := \mathrm{Id} - P_{\pi_0}$ is the Laplacian operator of π_0, and where $(\Delta^{-1})^* = \grave{\rho}^{-1}(\Delta^{-1})^\top \grave{\rho}$ is the adjoint of Δ^{-1} acting on $L^2_0(\rho)$.

Consequently, for $\gamma = 1$ in a deterministic environment, the optimal features are the largest eigenfunctions of $\Delta^{-1} + (\Delta^{-1})^*$.

Intuitively, the largest eigenfunctions of $\Delta^{-1} + (\Delta^{-1})^*$ correspond to the lowest frequencies (longest range variations) in the environment: for $\gamma = 1$, we want to keep information on the large-scale variations of the reward function.

For comparison, it has previously been suggested to use as features the smallest eigenfunctions of $\Delta + \Delta^*$ [WTN18], or equivalently, the largest eigenfunctions of $P_{\pi_0} + P^*_{\pi_0}$. The largest eigenfunctions of $P_{\pi_0} + P^*_{\pi_0}$ also convey a low-frequency intuition, but in a somewhat different way. Indeed, in general, symmetrizing does not commute with taking the inverse: $\Delta^{-1} + (\Delta^{-1})^* \neq (\Delta + \Delta^*)^{-1}$ in general. This equality can still happen, for instance if Δ is reversible, but reversibility of Δ is a very specific situation: for instance, this is never the case in kinematic environments. [8]

Therefore, in general, the optimal features are *not* the smallest eigenfunctions of $\Delta + \Delta^*$.

In contrast, for small γ, what matters are the *highest* frequencies of the reward function, as we show now.

Proposition 11 (Optimal features for $\gamma = 0$ in a deterministic environment). *Assume the environment is deterministic. For $\gamma = 0$ and $r \in L^2(\rho)$, the advantage kernel is given by*

$$r^\top \mathcal{A}_{\pi_0} r = \langle r, (\mathrm{Id} - P^*_{\pi_0} P_{\pi_0}) r\rangle_{L^2(\rho)} \tag{29}$$

[8] Indeed, reversibility implies that if a transition $(s, a) \to (s', a')$ is possible, then so is the reverse transition $(s', a') \to (s, a)$. This is not the case if speed is part of the state: if $s = (x, v)$ then the next state has $x' \approx x + \delta t\, v$, so the transition $(s', a') \to (s, a)$ would require $v' \approx -v$, namely, the ability to fully reverse speed with a single action.

Note that here we deal with reversibility in the language of Markov chain theory, not reversibility in the language of physics: in a physicist's language, classical mechanics are reversible (by changing v to $-v$).

where $P^*_{\pi_0} = \grave{\rho}^{-1} P^\top_{\pi_0} \grave{\rho}$ is the adjoint of P_{π_0} acting on $L^2(\rho)$.

Consequently, for $\gamma = 0$ in a deterministic environment, the optimal features are the smallest *eigenfunctions of $P^*_{\pi_0} P_{\pi_0}$.*

For $\gamma = 0$, the problem is a bandit problem: given an initial state s_0, just pick the action with the highest reward, after which the game ends. This relies on knowing how to compare the value of r on actions at the same state, which is high-frequency information about r. [9]

Finally, we give the general expression for $0 < \gamma < 1$ in a deterministic environment.

Theorem 12 (Optimal features for $0 < \gamma < 1$ in a deterministic environment). *Assume the environment is deterministic. For $0 < \gamma < 1$, the advantage kernel is given by*

$$r^\top \mathcal{A}_{\pi_0} r = \frac{1}{\gamma^2} \left\langle r, \left(\Delta^{-1} + (\Delta^{-1})^* - \operatorname{Id} - (1-\gamma^2)(\Delta^{-1})^* \Delta^{-1}\right) r \right\rangle_{L^2(\rho)}$$

where $\Delta := \operatorname{Id} - \gamma P_{\pi_0}$ is the Laplacian operator of π_0, and where $(\Delta^{-1})^* = \grave{\rho}^{-1} (\Delta^{-1})^\top \grave{\rho}$ is the adjoint of Δ^{-1} acting on $L^2(\rho)$.

Consequently, in a deterministic environment, the optimal features are the largest eigenfunctions of $\Delta^{-1} + (\Delta^{-1})^* - (1-\gamma^2)(\Delta^{-1})^* \Delta^{-1}$.

6 Proofs

Lemma 13. *Abbreviate $K(\pi, s) := \operatorname{TKL}(\pi(s) \,\|\, \pi_0(s))$ for the regularization term in Definition 1.*

Then

$$\operatorname{Bolt}_{\pi_0}(f)(s,a) = \pi_0(s,a) \left(1 + \frac{1}{T} f(s,a) - \frac{1}{T} \bar{f}(s)\right) + O(1/T^2) \tag{30}$$

and

$$K(\operatorname{Bolt}_{\pi_0}(f), s) = \frac{1}{2T} \mathbb{E}_{a \sim \pi_0(s)} \left(f(s,a) - \bar{f}(s)\right)^2 + O(1/T^2) \tag{31}$$

and consequently

$$\mathbb{E}_{s \sim \rho} \left[K(\operatorname{Bolt}_{\pi_0}(f), s)\right] = \frac{1}{2T} \|f\|^2_A + O(1/T^2) \tag{32}$$

Proof.

These follow from direct Taylor expansions. □

[9] This conclusion for $\gamma = 0$ heavily depends on the fact that we have defined rewards over *state-actions* all along. If rewards only depend on the state, there is no meaningful optimal behavior for $\gamma = 0$. One could study the $\gamma \to 0$ limit for reward models depending only on the state, but this is beyond the scope of the present work.

Proof of Theorem 2.
We want to estimate the regularized return G_r^π of $\pi = \mathrm{Bolt}_{\pi_0}(\hat{Q})$. By definition, it is the sum of the ordinary return and a penalty term,

$$G_r^\pi = \mathbb{E}_{s_0\sim\rho}[V_r^\pi(s_0)] - \mathbb{E}_{s_0\sim\rho}\,\mathbb{E}\left[\sum_{t\geqslant 0}\gamma^t K(\pi, s_t) \mid s_0, \pi\right] \tag{33}$$

where $K(\pi, s) := T\mathrm{KL}(\pi(s) \,\|\, \pi_0(s))$ is the penalty term as defined in Lemma 13.

