Appendix
Reinforcement Learning with Non-Exponential Discounting

A Convergence proof for the value function under hyperbolic discounting

In the following, we assume a hyperbolic survival function as presented in Eq. (2), i.e.,

$$S(t;\alpha,\beta) = \frac{1}{(\frac{t}{\beta}+1)^{\alpha}}.$$

Part I *If the reward function $R(\mathbf{x},\mathbf{u},t)$ is bounded above for all $(\mathbf{x},\mathbf{u},t) \in \mathcal{X} \times \mathcal{U} \times \mathbb{R}_0^+$, and $\alpha_0 > 1$, the value function defined in equation Eq. (6) is well-defined.*

We denote the supremum of the reward function $R(\mathbf{x},\mathbf{u},t)$ for all $(\mathbf{x},\mathbf{u},t) \in \mathcal{X} \times \mathcal{U} \times \mathbb{R}_0^+$ by r_{sup}. We find

$$\begin{aligned}
V^*(\mathbf{x},t) &= \max_{\mathbf{u}_{[t,\infty)}} \mathbb{E}\left[\int_t^\infty \frac{S(\tau)}{S(t)}\, R(\mathbf{X}(\tau),\mathbf{u}(\tau),\tau)\,\mathrm{d}\tau \,\middle|\, \mathbf{X}(t)=\mathbf{x}\right] \\
&\leq \int_t^\infty \frac{S(\tau)}{S(t)}\, r_{\text{sup}}\,\mathrm{d}\tau \\
&= \frac{r_{\text{sup}}}{S(t)} \int_t^\infty S(\tau)\,\mathrm{d}\tau \\
&= \frac{r_{\text{sup}}}{S(t)} \int_t^\infty \frac{1}{\left(\frac{\tau}{\beta}+1\right)^{\alpha}}\,\mathrm{d}\tau \\
&\leq \frac{r_{\text{sup}}}{S(t)} \int_t^\infty \frac{1}{\left(\frac{\tau}{\beta}\right)^{\alpha}}\,\mathrm{d}\tau \\
&= \frac{\beta^{\alpha}\, r_{\text{sup}}}{S(t)} \int_t^\infty \frac{1}{\tau^{\alpha}}\,\mathrm{d}\tau \\
&= \frac{\beta^{\alpha}\, r_{\text{sup}}}{S(t)} \left[\frac{\tau^{1-\alpha}}{1-\alpha}\right]_{\tau=t}^{\infty} \\
&= \frac{\beta^{\alpha}\, r_{\text{sup}}}{S(t)\,(1-\alpha)} \left[\tau^{1-\alpha}\right]_{\tau=t}^{\infty},
\end{aligned}$$

which is finite for $\alpha > 1$.

Part II *If $R(\mathbf{x},\mathbf{u},t)$ is bounded below for all $(\mathbf{x},\mathbf{u},t) \in \mathcal{X} \times \mathcal{U} \times \mathbb{R}_0^+$, and $\alpha_0 \leq 1$, the value function defined in equation Eq. (6) is not well-defined.*

License: CC BY 4.0 International - Creative Commons, Attribution
https://creativecommons.org/licenses/by/4.0/

We denote the infimum of the reward function $R(\mathbf{x}, \mathbf{u}, t)$ for all $(\mathbf{x}, \mathbf{u}, t) \in \mathcal{X} \times \mathcal{U} \times \mathbb{R}_0^+$ by r_{inf}. We find

