Per-Decision Option Discounting

Anna Harutyunyan [1 2] **Peter Vrancx** [3 2] **Philippe Hamel** [1] **Ann Nowé** [2] **Doina Precup** [1]

Abstract

In order to solve complex problems an agent must be able to reason over a sufficiently long horizon. Temporal abstraction, commonly modeled through *options*, offers the ability to reason at many timescales, but the horizon *length* is still determined by the discount factor of the underlying Markov Decision Process. We propose a modification to the options framework that naturally scales the agent's horizon with option length. We show that the proposed option-step discount controls a bias-variance trade-off, with larger discounts (counter-intuitively) leading to less estimation variance.

1. Introduction

Reinforcement learning agents have to solve the problem of reasoning about actions that improve long-term performance. This objective can be formulated either in the discounted setting (in which the agent optimizes the expectation of the discounted sum of rewards) or in the average-reward setting (in which the agent optimizes the average reward received per step over an infinite amount of time) (Schwartz, 1993; Bertsekas & Tsitsiklis, 1996; Mahadevan, 1996). Intuitively, if agents are to have a finite lifetime (but of unknown duration), the discounted reward framework is more appropriate. Indeed, an average reward agent is content to wait arbitrarily long before collecting rewards, so long as the average over the infinite horizon is favorable. Discounting, on the other hand, causes the agent to focus on collecting rewards early on – preferred behavior if there is a chance of termination. Even in cases where the average reward performance is the desired target, it can be approximated arbitrarily well by using a discount which approaches one (Tsitsiklis & Van Roy, 2002). We will hence focus on the discounted setting in this paper.

[1]DeepMind, London, UK [2]Vrije Universiteit Brussel, Brussels, Belgium [3]PROWLER.io, Cambridge, UK. Correspondence to: Anna Harutyunyan <harutyunyan@google.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

The discount factor γ itself is usually treated as something in between a mathematical convenience and a meaningful time horizon parameter. Indeed, the discounted *return* $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$ corresponds roughly to a time horizon of $\frac{1}{1-\gamma}$ steps, and the convergence speed of learning and planning algorithms is strongly related to this quantity (Bertsekas & Tsitsiklis, 1996; Jiang et al., 2015). While we wish for our agents to reason over long temporal horizons, this coupling makes it challenging for that to be efficient, and fairly low discounts are favored in practice.

Temporal abstraction is often considered key for reasoning over long temporal horizons (Sutton et al., 1999). Indeed, options induce reward and transition models that act as higher-order analogues of those from the primitive-action Markov Decision Process. However, the discount factor remains an integral part of these models – there are in fact clear parallels between options and multi-step temporal difference methods (Bacon & Precup, 2016; Harutyunyan et al., 2018). As such, in the classical options framework, an option that takes one hundred steps incurs a discount of γ^{100} at termination, and the information about its outcome can be significantly washed out. So, regardless of how sophisticated the options are, the agent remains at the mercy of the step-discount, and its high-level plan is unlikely to benefit from the foresight brought by the options. The temporal abstraction in the *behavior* offers little additional temporal representation power in the model or the value function. The aim of this work is to overcome this limitation. In particular:

- We generalize the options framework to better express temporal abstraction. Namely, we (1) decouple the step-discounts in the reward and transition models, and (2) introduce *per-decision* discounting that augments the transition model irrespectively of duration. This simple generalization allows for options to extend the agent's horizon, a property which can be thought of as "time dilation".
- We analyze the properties of planning with the new framework, and devise novel bias-variance bounds that apply to the classical options framework as a special case. Notably, we show that *larger* discounts in the transition model can *reduce* the variance of estimating values, which is contrary to the familiar intuitions about e.g. multi-step returns.

- We verify the shape of the bounds empirically on a classical task. Our results imply that in addition to extending the agent's horizon, time dilation can be a tool for better estimation of value functions.

The paper is organized as follows. After providing background and relevant notation, in Section 3 we motivate and in Section 4 formally introduce time dilation. In Section 5 we begin our analysis by deriving the equivalent step-discounted problem and providing a convergence result. Section 6 analyzes the bias-variance trade-off of the new framework in the approximate dynamic programming setting. All proofs from these sections are in the appendix. Finally, Section 7 provides supporting experimental results.

2. Background

A Markov Decision Process (MDP) is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ the set of discrete actions; $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ represents the environment dynamics, where $p(s'|s, a)$ is the probability of transitioning to state s' when a is taken in s; $r : \mathcal{S} \times \mathcal{A} \to [-r_{\max}, r_{\max}]$ is the reward function. A policy $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$ induces a reward Markov chain given by $p^\pi(s'|s) = \sum_{a\in\mathcal{A}} \pi(a|s)p(s'|s, a)$, and $r^\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s)r(s, a)$.

The general goal of a reinforcement learning (RL) agent is to find a policy that optimizes a cumulative measure of reward (Sutton & Barto, 2017). One standard way to define such a criterion is by using a scalar discount factor $\gamma \in (0, 1]$, which depreciates rewards received further in the future. More formally, the agent will aim to optimize:

$$\mathbb{E}_\pi \left[\sum_{i=0}^{\infty} \gamma^i r^\pi(S_{t+i}) | S_t = s \right], \forall s \in \mathcal{S}$$

Let $q : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ be a generic action-value (or Q-) function and define the one-step transition operator:

$$\mathcal{P}^\pi q(s, a) \stackrel{\text{def}}{=} \sum_{s'\in\mathcal{S}} \sum_{a'\in\mathcal{A}} p(s'|s, a)\pi(a'|s')q(s', a').$$

Using operator notation, the value of policy π under a discount γ is given by: $q^\pi_\gamma \stackrel{\text{def}}{=} \sum_{t=0}^{\infty} \gamma^t(\mathcal{P}^\pi)^t r = r+\gamma\mathcal{P}^\pi q^\pi_\gamma = (I-\gamma\mathcal{P}^\pi)^{-1}r$, where the inverse always exists if $\gamma < 1$. The corresponding one-step Bellman operator (Bellman, 1957) can be applied to any q:

$$\mathcal{T}^\pi q(s, a) \stackrel{\text{def}}{=} r(s, a) + \gamma\mathcal{P}^\pi q(s, a),$$

and its repeated applications are guaranteed to produce its fixed point q^π_γ (Puterman, 1994). Policy *evaluation* is concerned with estimating this quantity for a fixed π, while in control we seek the *optimal* policy π^*_γ, whose value $q^*_\gamma = \max_\pi q^\pi_\gamma$. The *state* value function averages q^π_γ w.r.t. the policy: $v^\pi_\gamma(s) = \sum_a \pi(a|s)q^\pi_\gamma(s, a)$, and $v^*_\gamma = \max_\pi v^\pi_\gamma$.