We first estimate the value function V_r^π, then turn to the penalty term.

By definition of Boltzmann policies, $\pi = \mathrm{Bolt}_{\pi_0}(\hat{Q})$ is $O(1/T)$-close to π_0.

For π close to π_0, the policy gradient theorem provides the expression of the derivative of V^π with respect to π. Writing the policy gradient theorem as a Taylor expansion around $\pi \approx \pi_0$, we obtain

$$V^\pi(s_0) - V^{\pi_0}(s_0) = \mathbb{E}_{(s_t,a_t)}\left[\sum_{t\geqslant 0}\gamma^t A^{\pi_0}(s_t, a_t)\,(\ln\pi(a_t|s_t) - \ln\pi_0(a_t|s_t)) \mid s_0, \pi_0\right] + O((\pi - \pi_0)^2) \tag{34}$$

where A^{π_0} is the advantage function of policy π_0.

Let us average this over $s_0 \sim \rho$. Since ρ is the invariant distribution of π_0, each s_t above is also distributed according to ρ, and therefore all values of t make the same contribution:

$$\mathbb{E}_{s_0\sim\rho}[V^\pi(s_0) - V^{\pi_0}(s_0)] = \frac{1}{1-\gamma}\mathbb{E}_{s\sim\rho, a\sim\pi_0(s)}\left[A^{\pi_0}(s, a)\,(\ln\pi(a|s) - \ln\pi_0(a|s))\right] + O((\pi - \pi_0)^2) \tag{35}$$

By Lemma 13, for $\pi = \mathrm{Bolt}_{\pi_0}(\hat{Q})$ we have $(\pi - \pi_0)^2 = O(1/T^2)$ and moreover

$$\ln\pi(s, a) = \ln\pi_0(s, a) + \frac{1}{T}(\hat{Q}(s, a) - \bar{\hat{Q}}(s)) + O(1/T^2) \tag{36}$$

and therefore

$$\begin{aligned}
&\mathbb{E}_{s_0\sim\rho}[V^\pi(s_0) - V^{\pi_0}(s_0)] \\
&\quad = \frac{1}{T(1-\gamma)}\mathbb{E}_{s\sim\rho, a\sim\pi_0(s)}\left[A^{\pi_0}(s, a)\left(\hat{Q}(s, a) - \bar{\hat{Q}}(s)\right)\right] + O(1/T^2) \\
&\quad = \frac{1}{T(1-\gamma)}\langle Q_r^{\pi_0}, \hat{Q}\rangle_A + O(1/T^2)
\end{aligned} \tag{37}$$

by definition of the advantage norm.

Let us now turn to the regularization term in G_r^π: we want to estimate

$$\mathbb{E}_{s_0\sim\rho}\,\mathbb{E}\left[\sum_{t\geqslant 0}\gamma^t K(\pi, s_t) \mid s_0, \pi\right]. \tag{38}$$

Let $p_{\pi,t}$ be the distribution of s_t under policy π and $s_0 \sim \rho_0$. Since $\pi = \mathrm{Bolt}_{\pi_0}(f)$ is $O(1/T)$-close to π_0, the associated transition matrices are also close: $P_\pi = P_{\pi_0} + O(1/T)$, and therefore $P_\pi^t = P_{\pi_0}^t + O(1/T)$. [10] So in turn, $p_{\pi,t} = p_{\pi_0,t} + O(1/T)$.

By Lemma 13, K itself is $O(1/T)$. So when computing the expectation of K under $s_t \sim p_{\pi,t}$, we can replace π with π_0 up to an $O(1/T^2)$ error:

$$\mathbb{E}_{s_t\sim p_{\pi,t}}[K(\pi, s_t)] = \mathbb{E}_{s_t\sim p_{\pi_0,t}}[K(\pi, s_t)] + O(\|K\| \, \|p_{\pi,t} - p_{\pi_0,t}\|) \tag{39}$$

$$= \mathbb{E}_{s_t\sim p_{\pi_0,t}}[K(\pi, s_t)] + O(1/T^2). \tag{40}$$

Since $s_0 \sim \rho$ and ρ is an invariant distribution of π_0, we have $p_{\pi_0,t} = \rho$. Therefore,

$$\sum_{t\geqslant 0} \gamma^t \mathbb{E}_{s_t\sim p_{\pi_0,t}}[K(\pi, s_t)] = \frac{1}{1-\gamma}\mathbb{E}_{s\sim\rho}[K(\pi, s)] \tag{41}$$

and since $\pi = \mathrm{Bolt}_{\pi_0}(\hat{Q})$, this is equal to $\frac{1}{2T(1-\gamma)}\left\|\hat{Q}\right\|_A^2 + O(1/T^2)$ by Lemma 13.

By collecting the return term and the penalty term, we have

$$G_r^\pi - \mathbb{E}_{s_0\sim\rho}[V^{\pi_0}(s_0)] = \frac{1}{T(1-\gamma)}\langle Q_r^{\pi_0}, \hat{Q}\rangle_A - \frac{1}{2T(1-\gamma)}\left\|\hat{Q}\right\|_A^2 + O(1/T^2) \tag{42}$$

$$= \frac{1}{2T(1-\gamma)}\left(\|Q_r^{\pi_0}\|_A^2 - \left\|Q_r^{\pi_0} - \hat{Q}\right\|_A^2\right) + O(1/T^2) \tag{43}$$

which ends the proof since $G_r^{\pi_0} = \mathbb{E}_{s_0\sim\rho}[V^{\pi_0}(s_0)]$. □

For the proof of Theorem 8 we need a preliminary result on the reward models.

