Time Consistent Reinforcement Learning for Optimal Consumption under Epstein-Zin Preferences

Matthew Dixon, Ivan Gvozdanovic, and Dominic O'Kane

April 18, 2025

Introduction

In this paper, the authors

- cast Epstein-Zin Utility preferences into a dynamic utility functional framework;
- formulate the dynamic utility function optimization as a Markov Decision Process;
- replace expectation in least square learning with a certainty equivalent operator;
- show how to learn time consistent (aka stationary) optimal policies using temporal difference learning.

Structure of the presentation

First I will introduce two preliminaries:

- Temporal Difference Learning
- Filtering

Then we will go into two topics with Epstein-Zin preferences:

- Dynamic Utility
- Q-Learning

Setup

We consider an MDP (r, Γ, β, P) with state space X and action space A.

- r is bounded and continuous reward function on $\mathsf{G} := \{(x, a) \in \mathsf{X} \times \mathsf{A} : a \in \Gamma(x)\}$;
- Γ is a nonempty, continuous and compact-valued feasible correspondence;
- $\beta \in (0, 1)$ is a discount factor;
- $P(x, a, \mathrm{d}x')$ is a distribution over next period states given current state x and action a.

We further assume that the map $(x, a) \mapsto \int v(x')P(x, a, dx')$ is continuous on G whenever $v \in bc\mathsf{X}$.

Temporal Difference Learning

To begin, we take the MDP model and, given $v \in \mathbb{R}^{\mathsf{X}}$, set

$$q(x, a) := r(x, a) + \beta \int v(x') P(x, a, dx') \qquad ((x, a) \in \mathsf{G}). \tag{1}$$

The function $q(x, a)$ represents the **action-value function** or **Q-function**. It estimates the expected total discounted future reward starting from state x, taking action a.

Given a policy σ, the σ-Q-function can be written as:

$$q_\sigma(x, a) = r(x, a) + \beta \int v_\sigma(x') P(x, a, dx')$$

where v_σ is the lifetime value of policy σ.

The **Q-function** $q^*(x, a)$, gives the maximum expected return achievable from (x, a):

$$q^*(x, a) = \max_{\sigma \in \Sigma} q_\sigma(x, a)$$

where $q_\sigma(x, a)$ is the value of taking action a in state x and then following policy σ.

The Q-function satisfies the **Bellman equation for Q-functions**:

$$q^*(x, a) = r(x, a) + \beta \int \max_{a' \in \Gamma(x')} q^*(x', a') P(x, a, dx')$$

Policy Improvement Theorem

Theorem 1: Policy Improvement Theorem

Let σ and σ' be two policies. such that for all $x \in \mathsf{X}$,

$$\begin{aligned} q_\sigma(x, \sigma'(x)) &:= r(x, \sigma'(x)) + \beta \int v_\sigma(x') P(x, \sigma'(x), dx') \\ &\geqslant r(x, \sigma(x)) + \beta \int v_\sigma(x') P(x, \sigma(x), dx') \\ &=: v_\sigma(x) \end{aligned}$$

Then,

$$v_{\sigma'}(x) \geqslant v_\sigma(x)$$

Moreover, if $q_\sigma(x, \sigma'(x)) > v_\sigma(x)$ for some $x \in \mathsf{X}$, then $v_{\sigma'}(x) > v_\sigma(x)$ for all $x \in \mathsf{X}$.

Proof.

To begin with, we can expand $v_\sigma(x)$ as follows:

$$
\begin{aligned}
v_\sigma(x) &= r(x, \sigma(x)) + \beta \int v_\sigma(x') P(x, \sigma(x), dx') \\
&\leqslant r(x, \sigma'(x)) + \beta \int q_\sigma(x, \sigma'(x)) P(x, \sigma'(x), dx') \\
&= r(x, \sigma'(x)) \\
&+ \beta \int \left[r(x', \sigma'(x')) + \beta \int v_\sigma(x'') P(x', \sigma'(x'), dx'') \right] P(x, \sigma'(x), dx')
\end{aligned}
$$

Continuing the expansion, we get:

$$
\begin{aligned}
v_\sigma(x) &\leqslant \sum_{t=0}^{\infty} \beta^t \int r(x_t, \sigma'(x_t)) P^t_{\sigma'}(x, dx_t) \\
&= v_{\sigma'}(x)
\end{aligned}
$$

where $P^t_{\sigma'}(x, dx_t)$ denotes the t-step transition probability under policy σ'.