The target of learning is often intended to be w.r.t. some very long horizon, whose corresponding discount factor we denote by γ_{eval}. Using a large discount factor can be inefficient during value function learning, and can pose problems during planning with approximate models (e.g. (Jiang et al., 2015; Lehnert et al., 2018)). Hence, in practice, γ used during learning is often treated as a parameter with the hope that it will lead to finding a good policy w.r.t. γ_{eval}. Similarly to the notions from Jiang et al. (2015) and Lehnert et al. (2018), we say that a discount γ is able to *represent* a policy w.r.t. $\gamma' > \gamma$ if $\pi^*_\gamma = \pi^*_{\gamma'}$.

The average-reward formulation aims to find the policy that optimizes the following criterion:

$$\max_\pi \mathbb{E}_\pi \left[\lim_{T\to\infty} \sum_{i=0}^{T} r^\pi(S_i) \right].$$

Tsitsiklis & Van Roy (2002) showed that both the expected and the transient behavior of the average reward criterion can be approximated arbitrarily well by a large enough γ.

An option o is a tuple $(\mathcal{I}^o, \beta^o, \pi^o)$, with $\mathcal{I}^o \subseteq \mathcal{S}$ the initiation set,[1] π^o the internal option policy, and $\beta^o : \mathcal{S} \to [0, 1]$ the termination condition, with $\beta^o(s)$ denoting the probability of option o terminating upon arriving in s (Sutton et al., 1999). Given a discount factor γ, an option o has an associated *semi*-MDP (Puterman, 1994) model given by:

$$\begin{aligned} P^o_\gamma(s'|s) &\stackrel{\text{def}}{=} \mathbb{E}_{D:s\to s'|o} \left[\gamma^D \mathbb{I}_{S_{t+D}=s'} | S_t = s\right] \\ &= \gamma p^{\pi^o}(s'|s)\beta^o(s') \\ &\quad + \gamma \sum_{s''} p^{\pi^o}(s''|s)(1 - \beta^o(s''))P^o_\gamma(s''|s'), \\ R^o_\gamma(s) &\stackrel{\text{def}}{=} \mathbb{E}_{D:s|o} \left[\sum_{i=0}^{D-1} \gamma^i r^{\pi^o}(S_{t+i}) | S_t = s \right] \\ &= r^{\pi^o}(s) + \gamma \sum_{s'} p^{\pi^o}(s'|s)(1 - \beta^o(s'))R^o_\gamma(s'), \end{aligned}$$

where $\mathbb{E}_{D:s|o}[\cdot]$ and $\mathbb{E}_{D:s\to s'|o}[\cdot]$ are the expectations of the option o's duration D from state s and the travel time between state s and s' in which o terminates, respectively, and where we index relevant quantities with the appropriate discount factor to make the dependence clear. For a finite set of options, we denote the $|\mathcal{S}| \times |\mathcal{O}|$ matrix collecting all R^o_γ by R_γ.

In the usual call-and-return model of option execution, an option is run until completion (according to its termination condition), upon which a new option choice is made (Precup

[1] For simplicity, we assume that $\mathcal{I}^o = \mathcal{S}, \forall o \in \mathcal{O}$.

et al., 1998). This suggests the following analogues of $\mathcal{P}^\pi$ and $\mathcal{T}^\pi$ for a policy over options μ and discount factor γ:

$$\mathcal{P}^\mu_\gamma q(s,o) \stackrel{\text{def}}{=} \sum_{s'} P^o_\gamma(s'|s) \sum_{o'} \mu(o'|s') q(s',o'),$$
$$\mathcal{T}^\mu_\gamma q(s,o) \stackrel{\text{def}}{=} R^o_\gamma(s) + \mathcal{P}^\mu_\gamma q(s,o). \quad (1)$$

Finally, we will write $\|\cdot\|$ to denote the L_∞ norm.

3. Motivation

In order to further understand the goal of our approach, consider an agent placed in a deterministic environment needing to choose between a closer, worse reward of z and a farther, better reward of $Z > z$, with all other rewards being zero. The rewarding states s_z and s_Z could be terminal goal states, or could in fact allow the agent to continue on some future path. Now, consider the role of the discount factor γ in this decision. In order for the agent to pick the higher reward, it would need to be the case that $\gamma^K Z > \gamma^k z$, where K and k are the distances from s_Z and s_z to the agent's location. Thus, there is a minimum value of $\gamma > \left(\frac{z}{Z}\right)^{\frac{1}{K-k}}$ required for the agent to display foresight (see Figure 1 for an illustration). The average-reward framework can fix this problem if the environment is a continuing task. However, if the goal states are absorbing (the agent stays in these states forever once it reaches them), then in fact all policies have an average reward of 0 and there is no useful signal for the agent to optimize. This illustrates that considering average rewards does not handle the trade-off of the two subgoals correctly in all cases.

Now, consider the same task, but with the choice being between two options (instead of primitive actions), going to each of the respective goals. Ideally, scaling the primitive time step should allow the agent to consistently exhibit foresight and choose the larger reward (since we can assume the primitive time step was a somewhat arbitrary choice anyway). However, the agent will only keep its preference if $P^{o_1}_\gamma(s_Z|s_0)Z > P^{o_2}_\gamma(s_z|s_0)z$, which still entirely depends on γ.[2] Hence, the policy over options remains tied critically to the magnitude of the discount factor applied at the primitive time step.