Proposition 14 (Second moment of the reward in the models). *For the random reward models of Section 3, the second moment $\mathbb{E}[rr^\top]$ satisfies:*

- *For the random Gaussian reward and random goal-reaching reward,*

$$\mathbb{E}[rr^\top] = \grave{\rho}^{-1}. \tag{44}$$

- *For the scattered random reward model with intensity κ, mean μ and variance σ^2,*

$$\mathbb{E}[rr^\top] = \kappa(\mu^2 + \sigma^2)\grave{\rho}^{-1} + (\kappa\mu)^2 \mathbb{1}\mathbb{1}^\top \tag{45}$$

 where $\mathbb{1}$ is the constant vector with components equal to 1.

[10] The constant in O is not uniform in t. It grows like t, since $P_1^t - P_2^t = \sum_{i=0}^{t-1} P_1^i(P_1 - P_2)P_2^{t-1-i}$, and P_1 and P_2 are stochastic matrices so are non-expanding in sup norm, so each term is $O(P_1 - P_2)$. So we have $P_1^t - P_2^t = O(t(P_1 - P_2))$ with uniform constants. The t factor is absorbed by γ^t in the cumulated return.

Proof of Proposition 14.
For the random Gaussian reward, the model is $\propto \exp(-\|r\|^2_{L^2(\rho)}/2) = \exp(-r^\top \grave{\rho} r/2)$ so the covariance matrix is $\grave{\rho}^{-1}$ by construction.

For random goal-reaching, we first sample a state-action $(s^\star, a^\star) \sim \rho$ then set the reward to $r = \mathbb{1}_{(s^\star,a^\star)}/\rho(s^\star, a^\star)$. Therefore, the expectation of $rr^\top$ is

$$\mathbb{E}[rr^\top] = \sum_{(s^\star,a^\star)} \rho(s^\star, a^\star) \frac{\mathbb{1}_{(s^\star,a^\star)} \mathbb{1}^\top_{(s^\star,a^\star)}}{\rho(s^\star, a^\star)^2} = \sum_{(s^\star,a^\star)} \frac{1}{\rho(s^\star, a^\star)} \mathbb{1}_{(s^\star,a^\star)} \mathbb{1}^\top_{(s^\star,a^\star)} = \grave{\rho}^{-1}. \tag{46}$$

Scattered random reward require more computation. We first sample an integer $N \sim \text{Poisson}(\kappa)$, then sample N state-actions $(s_i, a_i) \sim \rho$ and weights $w_i \sim p_w$, $i = 1, \ldots, N$, then set the reward to

$$r = \sum_{i=1}^{N} \frac{w_i}{\rho(s_i, a_i)} \mathbb{1}_{(s_i,a_i)}. \tag{47}$$

Therefore,

$$\mathbb{E}[rr^\top] = \mathbb{E}\left[\sum_{i=1}^{N} \frac{w_i^2}{\rho(s_i, a_i)^2} \mathbb{1}_{(s_i,a_i)} \mathbb{1}^\top_{(s_i,a_i)} + \sum_{i=1}^{N} \frac{w_i}{\rho(s_i, a_i)} \mathbb{1}_{(s_i,a_i)} \sum_{j \neq i} \frac{w_j}{\rho(s_j, a_j)} \mathbb{1}^\top_{(s_j,a_j)}\right] \tag{48}$$

Let us consider the first term. For each i the expectation is the same, and moreover the sampling of the weights w_i is independent from the sampling of the state-actions, so

$$\mathbb{E}\left[\sum_{i=1}^{N} \frac{w_i^2}{\rho(s_i, a_i)^2} \mathbb{1}_{(s_i,a_i)} \mathbb{1}^\top_{(s_i,a_i)}\right] = (\mathbb{E}[N])\,(\mathbb{E}[w_1^2])\,\mathbb{E}\left[\frac{1}{\rho(s_1, a_1)^2} \mathbb{1}_{(s_1,a_1)} \mathbb{1}^\top_{(s_1,a_1)}\right] \tag{49}$$

$$= \kappa(\mu^2 + \sigma^2)\grave{\rho}^{-1} \tag{50}$$

because the expectation of $N \sim \text{Poisson}(\kappa)$ is κ, and because $\mathbb{E}\left[\frac{1}{\rho(s_1,a_1)^2} \mathbb{1}_{(s_1,a_1)} \mathbb{1}^\top_{(s_1,a_1)}\right]$ is computed exactly as in the random goal-reaching reward case.

For the second term, each j term is independent from the i term. For each j term, we have

$$\mathbb{E}\left[\frac{w_j}{\rho(s_j, a_j)} \mathbb{1}^\top_{(s_j,a_j)}\right] = \mathbb{E}[w_j]\,\mathbb{E}\left[\frac{1}{\rho(s_j, a_j)} \mathbb{1}^\top_{(s_j,a_j)}\right] \tag{51}$$

$$= \mu \sum_{(s,a)} \rho(s, a) \frac{1}{\rho(s, a)} \mathbb{1}^\top_{(s,a)} \tag{52}$$

$$= \mu \mathbb{1}^\top \tag{53}$$

since the probability to sample (s, a) is $\rho(s, a)$, and since the weights w_j are sampled independently from the state-actions.

The same computation applies to the i term, which is independent from the j term. So conditionally to N, each pair (i, j) contributes $\mu^2 \mathbb{1}\mathbb{1}^\top$ to the expectation. Since there are $N(N-1)$ pairs (i, j), we have

$$\mathbb{E}\left[\sum_{i=1}^{N} \frac{w_i}{\rho(s_i, a_i)} \mathbb{1}_{(s_i,a_i)} \sum_{j\neq i} \frac{w_j}{\rho(s_j, a_j)} \mathbb{1}_{(s_j,a_j)}^\top\right] = \mathbb{E}[N(N-1)]\, \mu^2 \mathbb{1}\mathbb{1}^\top. \quad (54)$$

Finally, for a Poisson process with parameter κ, we have $\mathbb{E}[N(N-1)] = \mathbb{E}[N^2] - \mathbb{E}[N] = \mathrm{Var}[N] + (\mathbb{E}[N])^2 - \mathbb{E}[N] = \kappa^2$ since the expectation and variance of the Poisson distribution are both κ. This ends the proof. □

The following lemma expresses that constant rewards have zero advantages.