$$
\begin{aligned}
V^*(\mathbf{x}, t) &= \max_{\mathbf{u}_{[t,\infty)}} \mathbb{E}\left[\int_t^\infty \frac{S(\tau)}{S(t)}\, R(\mathbf{X}(\tau), \mathbf{u}(\tau), \tau)\,\mathrm{d}\tau \;\middle|\; \mathbf{X}(t) = \mathbf{x}\right] \\
&\geq \int_t^\infty \frac{S(\tau)}{S(t)}\, r_{\text{inf}}\,\mathrm{d}\tau \\
&= \frac{r_{\text{inf}}}{S(t)} \int_t^\infty S(\tau)\,\mathrm{d}\tau \\
&= \frac{r_{\text{inf}}}{S(t)} \int_t^\infty \frac{1}{\left(\frac{\tau}{\beta} + 1\right)^\alpha}\,\mathrm{d}\tau \\
&= \frac{r_{\text{inf}}}{S(t)} \int_t^\infty \frac{1}{\left(\frac{\tau+\beta}{\beta}\right)^\alpha}\,\mathrm{d}\tau \\
&= \frac{\beta^\alpha\, r_{\text{inf}}}{S(t)} \int_t^\infty \frac{1}{(\tau+\beta)^\alpha}\,\mathrm{d}\tau \\
&= \frac{\beta^\alpha\, r_{\text{inf}}}{S(t)} \int_{t+\beta}^\infty \frac{1}{\tau^\alpha}\,\mathrm{d}\tau \\
&= \frac{\beta^\alpha\, r_{\text{inf}}}{S(t)} \left[\frac{1}{\tau^\alpha}\right]_{\tau=t+\beta}^\infty \\
&= \frac{\beta^\alpha\, r_{\text{inf}}}{S(t)} \left[\frac{\tau^{1-\alpha}}{1-\alpha}\right]_{\tau=t+\beta}^\infty \\
&= \frac{\beta^\alpha\, r_{\text{inf}}}{S(t)\,(1-\alpha)} \left[\tau^{1-\alpha}\right]_{\tau=t+\beta}^\infty,
\end{aligned}
$$

in which the integral diverges for $\alpha \leq 1$.

B Full derivation of the HJB equation

In this section, we provide a full derivation for the HJB equation introduced in Section 4.2. We start with the value function defined in Eq. (6), i.e.,

$$
V^*(\mathbf{x}, t) = \max_{\mathbf{u}_{[t,\infty)}} \mathbb{E}\left[\int_t^\infty \frac{S(\tau)}{S(t)} R(\mathbf{X}(\tau), \mathbf{u}(\tau), \tau)\,\mathrm{d}\tau \;\middle|\; \mathbf{X}(t) = \mathbf{x}\right].
$$

First, we split the integral into two terms and obtain

$$
\begin{aligned}
V^*(\mathbf{x}, t) = \max_{\mathbf{u}_{[t,t+\Delta t]}} \mathbb{E}\Bigg[&\int_t^{t+\Delta t} \frac{S(\tau)}{S(t)} R(\mathbf{X}(\tau), \mathbf{u}(\tau), \tau)\,\mathrm{d}\tau \\
&+ \int_{t+\Delta t}^\infty \frac{S(\tau)}{S(t)} R(\mathbf{X}(\tau), \mathbf{u}(\tau), \tau)\,\mathrm{d}\tau \;\Bigg|\; \mathbf{X}(t) = \mathbf{x}\Bigg].
\end{aligned}
$$

By identifying the second term as the value function of state $x(t + \Delta t)$ at time $t + \Delta t$, we obtain the recursive formulation

$$
\begin{aligned}
V^*(\mathbf{x}, t) = \max_{\mathbf{u}_{[t,t+\Delta t]}} \mathbb{E}\Bigg[&\int_t^{t+\Delta t} \frac{S(\tau)}{S(t)} R(\mathbf{X}(\tau), \mathbf{u}(\tau), \tau)\,\mathrm{d}\tau \\
&+ \frac{S(t+\Delta t)}{S(t)} V^*(\mathbf{X}(t+\Delta t), t+\Delta t) \;\Bigg|\; \mathbf{X}(t) = \mathbf{x}\Bigg].
\end{aligned}
$$

Consider a small Δt, then the first term evaluates to

$$
\int_t^{t+\Delta t} \frac{S(\tau)}{S(t)} R(\mathbf{X}(\tau), \mathbf{u}(\tau), \tau)\,\mathrm{d}\tau = R(\mathbf{X}(t), \mathbf{u}(t), t) \cdot \Delta t + o(\Delta t).
$$