4. Options with Time Dilation

The example from the previous section highlights the fact that the option transition model is responsible for the effective discount. This leads us to propose two modifications to the classical transition model, aimed to ensure that options provide real temporal abstraction, which diminishes or eliminates completely the dependence on the primitive action time step:

1. We allow "time dilation" in the transitions by using a transition step-discount γ^o_p that is larger than the reward step-discount γ^o_r, thus weakening the dependence on option duration.
2. We add a *per-decision* state-dependent discount γ^o_d, thus reinforcing option-level reasoning over primitive-action policies.

The inner γ^o_r has exactly the same interpretation as before, but locally for each option. But when reasoning at the higher level, the precise duration of each option matters less, and so $\gamma^o_p > \gamma^o_r$, but the number of decisions steps among options becomes relevant, and so γ^o_d need not be 1. We denote the set of option discounts by $\Gamma^o = \{\gamma^o_r, \gamma^o_p, \gamma^o_d\}$. The new option transition model is then given by:

$$P^o_\Gamma(s'|s) \stackrel{\text{def}}{=} \gamma^o_d(s,s') \mathbb{E}_{D:s\to s'|o}\left[(\gamma^o_p)^D \mathbb{I}_{S_{t+D}=s'} | S_t = s\right].$$

Clearly, $P^o_\Gamma = P^o_\gamma$ and consistent with $R^o_{\gamma_r}$ if $\gamma^o_p = \gamma^o_r = \gamma$ and $\gamma^o_d = I$. Figure 2 plots a simple instance of the coefficients induced by this manner of discounting on a fixed stream of rewards. The two-timescale "spiky" structure is due to the discrepancy between γ^o_p and γ^o_r: the returns within an option are still discounted with γ^o_r, but upon termination, γ^o_p is invoked.

We will show that γ^o_p imposes a bias-variance trade-off on the complexity of estimating the transition model and

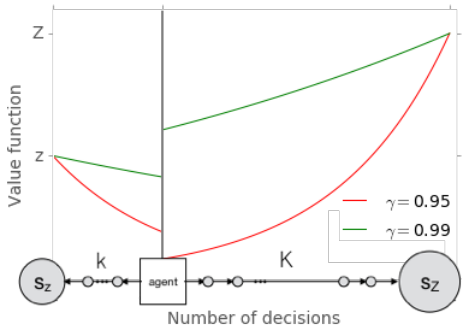

Figure 1. The value of the actions that lead towards the two goals plotted against the timestep distance towards these goals. The lines plot the values of the left and right actions in the states to the left and to the right of the agent, respectively, and the colors correspond to two different discounts. The preferred action in the current state thus depends on the discount: the red discounting scheme of $\gamma = 0.95$ is too short-sighted to prefer the correct goal s_Z. Note that for any discount $\gamma < 1$, the distances k and K can be proportionally increased (to $k + H$ and $K + H$ for some $H < \infty$) for γ to be insufficient to capture the correct ordering.

[2] To see this, note that $\gamma^{d_{\max}} \leq \|P^o_\gamma\| \leq \gamma^{d_{\min}}$, where $d_{\min}$ and $d_{\max}$ denote minimum and maximum duration of an option. If $\gamma^{d_{\max}} < \left(\frac{z}{Z}\right)^{\frac{1}{K-k}}$, the agent will flip its choice of option.

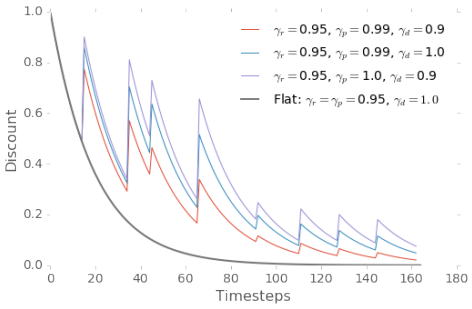

Figure 2. The discounting coefficient applied to the reward at time t under time dilation for various constant values of Γ and γ_p and for random option durations (drawn from a Poisson distribution with $\lambda = 10$). The spikes occur because the reward model (still discounted with γ) controls the discounting inside the option, and the transition model that at option decision points.

the value function, where the bias is w.r.t. a consistent discounting scheme $\gamma_p^o = \gamma_r^o$. In the extreme, if $\gamma_p^o = 1$, all of the variance of a given transition s, s' associated with the random variable D that determines the number of steps from s to s' is removed. This comes at the cost of introducing bias in terms of the difference between γ_r^o and γ_p^o. The additional option-level discount γ_d^o can help reduce this bias. In fact, we show that there is a value of γ_d^o for which the bias is zero, which occurs when γ_d^o captures P^o in a certain sense.

The new option model allows one to effectively redefine the primitive resolution of the agent, simply by considering $\gamma_p^o = 1$. This in turn provides options with the power to represent policies over horizons that would otherwise be too large to capture with a fixed step-discount. Indeed, consider the example from above but with $\gamma_p^o = 1$. Then, $\|P_\Gamma^o\| = \|\gamma_d^o\|$ becomes independent of the step-discount, and able to represent the same policy over options regardless of their length. This is a key motivation of our approach.

For notational simplicity, in the following we will take Γ^o to be the same for all options, and denote the relevant discounts simply by γ_r, γ_p and γ_d. All of the results trivially apply in the general case. Furthermore, instead of the full γ_d matrix, it is more practical to consider a diagonal γ_d that specifies an option-level discount that only depends on the arrival state s', regardless of where the option started.[3] This form of γ_d^o lets us rewrite P_Γ^o in an intuitive way:

$$P_\Gamma^o(s'|s) = \gamma_d(s')\beta^o(s')\Big(\gamma_p p^{\pi^o}(s'|s) + \gamma_p^2 \sum_{s''} p^{\pi^o}(s''|s)(1-\beta^o(s''))(p^{\pi^o}(s'|s'') + \ldots)\Big).$$

[3]The same can be achieved by considering a full row-wise constant matrix, but the diagonal form is more efficient and commonly used, see e.g. (White, 2017; Yu & Bertsekas, 2012).