Lemma 15. *For any policy π, one has $\mathcal{A}_\pi \mathbb{1} = 0$ where $\mathbb{1}$ is the constant vector with components 1.*

Proof. By definition, for any reward function r, one has

$$r^\top \mathcal{A}_\pi r = \|Q_r^\pi\|_A^2 = \mathbb{E}_{(s,a)\sim\rho}\left[(Q_r^\pi(s, a) - \mathbb{E}_{a'\sim\pi_0(s)} Q_r^\pi(s, a'))^2\right] \quad (55)$$

and the polar form of this quadratic form is the correlation between advantages,

$$r_1^\top \mathcal{A}_\pi r_2 = \mathbb{E}_{(s,a)\sim\rho}[(Q_{r_1}^\pi(s, a) - \mathbb{E}_{a'\sim\pi_0(s)} Q_{r_1}^\pi(s, a'))(Q_{r_2}^\pi(s, a) - \mathbb{E}_{a'\sim\pi_0(s)} Q_{r_2}^\pi(s, a'))]. \quad (56)$$

When $r_2 = \mathbb{1}$, one has $Q_{r_2}^\pi = \frac{1}{1-\gamma}\mathbb{1}$ for any policy π. Therefore, $Q_{r_2}^\pi(s, a) - \mathbb{E}_{a'\sim\pi_0(s)} Q_{r_2}^\pi(s, a') = 0$ and $r_1^\top \mathcal{A}_\pi r_2 = 0$ for any r_1, so that $\mathcal{A}_\pi r_2 = 0$. □

Proof of Theorem 8.

By Theorem 2, for a reward r, the expected return using policy $\mathrm{Bolt}_{\pi_0}(\hat{Q})$ is

$$G_r^\pi = G_r^{\pi_0} + \frac{1}{2T(1-\gamma)}\left(\|Q_r^{\pi_0}\|_A^2 - \left\|\hat{Q} - Q_r^{\pi_0}\right\|_A^2\right) + O(1/T^2). \quad (57)$$

and we want to compute the expectation of this when the reward follows one of the models in Section 3, and $\hat{Q}$ is the Q-function estimated by regularized successor features.

To obtain Theorem 8, we must compute the expectation over r of the $\frac{1}{2T(1-\gamma)}$ term.

Let $\hat{r}$ be the $L^2(\rho)$-orthogonal projection of the reward r onto the features φ. By Definition 4, successor features estimate $\hat{Q}$ as the Q-function of the estimated reward $\hat{r}$ for policy π_0, namely,

$$\hat{Q} = Q_{\hat{r}}^{\pi_0} \quad (58)$$

and therefore

$$\hat{Q} - Q_r^{\pi_0} = Q_{\hat{r}}^{\pi_0} - Q_r^{\pi_0} = Q_{\hat{r}-r}^{\pi_0} \tag{59}$$

since Q-functions are linear in the reward for a fixed policy.

By definition of the advantage kernel $\mathcal{A}$, the advantage norm of the Q-function of reward r is $r^\top \mathcal{A} r$, so

$$\|Q_r^{\pi_0}\|_A^2 = r^\top \mathcal{A}_{\pi_0} r = \mathrm{Tr}(\mathcal{A}_{\pi_0} r r^\top) \tag{60}$$

and

$$\left\|\hat{Q} - Q_r^{\pi_0}\right\|_A^2 = \left\|Q_{\hat{r}-r}^{\pi_0}\right\|_A^2 = (r-\hat{r})^\top \mathcal{A}_{\pi_0}(r-\hat{r}) = \mathrm{Tr}(\mathcal{A}_{\pi_0}(r-\hat{r})(r-\hat{r})^\top). \tag{61}$$

By definition of $\hat{r}$, we have $r - \hat{r} = (\mathrm{Id} - \Pi)r$ where Π is the $L^2(\rho)$-orthogonal projector onto the features. So

$$\mathbb{E}_r\left[\|Q_r^{\pi_0}\|_A^2 - \left\|\hat{Q} - Q_r^{\pi_0}\right\|_A^2\right] = \mathrm{Tr}(\mathcal{A}_{\pi_0}\mathbb{E}_r[rr^\top]) \\ - \mathrm{Tr}\left(\mathcal{A}_{\pi_0}(\mathrm{Id} - \Pi)\mathbb{E}_r[rr^\top](\mathrm{Id} - \Pi)^\top\right) \tag{62}$$

Since the features φ are $L^2(\rho)$-orthonormal, this projector is given by $\Pi r = \varphi w$ where $w = \mathbb{E}_{(s,a)\sim\rho}[r(s,a)\varphi(s,a)] = \varphi^\top \grave{\rho} r$ is the linear regression vector of r onto the features. Therefore the projector Π satisfies $\Pi r = \varphi\varphi^\top \grave{\rho} r$ so

$$\Pi = \varphi\varphi^\top \grave{\rho}. \tag{63}$$

Let us first consider the case of random Gaussian rewards and random goal-reaching rewards. In both cases, by Proposition 14 we have

$$\mathbb{E}[rr^\top] = \grave{\rho}^{-1} \tag{64}$$

so in that case

$$\mathbb{E}_r\left[\|Q_r^{\pi_0}\|_A^2 - \left\|\hat{Q} - Q_r^{\pi_0}\right\|_A^2\right] = \mathrm{Tr}(\mathcal{A}_{\pi_0}\grave{\rho}^{-1}) \\ - \mathrm{Tr}\left(\mathcal{A}_{\pi_0}(\mathrm{Id} - \Pi)\grave{\rho}^{-1}(\mathrm{Id} - \Pi)^\top\right). \tag{65}$$

Since $\mathrm{Id} - \Pi$ is an $L^2(\rho)$-orthogonal projector, one has

$$(\mathrm{Id} - \Pi)\grave{\rho}^{-1}(\mathrm{Id} - \Pi)^\top = (\mathrm{Id} - \Pi)\grave{\rho}^{-1} \tag{66}$$

which one can also check by a direction computation using $\Pi = \varphi\varphi^\top\grave{\rho}$ and $L^2(\rho)$-orthonormality of features ($\varphi^\top\grave{\rho}\varphi = \mathrm{Id}$). Therefore,

$$\mathbb{E}_r\left[\|Q_r^{\pi_0}\|_A^2 - \left\|\hat{Q} - Q_r^{\pi_0}\right\|_A^2\right] = \mathrm{Tr}(\mathcal{A}_{\pi_0}\grave{\rho}^{-1}) - \mathrm{Tr}\left(\mathcal{A}_{\pi_0}(\mathrm{Id} - \Pi)\grave{\rho}^{-1}\right) \tag{67}$$

$$= \mathrm{Tr}(\mathcal{A}_{\pi_0}\Pi\grave{\rho}^{-1}) \tag{68}$$

$$= \mathrm{Tr}(\mathcal{A}_{\pi_0}\varphi\varphi^\top) \tag{69}$$

$$= \mathrm{Tr}(\varphi^\top\mathcal{A}_{\pi_0}\varphi). \tag{70}$$

This yields the expression of the expected gain in Theorem 8 for the case of random Gaussian or goal-reaching rewards.