□

Q-Learning with TD(0)

Q-learning is a model-free RL algorithm that aims to learn the optimal Q-function q^* directly from experience / data (state transitions $(x, a, \tilde{r}, x')$) without needing the transition model P.

Here we use the **Temporal Difference (TD(0))** update rule. Let $q_k(x, a)$ be the estimate of the Q-function at iteration k. After observing a transition (x, a, r, x'), the update is:

$$q_{k+1}(x, a) \leftarrow (1 - \eta)\, q_k(x, a) + \eta \underbrace{\left(\tilde{r} + \beta \max_{a' \in \Gamma(x')} q_k(x', a') \right)}_{\text{TD Target}}$$

where

- $\eta \in (0, 1]$ is the learning rate.
- $\tilde{r} + \beta \max_{a' \in \Gamma(x')} q_k(x', a')$ is the **TD target**, an estimate of the value given best action in the next state x'.

The update can also be written using the **TD error**, δ_k:

$$\delta_k = \tilde{r} + \beta \max_{a' \in \Gamma(x')} q_k(x', a') - q_k(x, a)$$

$$q_{k+1}(x, a) \leftarrow q_k(x, a) + \eta \delta_k$$

Q-learning iteratively improves the Q-function estimates until they converge to the optimal values q^*.

TD(0) Algorithm

Algorithm 1: Q-Learning

Input: Learning rate $\eta \in (0, 1]$, discount factor $\beta \in (0, 1)$

Initialize $q(x, a) = 0$ for all $x \in \mathsf{X}$, $a \in \Gamma(x)$
for *each episode* **do**
  Initialize state x
  while x *is not terminal* **do**
    Choose $a \in \Gamma(x)$ from q (e.g., ε-greedy)
    Take action a, observe reward r and next state x'
    $y \leftarrow \tilde{r} + \beta \max_{a' \in \Gamma(x')} q(x', a')$ `// TD target`
    $q(x, a) \leftarrow (1 - \eta)\, q(x, a) + \eta y$
    $x \leftarrow x'$ `// Transition to next state`
return q

Connection to Stochastic Approximation

The Q-learning update rule can be viewed through the lens of **stochastic approximation**.

Recall the Bellman equation for q^*:

$$q^*(x, a) = r(x, a) + \beta \int \left[\max_{a' \in \Gamma(x')} q^*(x', a') \right] P(x, a, dx')$$

Let $(Tq)(x, a)$ denote the right-hand side (the optimal Bellman operator applied to q). Finding q^* is equivalent to finding the fixed point $q = Tq$, or finding the root of the function $F(q) = q - Tq = 0$.

The TD update uses a noisy sample of the target value:

- $\tilde{r}$ is a sample of the immediate reward.
- x' is a sample from the transition distribution $P(x, a, \cdot)$.
- $\max_{a'} q_k(x', a')$ uses the current estimate q_k.

The TD target $y_k = \tilde{r} + \beta \max_{a'} q_k(x', a')$ is a *stochastic estimate* of $(Tq_k)(x, a)$.

The update $q_{k+1}(x,a) \leftarrow q_k(x,a) + \eta(y_k - q_k(x,a))$ is a step in the direction of the sampled residual $y_k - q_k(x,a)$, which is a noisy estimate of $(Tq_k)(x,a) - q_k(x,a)$.

This resembles algorithms like Robbins-Monro for finding roots $G(\theta) = 0$ via $\theta_{k+1} = \theta_k - \eta_k \hat{G}(\theta_k)$, where $\hat{G}$ is a noisy observation of G (e.g., $\hat{G} = G(\theta) + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ where h is a deterministic function).

(Surely we might use a newton/SGD-like method, but I think it is not popular as we do not know the directional derivative of the TD error.)

In this paper ...

As we will see soon, the authors propose a Q-learning algorithm to solve Epstein-Zin preferences.

Issues?

- If it is not time-consistent, estimating q^* is not meaningful as it is a moving target, the optimal policy might not stationary.
- Can we generalize the "linear" expectation in the TD target to a non-linear expectation? i.e., replacing the

$$q^*(x,a) = r(x,a) + \beta \mathbb{E}_{x'} \left[\max_{a' \in \Gamma(x')} q^*(x',a') \right]$$

to a certainty equivalent operator?