This equation makes it evident that γ_d controls the discounting over the agent's decision points (i.e. where the agent chooses an option), while γ_p is the *intra-option* discount. We return to the full matrix view briefly in Section 6.2.

5. Convergence Analysis

In this section we will derive an equivalence from the model we propose to a step-discounted setting with consistent options (i.e. ones with $\gamma_p = \gamma_r$), and use it to prove expected convergence under mild conditions. Our analysis both here and in the next section is for policy evaluation given a fixed policy over options μ, but our experiments test the control setting, and illustrate the insights found in the theory. The new option operators that we will work with are given by:

$$\mathcal{P}_\Gamma^\mu q(s,o) = \sum_{s'} P_\Gamma^o(s'|s) \sum_{o'} \mu(o'|s') q(s',o'),$$
$$\mathcal{T}_\Gamma^\mu q(s,o) \stackrel{\text{def}}{=} R_{\gamma_r}^o(s) + \mathcal{P}_\Gamma^\mu q(s,o). \quad (2)$$

with $R_{\gamma_r}^o$ defined as before. We can show that $\mathcal{T}_\Gamma^\mu$ is a contraction (i.e. converges asymptotically), so long as a terminating state with a sub-unitary discount is reachable.

Assumption 5.1. *Options are finite:* $d_{\max} < \infty$.

This assumption is standard, and can be attained by stopping options after some maximum number of steps, since all option theory results hold for such semi-Markov terminations (Sutton et al., 1999). We additionally require that there is a chance for an option to terminate in a state whose discount is less than one.

Assumption 5.2. *For all* $o \in \mathcal{O}$ *and* $s \in \mathcal{S}$, $\exists s'$ *that is reachable by* π^o, *s.t.* $\beta^o(s') > 0$ *and* $\gamma_d^o(s') < 1$, *or* $\gamma_p < 1$.

The following theorem proves that $\mathcal{T}_\Gamma^\mu$ is a contraction, and derives the equivalent problem with a modified reward model, termination scheme, and a generalized step-discount.

Theorem 1. *If Assumptions 5.1 and 5.2 hold, the operator* $\mathcal{T}_\Gamma^\mu$ *from Eq. (2) is a contraction. The fixed point of* $\mathcal{T}_\Gamma^\mu$ *is equivalent to that of a* κ*-discounted options operator* $\mathcal{T}_\kappa^\mu$ *from Eq. (1) for*

$$\kappa(s,o,s') = \gamma_p(1-\beta^o(s')(1-\gamma_d(s'))) \leq \gamma_p,$$

w.r.t. modified reward and termination functions:

$$z^{\pi^o} = (I - \gamma_p p^{(1-\beta)\pi^o})(I - \gamma_r p^{(1-\beta)\pi^o})^{-1} r^{\pi^o}$$
$$\zeta^o(s') = \frac{\gamma_d(s')\beta^o(s')}{\gamma_d(s')\beta^o(s') + 1 - \beta^o(s')},$$

where $p^{(1-\beta)\pi^o}(s'|s) \stackrel{def}{=} (1-\beta^o(s'))p^{\pi^o}(s'|s)$.

The proof is in Appendix A. This theorem implies that the step discount is controlled by γ_d, γ_p, but also β^o. This

is appropriate, since β^o controls the inner timescale of an option. For example, if $\gamma_p = 1$ and $\gamma_d(s') = 0$ for some s', the discount at s' is $1 - \beta^o(s')$ exactly.

On the other hand, the value of γ_d directly impacts the new, implicitly induced termination scheme. For example, if $\gamma_d(s') = 0$ and no bootstrapping occurs, then ζ accounts for it by not permitting any termination. In general, any $\gamma_d(s') < 1$ implies $\zeta^o(s') < \beta^o(s')$, and hence the corresponding induced options are longer. This highlights the fact that our approach yields more effective temporal abstraction.

6. The Bias-Variance Tradeoff in the Option Transition Model

We will now analyze the computational effects of time dilation. We will show that a larger discount in the transition model can reduce estimation variance, at the cost of introducing bias, as compared to the consistent discounting scheme. The inter-option discount γ_d then helps control this bias, with a particular shape of γ_d removing it altogether.

More specifically, we will show that varying γ in $\mathcal{P}^\mu_\gamma$ from Eq. (1), and more generally replacing $\mathcal{P}^\mu_{\gamma_r}$ with $\mathcal{P}^\mu_\Gamma$ from Eq. (2) induces a novel bias-variance tradeoff on the approximation error when $\mathcal{P}^\mu_\Gamma$ is estimated from samples. Let $q^\mu_{\gamma_r} = (I - \mathcal{P}^\mu_{\gamma_r})^{-1} R_{\gamma_r}$, $q^\mu_\Gamma = (I - \mathcal{P}^\mu_\Gamma)^{-1} R_{\gamma_r}$, and $q^\mu_{\widehat{\Gamma}} = (I - \widehat{\mathcal{P}^\mu_\Gamma})^{-1} R_{\gamma_r}$, where $\widehat{\mathcal{P}^\mu_\Gamma}$ is the approximate transition model estimated from samples. The approximate loss has the following form:

$$\begin{aligned}\mathcal{E} &= \|q^\mu_{\widehat{\Gamma}} - q^\mu_{\gamma_r}\| = \|q^\mu_{\widehat{\Gamma}} - q^\mu_\Gamma + q^\mu_\Gamma - q^\mu_{\gamma_r}\| \\ &\leq \underbrace{\|q^\mu_{\widehat{\Gamma}} - q^\mu_\Gamma\|}_{\mathcal{E}_{estim}} + \underbrace{\|q^\mu_\Gamma - q^\mu_{\gamma_r}\|}_{\mathcal{E}_{targ}}, \end{aligned} \tag{3}$$

The first term $\mathcal{E}_{estim}$ is the estimation error that contains the variance, while the second term $\mathcal{E}_{targ}$ is the bias in the targets. We analyze them separately below.