For random scattered rewards, the computation has one more term. Let us start with (62). By Proposition 14,

$$\mathbb{E}[rr^\top] = \kappa(\mu^2 + \sigma^2)\grave{\rho}^{-1} + (\kappa\mu)^2 \mathbb{1}\mathbb{1}^\top. \tag{71}$$

The first term is proportional to $\grave{\rho}^{-1}$, so its contribution to (62) is the same as for random Gaussian rewards, up to the additional factor $\kappa(\mu^2+\sigma^2)$. The contribution of the second term $(\kappa\mu)^2\mathbb{1}\mathbb{1}^\top$ to (62) is $(\kappa\mu)^2$ times

$$\begin{aligned}\operatorname{Tr}(\mathcal{A}_{\pi_0}\mathbb{1}\mathbb{1}^\top) - \operatorname{Tr}(\mathcal{A}_{\pi_0}(\operatorname{Id}-\Pi)\mathbb{1}\mathbb{1}^\top(\operatorname{Id}-\Pi^\top)) \\ = \mathbb{1}^\top\mathcal{A}_{\pi_0}\mathbb{1} - ((\operatorname{Id}-\Pi)\mathbb{1})^\top\mathcal{A}_{\pi_0}(\operatorname{Id}-\Pi)\mathbb{1} \\ = 2(\Pi\mathbb{1})^\top\mathcal{A}_{\pi_0}\mathbb{1} - (\Pi\mathbb{1})^\top\mathcal{A}_{\pi_0}\Pi\mathbb{1}\end{aligned} \tag{72}$$

Now, by Lemma 15, one has $\mathcal{A}_{\pi_0}\mathbb{1} = 0$. Therefore, the above reduces to $-(\Pi\mathbb{1})^\top\mathcal{A}_{\pi_0}\Pi\mathbb{1}$, which is the term $\varphi_{\mathrm{cst}}^\top\mathcal{A}_{\pi_0}\varphi_{\mathrm{cst}}$ in Theorem 8, by definition of φ_{cst}. This ends the proof. □

Proof of Corollary 9.

The Poincaré separation theorem states that, given a symmetric matrix A and a rank-d orthogonal projector Π, the i-th largest eigenvalue of $\Pi A \Pi$ is at most the i-th largest eigenvalue of A. Consequently, the trace of $\Pi A\Pi$ is at most the sum of the largest d eigenvalues of A, which is achieved when Π coincides with the largest d eigendirections of A.

For our case, let us perform a change of basis $\varphi = \grave{\rho}^{-1/2}\tilde{\varphi}$: the $L^2(\rho)$-orthonormality of φ is equivalent to Euclidean orthonormality of $\tilde{\varphi}$. Then $\tilde{\Pi} := \tilde{\varphi}\tilde{\varphi}^\top$ is the Euclidean orthogonal projector onto $\tilde{\varphi}$. We can then rewrite

$$\begin{aligned}\operatorname{Tr}(\varphi^\top\mathcal{A}_{\pi_0}\varphi) &= \operatorname{Tr}(\tilde{\varphi}^\top\grave{\rho}^{-1/2}\mathcal{A}_{\pi_0}\grave{\rho}^{-1/2}\tilde{\varphi}) && (73)\\ &= \operatorname{Tr}(\grave{\rho}^{-1/2}\mathcal{A}_{\pi_0}\grave{\rho}^{-1/2}\tilde{\varphi}\tilde{\varphi}^\top) && (74)\\ &= \operatorname{Tr}(\grave{\rho}^{-1/2}\mathcal{A}_{\pi_0}\grave{\rho}^{-1/2}\tilde{\Pi}) && (75)\\ &= \operatorname{Tr}(\grave{\rho}^{-1/2}\mathcal{A}_{\pi_0}\grave{\rho}^{-1/2}\tilde{\Pi}^2) && (76)\\ &= \operatorname{Tr}(\tilde{\Pi}\grave{\rho}^{-1/2}\mathcal{A}_{\pi_0}\grave{\rho}^{-1/2}\tilde{\Pi}). && (77)\end{aligned}$$

Therefore, by the Poincaré separation theorem, this is maximal when $\tilde{\varphi}$ are the largest d eigenvectors of $\grave{\rho}^{-1/2}\mathcal{A}_{\pi_0}\grave{\rho}^{-1/2}$, or equivalently, when φ are the largest d eigenvectors of $\grave{\rho}^{-1}\mathcal{A}_{\pi_0}$.

Note that $\grave{\rho}^{-1}\mathcal{A}_{\pi_0}$ is self-adjoing in $L^2(\rho)$, because $\mathcal{A}_{\pi_0}$ is symmetric. Therefore, the largest d eigenvectors of $\grave{\rho}^{-1}\mathcal{A}_{\pi_0}$ are also the extrema of $\langle r, \grave{\rho}^{-1}\mathcal{A}_{\pi_0}r\rangle_{L^2(\rho)}/\left\|r\right\|^2_{L^2(\rho)}$, and since $\langle r, \grave{\rho}^{-1}\mathcal{A}_{\pi_0}r\rangle_{L^2(\rho)} = r^\top\mathcal{A}_{\pi_0}r$, they are also the extrema of $r^\top\mathcal{A}_{\pi_0}r/\left\|r\right\|^2_{L^2(\rho)}$. This proves the claim for the case of random Gaussian and random goal-reaching rewards.