For the second term, we apply a Taylor expansion and get

$$\begin{aligned}V^*(\mathbf{X}(t+\Delta t), t+\Delta t) &= V^*(\mathbf{X}(t), t) + \int_t^{t+\Delta t} \frac{\mathrm{d}}{\mathrm{d}\tau} V^*(\mathbf{X}(\tau), \tau)\,\mathrm{d}\tau + o(\Delta t) \\ &= V^*(\mathbf{X}(t), t) + \int_t^{t+\Delta t} V^*_{\mathbf{x}}(\mathbf{X}(\tau), \tau)\,\mathrm{d}\mathbf{X}(\tau) + \int_t^{t+\Delta t} V^*_t(\mathbf{X}(\tau), \tau)\,\mathrm{d}\tau + o(\Delta t).\end{aligned}$$

Here, the second term can be evaluated using Itô's formula as

$$\begin{aligned}\int_t^{t+\Delta t} V^*_{\mathbf{x}}(\mathbf{X}(\tau), \tau)\,\mathrm{d}\mathbf{X}(\tau) &= \int_t^{t+\Delta t} V^*_{\mathbf{x}}(\mathbf{X}(\tau), \tau)\, f(\mathbf{X}(\tau), \mathbf{u}(\tau), \tau)\,\mathrm{d}\tau \\ &\quad + \int_t^{t+\Delta t} \frac{1}{2}\operatorname{tr}\left\{V^*_{\mathbf{xx}}(\mathbf{X}(\tau), \tau)\, G(\mathbf{X}(\tau), \mathbf{u}(\tau), \tau)\, G(\mathbf{X}(\tau), \mathbf{u}(\tau), \tau)^T\right\}\mathrm{d}\tau \\ &\quad + \int_t^{t+\Delta t} V^*_{\mathbf{x}}(\mathbf{X}(\tau), \tau)\, G(\mathbf{X}(\tau), \mathbf{u}(\tau), \tau)\,\mathrm{d}\mathbf{W}(\tau) + o(\Delta t).\end{aligned}$$

Plugging in these terms into the equation above and dividing both sides by Δt yields

$$\begin{aligned}\frac{1 - \frac{S(t+\Delta t)}{S(t)}}{\Delta t} V^*(\mathbf{X}(t), t) &= \max_{\mathbf{u}_{[t, t+\Delta t]}} \mathbb{E}\left[\frac{1}{\Delta t}\int_t^{t+\Delta t} \frac{S(\tau)}{S(t)} R(\mathbf{X}(\tau), \mathbf{u}(\tau), \tau)\,\mathrm{d}\tau\right. \\ &\quad + \frac{1}{\Delta t}\int_t^{t+\Delta t} V^*_{\mathbf{x}}(\mathbf{X}(\tau), \tau)\, f(\mathbf{X}(\tau), \mathbf{u}(\tau), \tau)\,\mathrm{d}\tau + \int_t^{t+\Delta t} V^*_t(\mathbf{X}(\tau), \tau)\,\mathrm{d}\tau \\ &\quad + \frac{1}{\Delta t}\int_t^{t+\Delta t} \frac{1}{2}\operatorname{tr}\left\{V^*_{\mathbf{xx}}(\mathbf{X}(\tau), \tau)\, G(\mathbf{X}(\tau), \mathbf{u}(\tau), \tau)\, G(\mathbf{X}(\tau), \mathbf{u}(\tau), \tau)^T\right\}\mathrm{d}\tau \\ &\quad \left. + \frac{1}{\Delta t}\int_t^{t+\Delta t} V^*_{\mathbf{x}}(\mathbf{X}(\tau), \tau)\, G(\mathbf{X}(\tau), \mathbf{u}(\tau), \tau)\,\mathrm{d}\mathbf{W}(\tau) + \frac{o(\Delta t)}{\Delta t} \;\middle|\; \mathbf{X}(t) = \mathbf{x}\right].\end{aligned}$$

The factor on the l.h.s. in the limit $\Delta t \to 0$ can be recognized to be the hazard rate (cf. Eq. (1)),

$$\lim_{\Delta t \to 0} \frac{1 - \frac{S(t+\Delta t)}{S(t)}}{\Delta t} = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{S(t) - S(t+\Delta t)}{\Delta t} = \alpha(t).$$