6.1. Variance

It is widely known that larger discounts, like larger eligibility traces, incur more variance (Jiang et al., 2015; Petrik & Scherrer, 2009; Kearns & Singh, 2000). In the case of options, somewhat counter-intuitively, larger transition discounts γ_p may incur *less* estimation variance, when they are sufficiently large. This becomes evident when considering $\gamma_p = 1$, for which the variance in γ_p^D due to the random length of the trajectory is entirely removed. The reason this seems at odds with our knowledge of the properties of e.g. λ-returns is because the variance incurred by random option duration is not present there. The following result formalizes this intuition and hints at the type of problems in which it is particularly relevant:

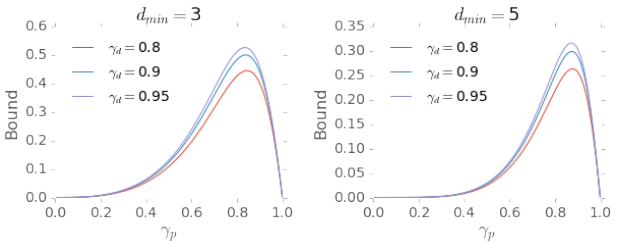

Figure 3. The numerator of Eq. (4) plotted against γ_p for $\Delta v = 0$, $w = 1$, $d_{\max} = 10$ and different values of $d_{\min}$. We see that there is a decrease in variance near $\gamma_p = 1$. Note that the lower values of γ_p that correspond to the other low-variance region may not be sufficient to represent complex policies.

Lemma 1. *Let $d_{\min}$ and $d_{\max}$ be the minimum and maximum option durations across the option set $\mathcal{O}$. Let $\mathcal{F}^o$ denote the set of possible terminating states of an option o. Let each P^o_Γ be estimated from n i.i.d. samples, and let $R^o_{\gamma_r}$ be given. Then, for any policy μ, with probability $1 - \delta$:*

$$\begin{aligned}\mathcal{E}_{estim} &= \|q^\mu_{\widehat{\Gamma}} - q^\mu_\Gamma\| \\ &\leq \frac{(\gamma_p^{d_{\min}} - \gamma_p^{d_{\max}})w + \gamma_p^{d_{\min}} \Delta v}{1 - \|\gamma_d\| \gamma_p^{d_{\min}}} \sqrt{\frac{1}{2n} \log \frac{2|\mathcal{S}||\mathcal{O}|}{\delta}}, \end{aligned} \tag{4}$$

where

$$\Delta v = \max_{o \in \mathcal{O}} \left(\max_{s \in \mathcal{F}^o} v^\mu_\Gamma(s) - \min_{s \in \mathcal{F}^o} v^\mu_\Gamma(s) \right)$$

is the maximum variation of value in terminating states, and $w = \max_{o \in \mathcal{O}} \min_{s \in \mathcal{F}^o} \gamma_d(s) v^\mu_\Gamma(s)$.

The lemma is proven in Appendix B. Intuitively, Δv is a measure of variability of an option's qualitative outcome, while w is a bound on that outcome together with its associated discount. The shape of the bound depends on the relationship between these two quantities. In particular, if Δv, the variation in the values of final states, is not too large, then $\mathcal{E}_{estim}$ is proportional to $\gamma_p^{d_{\min}} - \gamma_p^{d_{\max}}$, and monotonically decreases with γ_p when γ_p is large (see Figure 3 for example shapes). We will observe this behavior empirically on a control task. This class of options corresponds to the typical "goal-directed" options that terminate in a handful of similar states. On the other hand if Δv is large, heavier discounting is needed to account for the difference, and accounting for option duration becomes more important.

6.2. Bias

Now, let us turn to the error $\mathcal{E}_{targ}$ incurred by the discrepancy in the targets.

Lemma 2. *Let $d_{\min}$ and $d_{\max}$ be the minimum and maximum option durations across the option set $\mathcal{O}$. Let μ be*

a policy over options and consider the difference in the value of μ w.r.t. the option models $\{(R^o_{\gamma_r}, P^o_{\gamma_r})\}_{o\in\mathcal{O}}$ and $\{(R^o_{\gamma_r}, P^o_\Gamma)\}_{o\in\mathcal{O}}$. We have:

$$\mathcal{E}_{targ} = \|q^\mu_\Gamma - q^\mu_{\gamma_r}\|_\infty \leq \frac{r_{\max}\left((\gamma_p-\gamma_r)(\gamma_p^{d_{\min}}+1)+\gamma_p(1-\|\gamma_d\|)\right)}{(1-\gamma_r)^2(1-\|\gamma_d\|\gamma_p^{d_{\min}})}.$$

The lemma is proven in Appendix C. Consider the second factor in the numerator of the bound. It is composed of two terms, one that reflects the difference between γ_p and γ_r, and another additive term that has to do with the inter-option discount γ_d. If $\|\gamma_d\| = 1$, and there is no inter-option discounting, this term vanishes, and the error reduces to that incurred by the difference in the discounts. Otherwise, there is some bias introduced by $\|\gamma_d\| \neq 1$, and some bias introduced by $\gamma_p \neq \gamma_r$. Even though the worst-case bound is additive, these biases can sometimes be "in opposite directions", and reduce the overall error when compared to either one in isolation. In fact, there is a value of γ_d that reduces bias all the way to zero, even if $\gamma_p \neq \gamma_r$. The following proposition derives a sufficient condition for this (proof in Appendix D).