For the case of random scattered rewards, on top of the $\mathrm{Tr}(\varphi^\top \mathcal{A}_{\pi_0}\varphi)$ term, there is an additional term $-(\Pi\mathbb{1})^\top \mathcal{A}_{\pi_0}\Pi\mathbb{1} \leqslant 0$. So, a priori, the maximum is lower due to this term, and might be obtained with a different choice of φ. However, since $\mathcal{A}_{\pi_0}\mathbb{1} = 0$, the main eigendirections above do *not* include the constants, and are $L^2(\rho)$-orthogonal to the constants. Therefore, if we set φ to those eigendirections, then $\Pi\mathbb{1} = 0$ so the additional term is 0. Namely, the choice of φ that maximizes the $\mathrm{Tr}(\varphi^\top \mathcal{A}_{\pi_0}\varphi)$ term *also* sets the extra negative term to 0, so the maximum is still given by those eigendirections.

This ends the proof of Theorem 8. □

We now turn to the proof of Theorem 10. The first proposition relates three norms in an MDP: the advantage norm, the Dirichlet form $\langle f, \Delta f\rangle$, and the $L^2(\rho)$ norm.

Proposition 16. *In a deterministic environment, for any function f on state-actions, and any decay factor $0 \leqslant \gamma \leqslant 1$,*

$$\|f\|_A^2 = \|f\|_{L^2(\rho)}^2 - \|P_{\pi_0} f\|_{L^2(\rho)}^2 \tag{78}$$

and for $0 < \gamma \leqslant 1$ this is further equal to

$$\|f\|_A^2 = \frac{1}{\gamma^2}\left(2\langle f, \Delta f\rangle_{L^2(\rho)} - \|\Delta f\|_{L^2(\rho)}^2 - (1-\gamma^2)\,\|f\|_{L^2(\rho)}^2\right) \tag{79}$$

where $\Delta = \mathrm{Id} - \gamma P_{\pi_0}$ is the Laplacian operator of π_0.

Proof.

Let us denote by $\bar\rho$ the row vector of size S with components given by $\bar\rho_s := \rho(s)$, the marginal probability of state s under ρ. Let $f^{\cdot 2}$ denote the pointwise square applied to a vector f. Using this notation, we have

$$\|f\|_A^2 = \mathbb{E}_{s\sim\rho,\, a\sim\pi_0(s)}(f(s,a) - \mathbb{E}_{a'\sim\pi_0(s)} f(s,a'))^2 \tag{80}$$

$$= \mathbb{E}_{s\sim\rho,\, a\sim\pi_0(s)} f(s,a)^2 - \mathbb{E}_{s\sim\rho}(\mathbb{E}_{a\sim\pi_0(s)} f(s,a))^2 \tag{81}$$

$$= \bar\rho\pi_0 f^{\cdot 2} - \bar\rho(\pi_0 f)^{\cdot 2} \tag{82}$$

where we view π_0 as an $S \times (S \times A)$ matrix, as explained in the Notation.

Since ρ is the invariant distribution of π_0, we have $\bar\rho = \bar\rho\pi_0 P$. Therefore,

$$\|f\|_A^2 = \bar\rho\pi_0 f^{\cdot 2} - \bar\rho\pi_0 P(\pi_0 f)^{\cdot 2}. \tag{83}$$

An environment is deterministic if and only if the variance of $g(s_{t+1})$ knowing (s_t, a_t) is 0 for any function g. In other words, the environment is deterministic if and only if $Pg^{\cdot 2} - (Pg)^{\cdot 2} = 0$ for any g. Therefore, in a deterministic environment,

$$\bar\rho\pi_0 P(\pi_0 f)^{\cdot 2} = \bar\rho\pi_0(P\pi_0 f)^{\cdot 2} = \bar\rho\pi_0(P_{\pi_0} f)^{\cdot 2}. \tag{84}$$

Finally, for any function g, $\bar{\rho}\pi_0 g^{\cdot 2}$ is just another notation for $\|g\|^2_{L^2(\rho)}$. Therefore, we find

$$\|f\|_A^2 = \bar{\rho}\pi_0 f^{\cdot 2} - \bar{\rho}\pi_0 (P_{\pi_0} f)^{\cdot 2} = \|f\|^2_{L^2(\rho)} - \|P_{\pi_0} f\|^2_{L^2(\rho)} \tag{85}$$

as needed. The second statement follows from substituting $P_{\pi_0} = \frac{1}{\gamma}(\mathrm{Id} - \Delta)$ and expanding. □

Proof of Theorem 10.
Let $r \in L^2_0(\rho)$. Since r averages to 0, its Q-function $Q_r^{\pi_0}$ is well-defined for $\gamma = 1$. By definition of the advantage kernel, we have

$$r^\top \mathcal{A}_{\pi_0} r = \|A_r^{\pi_0}\|^2_{L^2(\rho)} = \|Q_r^{\pi_0}\|_A^2 \tag{86}$$

by definition of $\|\cdot\|_A$.

By Proposition 16 for $\gamma = 1$, we have

$$\|Q_r^{\pi_0}\|_A^2 = 2\langle Q_r^{\pi_0}, \Delta Q_r^{\pi_0}\rangle_{L^2(\rho)} - \|\Delta Q_r^{\pi_0}\|^2_{L^2(\rho)}. \tag{87}$$

Now, $Q_r^{\pi_0}$ satisfies the Bellman equation $Q_r^{\pi_0} = r + P_{\pi_0} Q_r^{\pi_0}$ for $\gamma = 1$. By definition of Δ, this rewrites as

$$\Delta Q_r^{\pi_0} = r. \tag{88}$$

Therefore,

$$\begin{aligned}
\|Q_r^{\pi_0}\|_A^2 &= 2\langle Q_r^{\pi_0}, r\rangle_{L^2(\rho)} - \|r\|^2_{L^2(\rho)} \\
&= 2\langle \Delta^{-1} r, r\rangle_{L^2(\rho)} - \|r\|^2_{L^2(\rho)} \\
&= \langle \Delta^{-1} r, r\rangle_{L^2(\rho)} + \langle r, (\Delta^{-1})^* r\rangle_{L^2(\rho)} - \|r\|^2_{L^2(\rho)} \\
&= \langle r, (\Delta^{-1} + (\Delta^{-1})^* - \mathrm{Id}) r\rangle_{L^2(\rho)}
\end{aligned}$$

as needed.