A brief introduction to Filtering

Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

- Ω: The set of all possible outcomes (sample space).
- $\mathcal{F}$: A σ-algebra on Ω, representing the set of all possible events.
- $\mathbb{P}$: A probability measure on $(\Omega, \mathcal{F})$.

A **filtration** is a sequence of sub-σ-algebras $\{\mathcal{F}_t\}_{t \in \mathcal{T}}$ of $\mathcal{F}$, indexed by time $t \in \mathcal{T}$ (e.g., $\mathcal{T} = \{0, 1, 2, \dots\}$), such that:

$$\mathcal{F}_s \subseteq \mathcal{F}_t \quad \text{for all } s \leqslant t.$$

Intuitively, $\mathcal{F}_t$ represents the information accumulated up to time t. Events in $\mathcal{F}_t$ are those whose occurrence (or non-occurrence) is known by time t.

Intuition (Can be bad!)

- A filtration $\{\mathcal{F}_t\}_{t\in\mathcal{T}}$ can be viewed as representing the flow of information over time.
- For any $A \in \mathcal{F}_t$, we can determine whether A has occurred by time t.
- The condition $\mathcal{F}_s \subseteq \mathcal{F}_t$ for $s \leqslant t$ expresses that information is never lost – what is known at time s remains known at all future times $t \geqslant s$.
- Given a stochastic process $\{X_t\}_{t\in\mathcal{T}}$, we can define its **natural filtration** as: $\mathcal{F}_t^X = \sigma(X_s : s \leqslant t)$, which is the σ-algebra generated by the random variables $X_0, X_1, \ldots, X_t$.

Adapted Stochastic Processes: Definition

- Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a filtration $\{\mathcal{F}_t\}_{t \in \mathcal{T}}$
- A stochastic process $\{X_t\}_{t \in \mathcal{T}}$ with real-valued random variables is **adapted** to filtration $\{\mathcal{F}_t\}$ if:
  - For every $t \in \mathcal{T}$, the random variable $X_t : \Omega \to \mathbb{R}$ is $\mathcal{F}_t$-measurable
  - **Definition of $\mathcal{F}_t$-measurability:** X_t is $\mathcal{F}_t$-measurable if and only if:

$$X_t^{-1}(B) = \{\omega \in \Omega : X_t(\omega) \in B\} \in \mathcal{F}_t$$

for every Borel set $B \in \mathcal{B}(\mathbb{R})$

Intuition (Can be bad?)

- $\mathcal{F}_t$ represents all information available up to time t
- If X_t is $\mathcal{F}_t$-measurable, then:
  - The value of X_t can be completely determined using only information in $\mathcal{F}_t$
  - No future information (beyond time t) is needed to determine X_t
  - For any Borel set B, the event $\{X_t \in B\}$ is in $\mathcal{F}_t$
- Adaptation expresses the principle of **non-anticipativity** – a process cannot "see into the future"

Example: Consider a stock price S_t at time t. Let $\mathcal{F}_t$ be the information available up to time t, which includes all past stock prices $S_0, S_1, \ldots, S_t$. Then the stochastic process $\{S_t\}$ is adapted to the filtration $\{\mathcal{F}_t\}$ because S_t is known at time t, and thus is $\mathcal{F}_t$-measurable.

The process $\{S_t\}_{t\in\mathcal{T}}$ is adapted to this filtration because:

- In the example, define a natural filtration:

$$\mathcal{F}_t = \sigma(S_0, S_1, \ldots, S_t)$$

- Each S_t is $\mathcal{F}_t$-measurable by construction, we have

$$S_t^{-1}(B) \in \mathcal{F}_t \quad \forall, B \in \mathcal{B}(\mathbb{R}_+), \tag{2}$$

- So S_t is adapted to $(\mathcal{F}_t)_{t\in\mathcal{T}}$.

Dynamic Utility

The paper first discusses the dynamic utility functional for Epstein-Zin preferences.

Usually we consider the state space to be a set of random variables, but in dynamic utility, we consider reward (utility) to be random variables.

Benefit: We do not assume utility function is given or invariant across time.

Let $\mathcal{T}$ be the set of time periods. Consider a filtration $\{\mathcal{F}_t\}_{t\in\mathcal{T}}$ of the state space X, such that $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$ for all $t \in \mathcal{T}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Let $\{\mathcal{Y}_t\}_{t\in\mathcal{T}}$ be a space of bounded p-integrable $\mathcal{F}_t$-measurable random variables.