Proposition 1. *If $\gamma^o_d(s,s') = \frac{P^o_{\gamma_r}(s'|s)}{P^o_{\gamma_p}(s'|s)}$, $\forall s, s' \in \mathcal{S}$, $\forall o \in \mathcal{O}$, there is no bias in the value function, for any γ_p. That is: $q^\mu_\Gamma = q^\mu_{\gamma_r}$ for any policy μ.*

This result suggests an interpretation of γ^o_d as a particular importance sampling ratio of the two option models.[4] While it is unlikely to be able to obtain it exactly, even an approximate γ^o_d may help balance the bias (Munos et al., 2016). We leave a precise characterization of the general case of this for the future, Finally, from Eq. (3) and Lemmas 1 and 2, we have our result:

Theorem 2. *Let $d_{\min}$ and $d_{\max}$ be the minimum and maximum option durations across the option set $\mathcal{O}$. Let $\mathcal{F}^o$ denote the set of possible terminating states of an option o. Let μ be a policy over options, and let each P^o_Γ be estimated from n i.i.d. samples. Then, with probability $1-\delta$, the error in the estimate $q^\mu_{\widehat{\Gamma}}$ is bounded by:*

$$\mathcal{E} = \|q^\mu_{\widehat{\Gamma}} - q^\mu_{\gamma_r}\| \leq \underbrace{\frac{(\gamma_p^{d_{\min}}-\gamma_p^{d_{\max}})w + \gamma_p^{d_{\min}}\Delta v}{1-\|\gamma_d\|\gamma_p^{d_{\min}}}\sqrt{\frac{1}{2n}\log\frac{2|\mathcal{S}||\mathcal{O}|}{\delta}}}_{variance} + \underbrace{r_{\max}\frac{(\gamma_p-\gamma_r)(\gamma_p^{d_{\min}}+1)+\gamma_p(1-\|\gamma_d\|)}{(1-\gamma_r)^2(1-\|\gamma_d\|\gamma_p^{d_{\min}})}}_{bias}$$

[4] Note that in order for the form of γ^o_d from this proposition to hold, γ^o_d must be a full (rather than diagonal) matrix, whose value is closely related to that of the option transition model, e.g. if $\gamma_p = 1$, $\gamma^o_d = P^o_{\gamma_r}$ exactly.

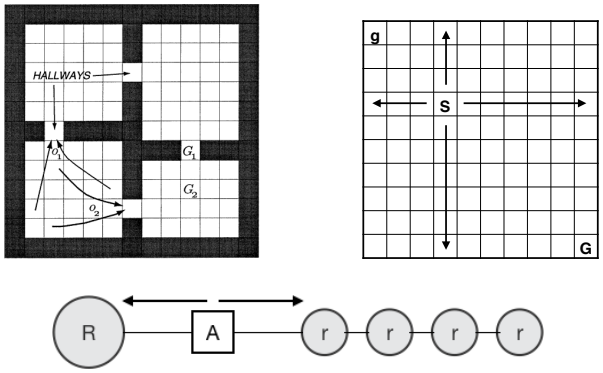

Figure 4. The domains used in our experiments. **Left** Four Rooms. The agent starts in the top left room, and aims to navigate to G_1 via options that navigate to hallways. The option policies are ϵ-soft and extremely noisy with $\epsilon = 0.5$. **Right** Growing Gridworld. The agent's task is to get from the start state S to the goal G. There is another distractor goal g with a smaller reward. **Bottom** Chain. The goal is to collect as much reward as possible in a limited time, by avoiding the distractor reward on the left, and pursuing the sequence of smaller rewards instead.

where

$$\Delta v = \max_{o\in\mathcal{O}}\left(\max_{s\in\mathcal{F}^o} v^\mu_\Gamma(s) - \min_{s\in\mathcal{F}^o} v^\mu_\Gamma(s)\right)$$

is the maximum variation of value in terminating states, and $w = \max_{o\in\mathcal{O}}\min_{s\in\mathcal{F}^o}\gamma_d(s)v^\mu_\Gamma(s)$.

7. Experiments

We illustrate empirically the key ideas of this paper:

1. the bias-variance tradeoff obtained in Theorem 2; and
2. the ability of time dilation to extend the agent's horizon and preserve far-sighted policies, irrespectively of the size of the environment;

Our approximate planning setting is similar to that described by Jiang et al. (2015). Following that work, and since the reward model is unaffected by our proposal, we do not estimate the reward model in these experiments, and instead use its true value (which can be computed exactly in these tasks). Finally, in 7.3, we evaluate the *learning* performance on an illustrative task with characteristic properties.

7.1. Bias-Variance

We investigate whether the analytical bias-variance tradeoff can be observed in practice in the control setting on the classical Four Rooms domain (Sutton et al., 1999). Here, the agent aims to navigate to a goal location via options that navigate from inside of each room to its hallways (see

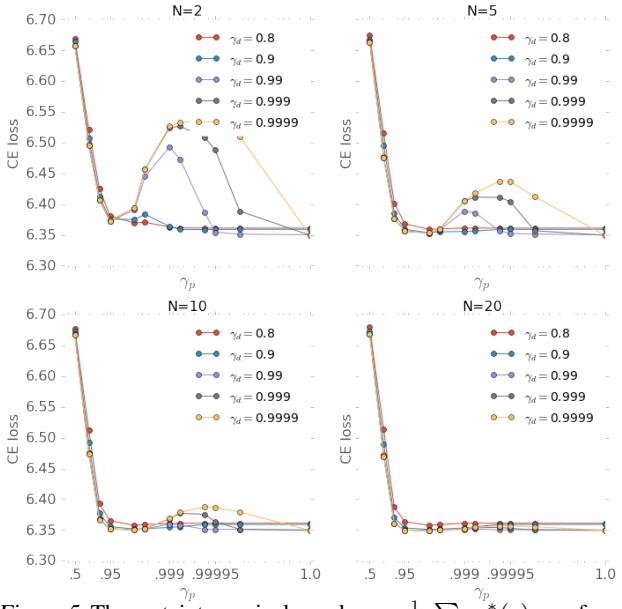

Figure 5. The certainty equivalence loss $-\frac{1}{|\mathcal{S}|}\sum_s v^*_{\widehat{\Gamma}}(s)$ as a function of γ_p, for different values of γ_d (lower is better). The reward model is known, the transition model is estimated from N samples, and $v^*_{\widehat{\Gamma}}$ is obtained from solving the corresponding MDP. Average of 100 independent runs. Notice the similarity with Fig. 3, which diminishes as N increases, since the effects of the variance then diminish, while the large error for small γ-s is due to a large bias. Note the log scale, where the value at 1.0 is biased to be finite.