Therefore, the functions $r \in L^2_0(\rho)$ that are the extrema of $r^\top \mathcal{A}_{\pi_0} r / \|r\|^2_{L^2(\rho)}$ are the largest eigenfunctions of $\Delta^{-1} + (\Delta^{-1})^* - \mathrm{Id}$, or equivalently the largest eigenfunctions of $\Delta^{-1} + (\Delta^{-1})^*$. □

Proof of Proposition 11.
Let $r \in L^2(\rho)$. By definition of the advantage kernel, we have

$$r^\top \mathcal{A}_{\pi_0} r = \|A_r^{\pi_0}\|^2_{L^2(\rho)} = \|Q_r^{\pi_0}\|_A^2 \tag{89}$$

by definition of $\|\cdot\|_A$.

But for $\gamma = 0$ we have $Q_r^{\pi_0} = r$, so

$$r^\top \mathcal{A}_{\pi_0} r = \|r\|_A^2. \tag{90}$$

By Proposition 16 for $\gamma = 0$, we have

$$\|r\|_A^2 = \|r\|_{L^2(\rho)}^2 - \|P_{\pi_0} r\|_{L^2(\rho)}^2 \tag{91}$$

$$= \langle r, r\rangle_{L^2(\rho)} - \langle P_{\pi_0} r, P_{\pi_0} r\rangle_{L^2(\rho)} \tag{92}$$

$$= \langle r, (\mathrm{Id} - P_{\pi_0}^* P_{\pi_0}) r\rangle_{L^2(\rho)} \tag{93}$$

as needed.

Therefore, the extrema of $r^\top \mathcal{A}_{\pi_0} r / \|r\|_{L^2(\rho)}^2$ are the largest eigenfunctions of $\mathrm{Id} - P_{\pi_0}^* P_{\pi_0}$, namely, the smallest eigenfunctions of $P_{\pi_0}^* P_{\pi_0}$. □

Proof of Theorem 12.
Let $r \in L^2(\rho)$. By definition of the advantage kernel, we have

$$r^\top \mathcal{A}_{\pi_0} r = \|A_r^{\pi_0}\|_{L^2(\rho)}^2 = \|Q_r^{\pi_0}\|_A^2 \tag{94}$$

by definition of $\|\cdot\|_A$.

By Proposition 16 for $0 < \gamma < 1$, we have

$$\gamma^2 \|Q_r^{\pi_0}\|_A^2 = 2\langle Q_r^{\pi_0}, \Delta Q_r^{\pi_0}\rangle_{L^2(\rho)} - \|\Delta Q_r^{\pi_0}\|_{L^2(\rho)}^2 - (1-\gamma^2)\|Q_r^{\pi_0}\|_{L^2(\rho)}^2 . \tag{95}$$

Now, $Q_r^{\pi_0}$ satisfies the Bellman equation $Q_r^{\pi_0} = r + \gamma P_{\pi_0} Q_r^{\pi_0}$. By definition of Δ, this rewrites as

$$\Delta Q_r^{\pi_0} = r, \qquad Q_r^{\pi_0} = \Delta^{-1} r. \tag{96}$$

Therefore,

$$\begin{aligned}
\gamma^2 \|Q_r^{\pi_0}\|_A^2 &= 2\langle Q_r^{\pi_0}, r\rangle_{L^2(\rho)} - \|r\|_{L^2(\rho)}^2 - (1-\gamma^2)\|Q_r^{\pi_0}\|_{L^2(\rho)}^2 \\
&= 2\langle \Delta^{-1} r, r\rangle_{L^2(\rho)} - \|r\|_{L^2(\rho)}^2 - (1-\gamma^2)\left\|\Delta^{-1} r\right\|_{L^2(\rho)}^2 \\
&= \langle r, \left(\Delta^{-1} + (\Delta^{-1})^* - \mathrm{Id} - (1-\gamma^2)(\Delta^{-1})^*\Delta^{-1}\right) r\rangle_{L^2(\rho)}
\end{aligned}$$

as needed.

Therefore, the extrema of $r^\top \mathcal{A}_{\pi_0} r / \|r\|_{L^2(\rho)}^2$ are the largest eigenfunctions of $\Delta^{-1} + (\Delta^{-1})^* - \mathrm{Id} - (1-\gamma^2)(\Delta^{-1})^*\Delta^{-1}$, or equivalently the largest eigenfunctions of $\Delta^{-1} + (\Delta^{-1})^* - (1-\gamma^2)(\Delta^{-1})^*\Delta^{-1}$.

Note that $\Delta^{-1} + (\Delta^{-1})^* - \mathrm{Id} - (1-\gamma^2)(\Delta^{-1})^*\Delta^{-1}$ is a non-negative operator, since $\|Q_r^{\pi_0}\|_A^2 \geqslant 0$.

Alternatively, one can start with the first expression in Proposition 16, namely $\|Q_r^{\pi_0}\|_A^2 = \|Q_r^{\pi_0}\|_{L^2(\rho)}^2 - \|P_{\pi_0} Q_r^{\pi_0}\|_{L^2(\rho)}^2$. By a similar derivation, this leads to a slightly different expression for the same quantities, mixing Δ and P_{π_0}:

$$r^\top \mathcal{A}_{\pi_0} r = \langle r, (\Delta^{-1})^* (\mathrm{Id} - P_{\pi_0}^* P_{\pi_0}) \Delta^{-1} r\rangle_{L^2(\rho)}. \tag{97}$$

□

Proof of Proposition 6.
By definition, successor features estimate $\hat{Q}$ as the Q-function of the estimated reward $\hat{r}$,

$$\hat{Q} = \hat{r} + \gamma P_{\pi_0} \hat{Q} \tag{98}$$

where

$$\hat{r}(s,a) = z^\top \varphi(s,a) \tag{99}$$

is the $L^2(\rho)$-orthogonal projection of r onto the features φ, with z given by (16). Therefore

$$\left\| \hat{Q} - r - \gamma P_{\pi_0} \hat{Q} \right\|^2_{L^2(\rho)} = \|\hat{r} - r\|^2_{L^2(\rho)} = \|(\mathrm{Id} - \Pi) r\|^2_{L^2(\rho)} \tag{100}$$

where Π is the $L^2(\rho)$-orthogonal projector onto the features.