Taking the limit $\Delta t \to 0$ on both sides and calculating the expectation w.r.t. $\mathbf{W}(t)$, we obtain the HJB equation

$$\begin{aligned}\alpha(t) V^*(\mathbf{x}, t) = \max_{\mathbf{u}} \Big[&R(\mathbf{x}, \mathbf{u}, t) + V^*_t(\mathbf{x}, t) + V^*_{\mathbf{x}}(\mathbf{x}, t)\, f(\mathbf{x}, \mathbf{u}, t) \\ &+ \frac{1}{2}\operatorname{tr}\left\{V^*_{\mathbf{xx}}(\mathbf{x}, t)\, G(\mathbf{x}, \mathbf{u}, t)\, G(\mathbf{x}, \mathbf{u}, t)^T\right\}\Big].\end{aligned}$$

C Bellman equation for discrete time

We consider the discrete-time setting, in which the objective is given as

$$J(\mathbf{u}_0, \mathbf{u}_1, \dots) = \mathbb{E}\left[\sum_{\tau=0}^{\infty} S(\tau)\, R(\mathbf{X}_\tau, \mathbf{u}_\tau, \tau)\right].$$

As in the continuous-time case, we can define the value function as

$$V(\mathbf{x}, t) = \max_{\mathbf{u}_t, \mathbf{u}_{t+1}, \dots} \mathbb{E}\left[\sum_{\tau=t}^{\infty} \frac{S(\tau)}{S(t)}\, R(\mathbf{X}_\tau, \mathbf{u}_\tau, \tau) \;\middle|\; \mathbf{X}_t = \mathbf{x}\right].$$

By identifying the recursive definition of the value function and evaluating terms, we obtain the Bellman equation

$$V(\mathbf{x}, t) = \max_{\mathbf{u}} \left\{R(\mathbf{x}, \mathbf{u}, t) + \lambda(t)\,\mathbb{E}\left[V(\mathbf{X}_{t+1}, t+1) \mid \mathbf{X}_t = \mathbf{x}\right]\right\},$$

with $\lambda(t) = S(t+1)/S(t)$ being the hazard probability at time t.

D Value function approximation and collocation method

In the collocation method in Algorithm 1, we need to sample random states $\hat{\mathbf{x}}_i$ and time points $\hat{t}_i$ for minimizing $\sum_i E(V^\psi, \hat{\mathbf{x}}_i, \hat{t}_i)^2$. If we assume a bounded state space $\mathcal{X} \in \mathbb{R}$, we can sample $\hat{\mathbf{x}}_i$ uniformly from this space. The time points $\hat{t}_i \in \mathbb{R}_0^+$ can be sampled from an exponential distribution. To do so, we first draw $\hat{y}_i \sim \mathrm{Uniform}(0,1)$ and compute $\hat{t}_i = -\log(1-\hat{y}_i)/\lambda$. To feed a normalized value of time into the network, we use $\hat{y}_i$ instead of $\hat{t}_i$ as input to the network. We denote the value function network depending on y by $\tilde{V}(\mathbf{x}, y)$. Given a specific time value t, we can compute its representation via $y(t) = 1 - \exp(-\lambda t)$.

When computing the partial derivative V_t, we have to take this reparametrization into account. By the chain rule, we find

$$V_t(\mathbf{x}, t) = \tilde{V}_y(\mathbf{x}, t)\, y_t(t),$$

for which we have with the chosen parametrization

$$y_t(t) = \lambda \exp(-\lambda t).$$

In general, there are multiple solutions to the HJB equation and the encountered solution depends on the initialization of the function approximator [29, 31]. In other work, this problem has been dealt with by omitting stochastic terms in the first episodes of training or annealing the discount factor [29, 31, 33]. We adopt the second approach and move from short to far-sighted discounting to converge to the desired solution. For hyperbolic discounting, we initially add an offset to α_0, leading to a high expected hazard rate. Over time, we decrease the offset to converge to the desired solution.