Fig. 4: Left). The reward is zero everywhere, except for the goal, where it is 10. To evaluate the effects of varied option duration, we add ϵ-noise to the typically deterministic option policies. That is: an option takes an action recommended by its original π^o w.p. $1-\epsilon$, and a random action w.p. ϵ. To obtain a clear picture, we consider a very noisy case of $\epsilon = 0.5$.

For each option o, and for each state $s \in \mathcal{I}^o$, we sample N trajectories to obtain an estimate $\widehat{P^o_\Gamma}$ of P^o_Γ. We then perform policy iteration w.r.t. the approximate models $\widehat{P^o_\Gamma}$ and the true reward models $R^o_{\gamma_r}$ to obtain the approximate optimal policy $\pi^*_{\widehat{\Gamma}}$. We then report the certainty equivalence (CE) loss $-\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} v^*_{\widehat{\Gamma}}(s)$ for the value of this $\pi^*_{\widehat{\Gamma}}$. See Fig. 5 for the results. Notice how the loss curves mimic the bound on the variance term from Lemma 1 closely for reasonably high γ_p, while the bias term dominates the performance of the low γ_p-s. Note that because options terminate at exactly one state (the hallway), Δv is zero, and the variance is entirely eliminated at $\gamma_p = 1$.

7.2. Horizon Invariance

Recall the scenario described in Sec. 3. We simulate an experiment that mimics this scenario and observe that our intuitions hold in practice numerically. In particular, we consider a simple Growing Gridworld task (Fig. 4: Right).

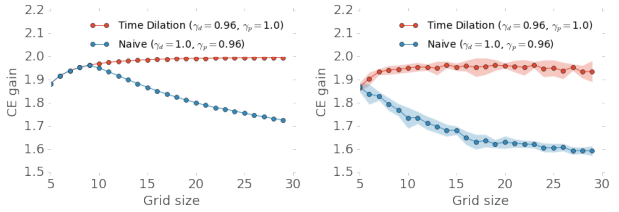

Figure 6. The certainty equivalence *gain* $\frac{1}{|\mathcal{S}|}\sum_s v(s)$ as a function of the grid size (higher is better). The value function v is the value w.r.t. a high γ_{eval} of the optimal policies w.r.t. **Left** the exact model P^o_Γ **Right** the approximate model $\widehat{P^o_\Gamma}$. The shaded area denotes standard deviation. In both cases the performance of the variant $\gamma_p < 1$ deteriorates with the size of the grid, while the variant with $\gamma_p = 1, \gamma_d < 1$ is indifferent to the size of the grid. Note that this pattern is irrespective of the chosen value of γ_p and would occur for some grid size for any γ_p.

There are two terminal states: g with a smaller reward (of 1) and G with a larger reward (of 2). The preference of the agent between them is entirely determined by its discount factor γ_p. As before, we estimate $\widehat{P^o_\Gamma}$ from N samples, and obtain $\pi^*_{\widehat{\Gamma}}$ by policy iteration. We take the value of N to be 2 here. For the estimation to be less trivial, we consider ϵ-soft option policies, as described above, with $\epsilon = 0.05$. We then consider both the value $v^{\widehat{\pi^*}}_{\gamma_{eval}}$ of the optimal policy $\pi^*_{\widehat{\Gamma}}$ w.r.t. approximate model, and the value $v^{\pi^*}_{\gamma_{eval}}$ of the optimal policy π^*_Γ w.r.t. the true model P^o_Γ, both evaluated with a very high $\gamma_{eval} = 1 - 10^{-8}$.

We compare two variants: one with $\gamma_p < 1, \gamma_d = 1$ (corresponding to the classical option model), and the other with $\gamma_p = 1, \gamma_d < 1$ (exploiting time dilation). The reward model is computed with the same value of $\gamma_r < 1$ for both cases. Figure 6 reports the certainty equivalence gain $\frac{1}{|\mathcal{S}|}\sum_s v(s)$ for both the exact and approximate optimal values of these variants. The same pattern is induced in both cases, and the values of the optimal policies diminish, as the size of the grid gets larger. Time dilation on the other hand allows the options to maintain the same performance regardless of the size of the grid.

7.3. Learning Performance

We now study the learning performance of an agent in relation to time dilation. We use a simple illustrative task imbued with the realistic properties that motivated our proposal: the chain given in Fig. 4 (bottom). The agent must choose between greedily pursuing a large reward R or a sequence of smaller rewards, whose sum is larger than R. The setting is continuing, but there is a small chance of death which doesn't allow for collecting all the rewards. There is a variable amount of steps between the rewards, and each reward is only collectible once. The agent has two options, one for each movement direction, which terminate when a reward is reached. At each primitive time step, there is zero-

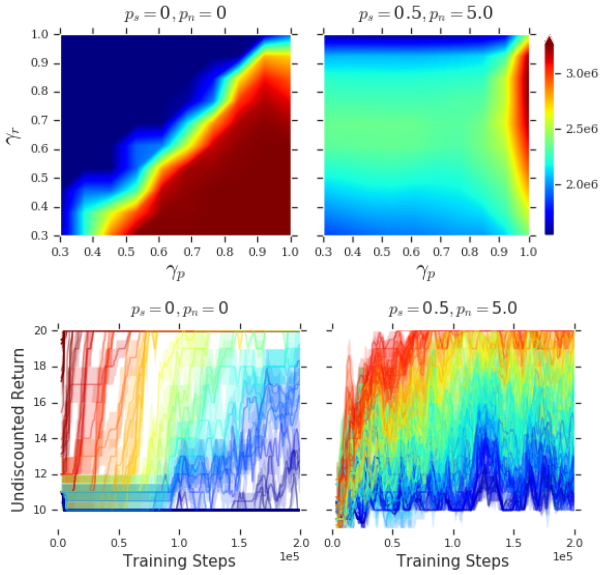

Figure 7. **Top:** Cumulative performance (total sum of undiscounted reward collected during training) for all of the γ_p, γ_r settings, plotted with linear interpolation. **Bottom:** Learning curves for each setting showing the mean undiscounted return per episode (with colors corresponding to the heatmap above).

mean reward-noise with parameterized standard deviation p_n, and a "stickiness" probability p_s, with which the agent stays in place.