Therefore, we have

$$\begin{aligned}
\mathbb{E}_r \left\| \hat{Q} - r - \gamma P_{\pi_0} \hat{Q} \right\|^2_{L^2(\rho)} &= \mathbb{E}_r \|(\mathrm{Id} - \Pi) r\|^2_{L^2(\rho)} && (101)\\
&= \mathbb{E}_r \left[\langle (\mathrm{Id} - \Pi) r, (\mathrm{Id} - \Pi) r \rangle_{L^2(\rho)} \right] && (102)\\
&= \mathbb{E}_r \left[\langle r, (\mathrm{Id} - \Pi) r \rangle_{L^2(\rho)} \right] && (103)\\
&= \mathbb{E}_r \left[r^\top \grave{\rho} (\mathrm{Id} - \Pi) r \right] && (104)\\
&= \mathbb{E}_r \operatorname{Tr} \left(r r^\top \grave{\rho} (\mathrm{Id} - \Pi) \right) && (105)\\
&= \operatorname{Tr} \left(\mathbb{E}_r [r r^\top] \grave{\rho} (\mathrm{Id} - \Pi) \right). && (106)
\end{aligned}$$

By Proposition 14, we have $\mathbb{E}_r[r r^\top] = \grave{\rho}^{-1}$ in the random Gaussian reward and random goal-reaching reward models. Therefore the above is

$$\begin{aligned}
\operatorname{Tr} \left(\mathbb{E}_r [r r^\top] \grave{\rho} (\mathrm{Id} - \Pi) \right) &= \operatorname{Tr} (\mathrm{Id} - \Pi) && (107)\\
&= \#S \times \#A - d && (108)
\end{aligned}$$

since Π is a projector of rank d. This proves the result.

The same proof works for the universal successor features as in [BBQ+18], where $\hat{Q}$ satisfies a Bellman equation with respect to π_z instead of π_0. □

References

[ALG+21] Majid Abdolshah, Hung Le, Thommen Karimpanal George, Sunil Gupta, Santu Rana, and Svetha Venkatesh. A new representation of successor features for transfer across dissimilar environments. In *International Conference on Machine Learning*, pages 1–9. PMLR, 2021.

[BBQ+18] Diana Borsa, André Barreto, John Quan, Daniel Mankowitz, Rémi Munos, Hado van Hasselt, David Silver, and Tom Schaul. Universal successor features approximators. *arXiv preprint arXiv:1812.07626*, 2018.

[BDM+17] André Barreto, Will Dabney, Rémi Munos, Jonathan J Hunt, Tom Schaul, David Silver, and Hado P van Hasselt. Successor features for transfer in reinforcement learning. In *NIPS*, 2017.

[BO21] Léonard Blier and Yann Ollivier. Unbiased methods for multi-goal reinforcement learning. *arXiv preprint arXiv:2106.08863*, 2021.

[BTO21] Léonard Blier, Corentin Tallec, and Yann Ollivier. Learning successor states and goal-dependent values: A mathematical viewpoint. *arXiv preprint arXiv:2101.07123*, 2021.

[EGIL18] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. In *International Conference on Learning Representations*, 2018.

[GB20] Dibya Ghosh and Marc G Bellemare. Representations for stable off-policy reinforcement learning. In *International Conference on Machine Learning*, pages 3556–3565. PMLR, 2020.

[GHB+19] Christopher Grimm, Irina Higgins, Andre Barreto, Denis Teplyashin, Markus Wulfmeier, Tim Hertweck, Raia Hadsell, and Satinder Singh. Disentangled cumulants help successor representations transfer to new tasks. *arXiv preprint arXiv:1911.10866*, 2019.

[GRW16] Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control. *arXiv preprint arXiv:1611.07507*, 2016.

[HDB+19] Steven Hansen, Will Dabney, Andre Barreto, Tom Van de Wiele, David Warde-Farley, and Volodymyr Mnih. Fast task inference with variational intrinsic successor features. *arXiv preprint arXiv:1906.05030*, 2019.

[LA21] Hao Liu and Pieter Abbeel. APS: Active pretraining with successor features. In *International Conference on Machine Learning*, pages 6736–6747. PMLR, 2021.

[LKTF20] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

[MM07] Sridhar Mahadevan and Mauro Maggioni. Proto-value functions: A Laplacian framework for learning representation and control in Markov decision processes. *Journal of Machine Learning Research*, 8(10), 2007.

[Oll18] Yann Ollivier. Approximate temporal difference learning is a gradient descent for reversible policies. *arXiv preprint arXiv:1805.00869*, 2018.

[RZL+22] Tongzheng Ren, Tianjun Zhang, Lisa Lee, Joseph E Gonzalez, Dale Schuurmans, and Bo Dai. Spectral decomposition representation for reinforcement learning. *arXiv preprint arXiv:2208.09515*, 2022.

[SB18] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018. 2nd edition.

[SLA+15] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

[SWD+17] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[TO21] Ahmed Touati and Yann Ollivier. Learning one representation to optimize all rewards. *Advances in Neural Information Processing Systems*, 34, 2021.

[TRO23] Ahmed Touati, Jérémy Rapin, and Yann Ollivier. Does zero-shot reinforcement learning exist? In *ICLR*. OpenReview.net, 2023.

[TVR97] John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.

[WTN18] Yifan Wu, George Tucker, and Ofir Nachum. The Laplacian in RL: Learning representations with efficient approximations. In *International Conference on Learning Representations*, 2018.

[ZSBB17] Jingwei Zhang, Jost Tobias Springenberg, Joschka Boedecker, and Wolfram Burgard. Deep reinforcement learning with successor features for navigation across similar environments. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2371–2378. IEEE, 2017.