Prospect Theoretic Q-Learning

Siddharth Chandak, 17D070019
Guide - Prof. Vivek S. Borkar

17$^{\text{th}}$ December 2020

Electrical Engineering
IIT Bombay

Outline

- Introduction
- Modified Q-Learning Scheme
- Convergence
- Equilibrium Points
- Numerical Experiments
- Alternate Formulation

Introduction

Reinforcement Learning

- Reinforcement Learning: Actions taken by a rational agent in order to maximize its expected rewards
- Typically modeled using Markov Decision Processes
- Useful and well-developed model for human decision making
- Economics, Control theory, Robotics and Games

Markov Decision Processes

- Consider finite state space S and finite action space A
- At each time step n, agent chooses action $Z_n \in A$ when it is in state $X_n \in S$
- Markov control policy:

$$P(X_{n+1} = j | X_m, Z_m, m \leq n) = p(j | X_n, Z_n) \; \forall n,$$

Q-Learning

- Q-learning: A reinforcement learning algorithm for MDPs
- Q-learning iteration:

$$Q_{n+1}(i,u) = Q_n(i,u) + a(n)I\{X_n = i, Z_n = u\} \times \left(k(i,u) + \alpha \max_a Q_n(X_{n+1}, a) - Q_n(i,u)\right)$$

- α: Discount factor for future rewards
- $a(n)$: Learning rate
- $k(i,u)$: Current reward
- Agent updates $Q(i,u)$ based on next state X_{n+1} and action a which is optimal for current estimate of Q-value

Convergence of Q-Learning Scheme

- Under appropriate conditions[1], $Q_n \to Q^*$ where Q^* is a solution of

$$Q(i,u) = k(i,u) + \alpha \sum_j p(j|i,u) \max_a Q(j,a),$$

- Q^* is the expected discounted reward of executing action u at state i
- Minimizer of $Q^*(i,:)$ yields an optimal choice of control in state i

[1]Stochastic approximation: a dynamical systems view-point by Vivek S. Borkar

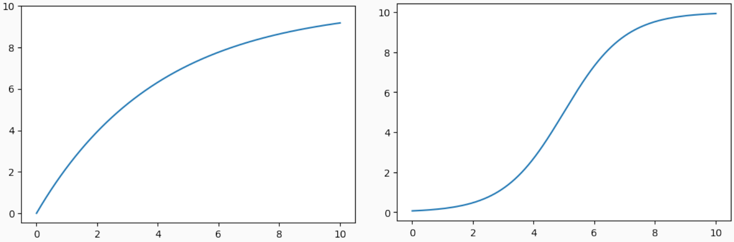
Risk and Prospect Theory

- Reinforcement Learning: Actions taken by a **rational** agent in order to maximize its expected rewards
- When faced with risk, humans don't always behave rationally
- Reinforcement Learning has been widely studied under risk-neutral and risk-averse policies
- But according to Prospect Theory, humans perceive risk differently in different scenarios

Prospect Theory

- Aims to model actual behavior of people
- A valuation map over gains and losses defined with respect to a reference point
- s-shaped valuation map:
  - Marginal impact of change in value diminishes with distance from the reference point
  - Concavity for gains contributes to risk aversion for gains
  - Convexity for losses contributes to risk seeking behavior

Example



(a) Risk-averse utility function

(b) Prospect theoretic valuation map

Figure 1

Motivation

- We study classical Q-learning from a prospect theoretic viewpoint
- Future returns are distorted using a s-shaped valuation map
- Previous works[2] applying such prospect theoretic valuation maps worked with certain restricting assumptions
    - Doesn't allow steep valuation maps and high discount factors for future rewards

[2] Shen et al., Risk-sensitive reinforcement learning, 2014

Modified Q-Learning Scheme

Q-Learning Iteration

- Prospect theoretic Q-learning iteration:

$$\begin{aligned} Q_{n+1}(i,v) &= Q_n(i,v) + a(n)I\{X_n = i, Z_n = v\}\Big(k(i,v) + \\ &\alpha u(Q_n(X_{n+1}, Z_{n+1}) - \xi_n(X_{n+1}, Z_{n+1})) - Q_n(i,v)\Big) \end{aligned}$$

- $\{X_n\}$: Controlled Markov chain on a finite state space S, $|S| = s$
- $\{Z_n\}$: Control Process in a finite action space A, $|A| = r$
- $\alpha \in (0,1)$: Discount factor
- $a(n) \in [0,1]$: Positive learning rate
- $k > 0$: The running reward
- $u(\cdot)$: s-shaped strictly increasing continuously differentiable map

Parameters (cont.)