We use Q-learning over options with call-and-return execution. When an option terminates after D steps, it receives a γ_r-discounted D-step return, and bootstraps with a discount of $\gamma_d \gamma_p^D$. Fig. 7 summarizes the results across a range of γ_r and γ_p for two settings of the task: noise-free and with noise as described above.[5] Qualitatively, there are three levels of performance: convergence to the optimal policy π_r^* (red), convergence to the suboptimal policy π_R^* (blue), and slower convergence to π_r^* (green). In the noise-free case, we observe a stark effect of the importance of the γ_r, γ_p relationship. When $\gamma_p > \gamma_r$ (i.e. with time dilation), the agent is able to quickly discover π_r^*, even for very low values of γ_r. Reward noise has the effect of yielding intermediate performance for more settings (i.e. a larger green region). In this more challenging case, settings with intermediate values of γ_r and high values of γ_p yield the best performance, showcasing the effectiveness of time dilation.

8. Related Work

The analysis we provided is closely related to that by Jiang et al. (2015), and earlier results along the same lines of Petrik & Scherrer (2009). In both works, the authors consider the tradeoff on estimation variance vs target bias in the quality of the approximate planning solution, due to using a lower discount factor. Jiang et al. (2015) show that it is beneficial to use a lower discount when the number of samples used to estimate the transition model is small. These implications carry over to γ_d in the context of options, while γ_p controls a more subtle tradeoff that has to do with the random duration of options.

Petrik & Scherrer (2009) focus on the bias aspect, and show that in some problems the bias due to using a lower discount can be lower than predicted by the worst case, especially in problems whose rewards are sparse. It is interesting to identify a similar structure for options.

General transition-based discounting is introduced by White (2017), where it is proposed to use the discount as a formalism for reinforcement learning *tasks*, and shown that each option then represents a task, since the termination condition together with the step discount incurs a transition discount. We alter the option transition model, and hence incur *option*-transition discounts.

9. Discussion

We proposed to provide options with autonomy over their own timescale, by introducing time dilation into the option transition model. We analyzed the bias-variance incurred by doing so, and verified the analytical predictions empirically. These insights are immediately applicable to any setting using the options framework.

While our experiments are in the control setting, the analysis applies to a fixed policy μ. In order to extend Lemma 1 to apply to all policies, we need to consider the relationship of the number of optimal policies under a given model $|M_\Gamma|$ to γ_d and γ_p. Jiang et al. (2015) give an interesting interpretation of the discount as a policy complexity control parameter, and show that this quantity grows monotonically with γ. This analysis is trickier with options, due to there being two parameters γ_d and γ_p controlling the discount.

This work offers a formal foundation, from which there are many possible future directions. *Learning* the appropriate discount parameters for the given set of options, or even as a way to *discover* options, is the obvious next step. The non-homogeneous discounting structure may act similarly to a deliberation cost (Harb et al., 2018) and encourage non-trivial option discovery. On the other hand, the horizon invariance property lends itself well to the transfer setting, which is where hierarchical methods are most appropriate. Namely, because our proposal removes the strict dependencies on the primitive timescale, it may be possible to successively transfer option models and policies learned on smaller instances of a task to larger instances of the same task, as a form of *scaffolding*.

[5] The per-decision γ_d in this task does not affect the results qualitatively, and we report results with $\gamma_d = 1$.

Acknowledgments

The authors thank Hado van Hasselt, Adam White, Will Dabney, Gheorghe Comanici and Andre Barreto for their feedback on earlier manuscripts of this paper, and Mohammad Azar for helping with one of the bounds.

References

Bacon, P.-L. and Precup, D. A matrix splitting perspective on planning with options. *arXiv preprint arXiv:1612.00916*, 2016.

Bellman, R. *Dynamic programming*. Princeton University Press, 1957.

Bertsekas, D. P. and Tsitsiklis, J. N. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

Harb, J., Bacon, P.-L., Klissarov, M., and Precup, D. When waiting is not an option: Learning options with a deliberation cost. In *Proceedings of 32nd AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

Harutyunyan, A., Vrancx, P., Bacon, P.-L., Precup, D., and Nowé, A. Learning with options that terminate off-policy. In *Proceedings of 32nd AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

Jiang, N., Kulesza, A., Singh, S., and Lewis, R. The dependence of effective planning horizon on model accuracy. In *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS-15)*, pp. 1181–1189, 2015.

Kearns, M. J. and Singh, S. P. Bias-variance error bounds for temporal difference updates. In *Proceedings of the 13th International Conference on Computational Learning Theory (COLT-00)*, pp. 142–147, 2000.

Lehnert, L., Laroche, R., and van Seijen, H. On value function representation of long horizon problems. In *Proceedings of 32nd AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

Mahadevan, S. Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine learning*, 22(1-3):159–195, 1996.

Munos, R., Stepleton, T., Harutyunyan, A., and Bellemare, M. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 1046–1054, 2016.

Petrik, M. and Scherrer, B. Biasing approximate dynamic programming with a lower discount factor. In *Advances in Neural Information Processing Systems 21*, pp. 1265–1272, 2009.

Precup, D., Sutton, R. S., and Singh, S. Theoretical results on reinforcement learning with temporally abstract options. In *Proceedings of the 10th European conference on machine learning (ECML-98)*, pp. 382–393, 1998.

Puterman, M. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994. ISBN 0471619779.

Schwartz, A. A reinforcement learning method for maximizing undiscounted rewards. In *Proceedings of the 10th International Conference on Machine Learning (ICML-93)*, 1993.

Sutton, R. and Barto, A. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2017.

Sutton, R. S., Precup, D., and Singh, S. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2): 181–211, 1999.

Tsitsiklis, J. N. and Van Roy, B. On average versus discounted reward temporal-difference learning. *Machine Learning*, 49:179–191, 2002.

White, M. Unifying task specification in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML-17)*, 2017.

Yu, H. and Bertsekas, D. P. Weighted bellman equations and their applications in approximate dynamic programming. Technical report, Citeseer, 2012.