E Experiments

Investment problem

- State space $\mathcal{X} = [0,1] \times [0,1]$, modeling account balance and interest rate, i.e., $\mathbf{x} = [x_b, x_i]$
- Action space $\mathcal{U} = \{\textit{spend}, \textit{invest}\}$
- Dynamics model

$$f(\mathbf{x}, \mathbf{u}) = \begin{cases} [0,0]^T & \text{if } \mathbf{u} \text{ is } \textit{spend} \\ [0.1, 0]^T & \text{if } \mathbf{u} \text{ is } \textit{invest} \end{cases}$$

$$G(\mathbf{x}, \mathbf{u}) = \begin{pmatrix} 0 & 0 \\ 0 & 0.01 \end{pmatrix}$$

- Reward function

$$R(\mathbf{x}, \mathbf{u}) = R^{\mathbf{x}}(\mathbf{x}) + R^{\mathbf{u}}(\mathbf{u})$$
$$R^{\mathbf{x}}([x_b, x_i]) = x_b \cdot x_i$$
$$R^{\mathbf{u}}(\mathbf{u}) = \begin{cases} 0.1 & \text{if } \mathbf{u} \text{ is } \textit{spend} \\ 0 & \text{if } \mathbf{u} \text{ is } \textit{invest} \end{cases}$$

- Initial belief of hazard rate $\alpha_0 = 3, \beta_0 = 1$ (visualized in Fig. 3)

Line problem

- State space $\mathcal{X} = [-1, 1]$
- Action space $\mathcal{U} = \{\textit{left}, \textit{stay}, \textit{right}\}$
- Dynamics model

$$f(\mathbf{x}, \mathbf{u}) = \begin{cases} -1 & \text{if } \mathbf{u} \text{ is } \textit{left} \\ 0 & \text{if } \mathbf{u} \text{ is } \textit{stay} \\ 1 & \text{if } \mathbf{u} \text{ is } \textit{right} \end{cases}$$

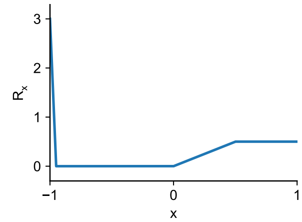
$$G(\mathbf{x}, \mathbf{u}) = \begin{cases} 0.05 & \text{if } \mathbf{u} \in \{\textit{left}, \textit{right}\} \\ 0 & \text{if } \mathbf{u} \text{ is } \textit{stay} \end{cases}$$

- Reward function

$$R(\mathbf{x}, \mathbf{u}) = R^{\mathbf{x}}(\mathbf{x}) + R^{\mathbf{u}}(\mathbf{u})$$

$$R^{\mathbf{x}}(\mathbf{x}) = \begin{cases} 0.5 & \text{if } \mathbf{x} \geq 0.5 \\ \mathbf{x} & \text{if } 0 \leq \mathbf{x} < 0.5 \\ 0 & \text{if } -0.95 \leq \mathbf{x} < 0 \\ -60\mathbf{x} - 57 & \text{if } \mathbf{x} < 0.95 \end{cases}$$

$$R^{\mathbf{u}}(\mathbf{u}) = \begin{cases} 0.1 & \text{if } \mathbf{u} \in \{\textit{left}, \textit{right}\} \\ 0 & \text{if } \mathbf{u} \text{ is } \textit{stay} \end{cases}$$



- Initial belief of hazard rate $\alpha_0 = 5, \beta_0 = 1$ (visualized in Fig. 3)

F Hyperparameters, implementation, and computing resources

Throughout the experiments, we have used the following hyperparameters:

- The neural networks are parametrized as

```
layers = (nn.Linear(input_dim, layer_size),
          nn.Sigmoid(),
          nn.Linear(layer_size, layer_size),
          nn.Sigmoid(),
          nn.Linear(layer_size, output_dim))
model = nn.Sequential(*layers)
```

- For the neural network representing V, we used

```
input_dim = x_dim + 1
output_dim = 1
```

- For the neural network representing V_θ, we used

```
input_dim = x_dim + 1
output_dim = theta_dim
```

- We set $\lambda = 0.2$.
- For the collocation method, we used 10.000 samples in each iteration and 125.000 episodes for the investment problem and 100.000 episodes for the line problem. The initial offset of α_0 was set to 50 and linearly decreased to zero over 50.000 episodes.
- We used Adam optimizer with learning rate 0.003.
- For the runs with exponential discounting, the mean of the initial belief over the hazard rate was taken for λ, i.e., 3 for the investment problem and 5 for the line problem.