We denote the tail sequence $\mathcal{Y}_{t,T} = \times_{s=t}^{T} \mathcal{Y}_s$.

Loosely speaking, we can consider $Y := (Y_{t_1}, \ldots, Y_T) \in \mathcal{Y}_{t_1,T}$ as sequences of **rewards** obtained over time from following a feasible policy σ_Y, i.e.,

$$Y_\tau = r(X_\tau, \sigma_Y(X_\tau)) \tag{3}$$

for all $\tau = t_1, \ldots, T$, where X_τ is a random variable representing the state at time τ.

Definition 1: Dynamic Utility Function

A **dynamic utility function** is a sequence $\{\tilde{v}_{t,T}\}_{t\in\mathcal{T}}$ where each utility

$$\tilde{v}_{t,T} : \mathcal{Y}_{t,T} \to \mathcal{Y}_t \tag{4}$$

is $\mathcal{F}_t$-measurable and satisfies the **monotonicity property**:

$$\tilde{v}_{t,T}(Y) \leqslant \tilde{v}_{t,T}(Z) \quad \forall\, Y, Z \in \mathcal{Y}_{t,T} \tag{5}$$

such that $Y_\tau \leqslant Z_\tau, \forall\, t \leqslant \tau \leqslant T$ almost surely.

- $\tilde{v}_{t,T}(Y) := \tilde{v}_{t,T}(Y_t, \ldots, Y_T)$ can be consisdered as **continuation value** contingent on the future stream of rewards $(Y_t, \ldots, Y_T)$ under strategy σ_Y.

Definition 2: Time Consistency

A dynamic utility function $\{\tilde{v}_{t,T}\}_{t\in\mathcal{T}}$ is said to be **time-consistent** iff for any sequence $Y, Z \in \mathcal{Y}_{t,T}$ and any $t_1, t_2 \in \mathcal{T}$ such that $0 \leqslant t_1 < t_2 \leqslant T$,

- $Y_\tau = Z_\tau, \quad \forall \tau = t_1, \ldots, t_2 - 1$
- $\tilde{v}_{t_2,T}(Y_{t_2}, \ldots, Y_T) \leqslant \tilde{v}_{t_2,T}(Z_{t_2}, \ldots, Z_T)$

implies that

$$\tilde{v}_{t_1,T}(Y_{t_1}, \ldots, Y_T) \leqslant \tilde{v}_{t_1,T}(Z_{t_1}, \ldots, Z_T) \tag{6}$$

Theorem: Time Consistency Characterization

Below is a theorem from Ruszczyński (2010).

Theorem 2: Time Consistency Characterization

Let $\{\tilde{v}_{t,T}\}_{t\in\mathcal{T}}$ be a dynamic utility function satisfying:

- $\tilde{v}_{t,T}(Y_t, Y_{t+1}, \ldots, Y_T) = Y_t + \tilde{v}_{t,T}(0, Y_{t+1}, \ldots, Y_T)$
- $\tilde{v}_{t,T}(0, 0, \ldots, 0) = 0$

for any $Y \in \mathcal{Y}_{t,T}$, $t \in \mathcal{T}$.

Then $\{\tilde{v}_{t,T}\}_{t\in\mathcal{T}}$ is **time-consistent** iff for any $0 \leqslant t_1 \leqslant t_2 \leqslant T$ and $Y \in \mathcal{Y}_{0,T}$ we have:

$$\tilde{v}_{t_1,T}(Y_{t_1}, \ldots, Y_T) = \tilde{v}_{t_1,t_2}(Y_{t_1}, \ldots, Y_{t_2-1}, \underbrace{\tilde{v}_{t_2,T}(Y_{t_2}, \ldots, Y_T)}_{\text{Continuation Value at } t_2})$$

This is easy to show from Definition 2.