More information about Implementation and computing resources:

- Methods were implemented in Python using the PyTorch framework [71], which has been published under a BSD license.
- Resources used: Intel® Xeon® Platinum 9242 Processor, using 8 cores per run.
- Network training took ~50 min. for the investment problem and ~30 min. for the line problem.

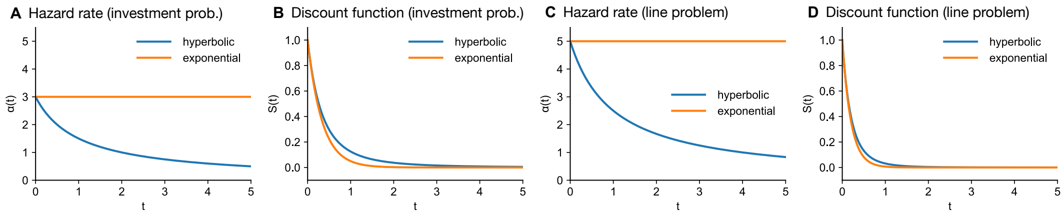


Figure 3: **Hazard rates and Discounting functions. A** Expected hazard rate for the investment problem. For hyperbolic discounting, the expected risk of termination is decreasing over time, while for exponential discounting, the hazard rate is constant. **B** Expected discount function for the investment problem in comparison to an exponential discount function. **C** Expected hazard rate for the line problem for hyperbolic discounting in comparison with the constant hazard rate when applying exponential discounting **D** Expected discount function for the line problem in comparison to an exponential discount function.

G Derivation of the hyperbolic discount function as uncertainty over the constant hazard rate

We assume $P(T > t \mid \lambda) = \exp(-\lambda t)$ and a belief $\lambda \sim \text{Gamma}(\lambda; \alpha, \beta)$. For the expected survival function, we calculate

$$\begin{aligned} S(t) &= \int_\lambda \exp\left(-\lambda t\right) p\left(\lambda\right)\, \mathrm{d}\lambda \\ &= \int_\lambda \exp\left(-\lambda t\right) \frac{\beta^\alpha \lambda^{\alpha-1} \exp(-\beta\lambda)}{\Gamma(\alpha)}\, \mathrm{d}\lambda \\ &= \int_\lambda \frac{\beta^\alpha \lambda^{\alpha-1} \exp(-(\beta+t)\lambda)}{\Gamma(\alpha)}\, \mathrm{d}\lambda \\ &= \frac{\beta^\alpha}{(\beta+t)^\alpha} \int_\lambda \text{Gamma}(\lambda; \alpha, \beta+t)\, \mathrm{d}\lambda \\ &= \frac{1}{\left(\frac{t}{\beta}+1\right)^\alpha}. \end{aligned}$$

H Interpretation of the discount factor as transition to terminal state

A Markov decision process (MDP) with discounting can be converted to an MDP without discounting by adding an additional terminal state Υ [13]. From each state with a certain probability γ, one transitions to the terminal state, and the remaining transition probabilities are renormalized. At the terminal state there is no possibility to transition to any other state and a reward of zero is given. In continuous time, the same formalization can be applied, but we consider a rate instead at which one transitions to the terminal state. Further, we assume in the following that the rate depends on time and denote it by $\lambda(t)$. The probability to be in the terminal states at time Υ is given by the cumulative distribution function (CDF),

$$P(\mathbf{X}(t) = \Upsilon) = P(T < t).$$

The probability of not having terminated yet is given by the complementary cumulative distribution function (CCDF),

$$\begin{aligned} P(\mathbf{X}(t) \neq \Upsilon) &= 1 - P(\mathbf{X}(t) = \Upsilon) \\ &= P(T \geq t) \\ &= S(t), \end{aligned}$$

which is equal to the discount function.

For a constant termination rate, one obtains the CDF and CCDF of the exponential distribution, respectively:

$$\begin{aligned} P(\mathbf{X}(t) = \Upsilon) &= \lambda \int_0^t \exp(-\lambda\tau)\,\mathrm{d}\tau \\ &= 1 - \exp(-\lambda t) \end{aligned}$$

$$\begin{aligned} P(\mathbf{X}(t) \neq \Upsilon) &= 1 - P(\mathbf{X}(t) = \Upsilon) \\ &= \exp(-\lambda t) \end{aligned}$$