Lemma 1: Recursive Relationship

A consequence of Theorem 2 is the recursive relationship:

$$\begin{aligned}\tilde{v}_{t,T}(Y_t, \ldots, Y_T) = Y_t + \tilde{v}_t(Y_{t+1} + \tilde{v}_{t+1}(Y_{t+2} + \ldots \\ + \tilde{v}_{T-2}(Y_{T-1} + \tilde{v}_{T-1}(Y_T))\ldots))\end{aligned} \tag{7}$$

where $\tilde{v}_t : \mathcal{Y}_{t+1} \to \mathcal{Y}_t$ is defined as:

$$\tilde{v}_t(Y) = \tilde{v}_{t,t+1}(0, Y), \quad \text{for } Y \in \mathcal{Y}_{t+1}.$$

Next, the paper wants to show that Epstein-Zin preferences can be written recursively.

Idea:

$$\text{Time Consistency} \iff \underbrace{\text{Dynamic Utility Function}}_{\text{Monotonicity}} + \text{Recursive Structure}$$

Epstein-Zin Utility

The following is a definition in DP1:

Definition 3: Epstein–Zin Koopmans operator

The **Epstein–Zin Koopmans operator** is defined as:

$$(Kv)(x) = \left[(1-\beta)\, c(x)^{\alpha} + \beta \left(\int v(x')^{\gamma} P(x, dx') \right)^{\alpha/\gamma} \right]^{1/\alpha}$$

- Why can we write it recursively?
- Will this recursive property holds outside of Markovian setting?

The paper gets complicated below so I will reserve it next time if there is an interest.

Aggregator Function View

Below defines the aggregator function W_{u_ρ} in the paper, which is in the similar spirit as the EZ Koopmans operator.

Definition 4: Aggregator Function

Let $u_\rho(c) = c^\rho$ be the utility of consumption c, with inverse $u_\rho^{-1}(y) = y^{1/\rho}$. The **aggregator function** $W_{u_\rho} : C_t \times C_{t+1} \to C_t$ combines current consumption utility C_t and future utility C_{t+1} (expressed in consumption units):

$$W_{u_\rho}(C_t, C_{t+1}) = u_\rho^{-1}\left[(1-\beta)\,u_\rho(C_t) + \beta\,u_\rho(C_{t+1})\right] = \left[(1-\beta)\,C_t^\rho + \beta\,C_{t+1}^\rho\right]^{1/\rho}$$

Epstein-Zin Dynamic Utility Function

Define $V_{t,T} = \tilde{v}_{t,T}(C_t, \ldots, C_T)$ be the dynamic utility function, and $\tilde{v}^{\alpha}(Y) = (\mathbb{E}_t[Y^{\alpha}])^{1/\alpha}$ (Kreps and Porteus certainty equivalent).

Definition 5: Epstein-Zin Dynamic Utility Function

The dynamic utility $\tilde{v}^{\alpha}_{t,T} : C_{t,T} \to C_t$ over a consumption stream $(C_t, \ldots, C_T)$ is defined recursively:

$$\tilde{v}^{\alpha}_{t,T}(C_t, \ldots, C_T) = W_{u_\rho}\left(C_t, \tilde{v}^{\alpha}_{t+1,T}(C_{t+1}, \ldots, C_T)\right)$$

where $\tilde{v}^{\alpha}_{t+1,T}(C_{t+1}, \ldots, C_T)$ can be seen as the Kreps and Porteus certainty equivalent of the continuation value $V_{t+1,T}$.

That is to say, the dynamic utility function is the aggregator function applied to the current consumption and the transformed continuation value.

Next lemma claims that EZ dynamic utility function satisfies the assumptions in Theorem 2.

Lemma 2: Properties of EZ Dynamic Utility

Let $\tilde{v}^{\psi}_{t,T} : \mathcal{Y}_{t,T} \to \mathcal{Y}_t$ be the corresponding dynamic utility function acting on the transformed space $\mathcal{Y}$ (where $Y_t = u_\rho(C_t)$ and $\psi = \alpha/\rho$). It satisfies:

$$\tilde{v}^{\psi}_{t,T}(Y_t, \ldots, Y_T) = \underbrace{(1-\beta)\, Y_t}_{\text{Current Utility}} + \underbrace{\beta \tilde{v}^{\psi}_{t,T}(0, Y_{t+1}, \ldots, Y_T)}_{\text{Continuation Value}}$$

$$\tilde{v}^{\psi}_{t,T}(0, \ldots, 0) = 0$$

The one-step certainty equivalent relationship is:

$$\tilde{v}^{\psi}_{t,t+1}(Y_t, Y_{t+1}) = (1-\beta)\, Y_t + \beta \tilde{v}^{\psi}_{t}(Y_{t+1})$$

where $\tilde{v}^{\psi}_{t}(Y_{t+1}) = \tilde{v}^{\psi}_{t,t+1}(0, Y_{t+1})$.

Epstein-Zin Dynamic Utility Function

Definition 6: Epstein-Zin Dynamic Utility Function

The dynamic utility $\tilde{v}^{\alpha}_{t,T} : C_{t,T} \to C_t$ over a consumption stream $(C_t, \ldots, C_T)$ is defined recursively:

$$\tilde{v}^{\alpha}_{t,T}(C_t, \ldots, C_T) = W_{u_\rho}\left(C_t, \tilde{v}^{\alpha}_{t+1,T}(C_{t+1}, \ldots, C_T)\right)$$

Theorem 3: Dynamic Epstein-Zin Utility Function

Let $Y_t = u_\rho(C_t)$ be the transformed rewards/consumption, forming sequences $Y_{t,T} \in \mathcal{Y}_{t,T}$. Let $\beta \in (0,1)$ and $\psi = \alpha/\rho > 0$. Then, the corresponding Epstein-Zin dynamic utility function $\tilde{v}^{\psi}_{t,T} : \mathcal{Y}_{t,T} \to \mathcal{Y}_t$

$$\tilde{v}^{\psi}_{t,T}(Y_t, \ldots, Y_T) = (1-\beta)\, Y_t + \beta \tilde{v}^{\psi}_{t,T}(0, Y_{t+1}, \ldots, Y_T),$$

satisfies the **monotonicity property** and is **time-consistent**.

Connecting Koopmans Operator to Q-Learning

Recall the EZ Koopmans operator for the value function $v(x)$:

$$(Kv)(x) = \left[(1-\beta)\,c(x)^{\alpha} + \beta\left(\mathbb{E}[v(x')^{\gamma}]\right)^{\alpha/\gamma}\right]^{1/\alpha}$$

We want a Bellman equation for the optimal Q-function $q^*(x, a)$ (value of taking action a then acting optimally).

- The immediate "consumption" is the reward $r(x, a)$.
- The optimal value from the next state onwards is $v^*(x') = \max_{a' \in \Gamma(x')} q^*(x', a')$.

Replacing $c(x)$ with $r(x, a)$ and $v(x')$ with $\max_{a'} q^*(x', a')$ in the Koopmans structure gives a Bellman equation for q^* in consumption units.

EZ Bellman Equation for Q-functions

Let $q^*(x, a)$ be the optimal Q-function in consumption units. The Bellman equation analogous to the Koopmans operator is:

$$q^*(x,a) = \left[(1-\beta)r(x,a)^\alpha + \beta\left(\mathbb{E}_{x'}\left[\left(\max_{a'\in\Gamma(x')} q^*(x',a')\right)^\gamma\right]\right)^{\alpha/\gamma}\right]^{1/\alpha}$$

where

- $r(x, a)^\alpha$ is the utility of the immediate reward.
- $\mathbb{E}_{x'}[\cdot] = \int(\cdot)P(x, a, dx')$.
- The term $\mathbb{E}_{x'}[(\max_{a'} q^*(x', a'))^\gamma]$ is the expected γ-power of the optimal value from the next state.
- The outer structure mirrors the Koopmans operator.

EZ TD(0) Update

Based on (33), the TD target for $q_k(x, a)$ after observing (x, a, r, x') involves estimating the expectation. Let $G(x, a; \theta)$ be a stochastic approximation of the expectation term:

$$G(x, a; \theta) \approx \mathbb{E}_{x'} \left[\left(\max_{a' \in \Gamma(x')} q_k(x', a') \right)^{\gamma} \right]$$

The approximate TD target $\hat{y}_k$ is:

$$\hat{y}_k = \left[(1 - \beta) r^{\alpha} + \beta G(x, a; \theta)^{\alpha/\gamma} \right]^{1/\alpha}$$

The EZ TD(0) update rule is:

$$q_{k+1}(x, a) \leftarrow (1 - \eta) q_k(x, a) + \eta \hat{y}_k$$

Remarks

Missing in the paper:

- Under what condition does EZ Q-learning converges to the optimal Q-function (perhaps global stability)?

References

Andrzej Ruszczyński. Risk-averse dynamic programming for markov decision processes. *Mathematical programming*, 125:235–261, 2010.