- Noise:
  - $\{\xi_n = [[\xi_n(i,v)]]\}$: $\mathcal{R}^{sr}$-valued zero mean i.i.d. noise
  - Each $\xi_n(i,v)$ is distributed according to a continuously differentiable density $\varphi(\cdot)$ concentrated on a finite interval $[-c, c]$
  - $c \in [0, k_{min}]$ where $k_{min} = \min_{i,v} k(i,v)$
- Choice of Z_{n+1}:
  - Need ϵ-randomization to ensure adequate exploration
  - Use epsilon-greedy policy:

$$Z_{n+1} = \begin{cases} w^*_{n+1} & \text{w.p. } (1-\epsilon) \\ w \neq w^*_{n+1} & \text{w.p. } \frac{\epsilon}{r-1} \text{ each} \end{cases}$$

  - $w^*_{n+1} = \arg\max_w (Q_n(X_{n+1}, w) - \xi_n(X_{n+1}, w))$
- Define $K := \frac{k_{max}}{1-\alpha}$ where $k_{max} = \max_{i,v} k(i,v)$
- $u : [0, K+c] \mapsto [0, K]$.

Boundedness

Lemma 2.1

When initiated in the set $\mathcal{S} := [k_{min}, K]^{sr}$, the Q-learning iteration stays in the set $\mathcal{S}$.

Proof (Outline):

- Note that $Q_{n+1}(i, v)$ can be written as the convex combination of $Q_n(i, v)$ and U
  where $U := k(i, v) + \alpha u(Q_n(X_{n+1}, Z_{n+1} - \xi_n(X_{n+1}, Z_{n+1}))$

$$Q_{n+1}(i, v) = \Big(1 - a(n)I\{X_n = i, Z_n = v\}\Big)Q_n(i, v) + a(n)I\{X_n = i, Z_n = v\}U$$

- U can be bounded as follows:

$$k_{min} \leq U \leq k_{max} + \alpha u(K + c) = k_{max} + \alpha K = K$$

- $Q_n \in \mathcal{S} \Rightarrow Q_{n+1} \in \mathcal{S}$

Convergence

Limiting O.D.E.

- Need the following restriction on $a(n)$:

$$\sum a(n) = \infty, \sum a(n)^2 < \infty$$

- Since $u(\cdot)$ is Lipschitz continuous and $\sup_n \|Q_n\|_\infty \leq K < \infty$, the Q-learning iteration converges to the following o.d.e.:

$$\begin{aligned}
&\frac{d}{dt} q_t(i,v) = h_{i,v}(q_t) \\
&= \quad F_{i,v}(q_t) - q_t(i,v) \\
&:= \quad k(i,v) + \alpha \int_{\mathcal{R}^{sr}} \Bigg(\sum_j p(j|i,v) \Big((1-\epsilon) \max_w \big(u(q_t(j,w) - y_{j,w}) \big) \\
&+ \quad \frac{\epsilon}{r-1} \sum_{w \neq w^*_{q_t,y,j}} \big(u(q_t(j,w) - y_{j,w}) \big) \Big) \Bigg) \prod_{j,w} \varphi(y_{j,w}) dy_{j,w} - q_t(i,v).
\end{aligned}$$

- where $w^*_{q_t,y,j} = \arg\max_w (q_t(j,w) - y_{j,w})$.

Properties of O.D.E.

- h and F are continuously differentiable
- Jacobian matrix of h (resp., F) at q is $J(q) - I$ (resp., $J(q)$):

$$
\begin{aligned}
J(q)_{(i,v),(j,w)} =& p(j|i,v)\alpha \\
&\times \int \Bigg[\Big((1-\epsilon) u'(q(j,w) - y_{j,w}) \mathbb{1}_{q,j,w} \\
&+ \frac{\epsilon}{r-1} u'(q(j,w) - y_{j,w}) \big(1 - \mathbb{1}_{q,j,w}\big) \Big) \\
&\times \prod_w \varphi(y_{j,w}) dy_{j,w} \Bigg]
\end{aligned}
$$

  - where $\mathbb{1}_{q,j,w} = 1$ if $q(j,w) - y_{j,w} > q(j,w') - y_{j,w'}\ \forall\ w' \neq w$ and 0 otherwise

Cooperative O.D.E.

Definition 3.1

(Cooperative o.d.e) An o.d.e. of the form $\dot{x} = h(x(t))$ is a cooperative o.d.e. if the Jacobian matrix for h is irreducible and

$$\frac{\partial h_i}{\partial x_j} \geq 0, \;\; j \neq i.$$

Cooperative O.D.E.

Lemma 3.1

When the controlled Markov chain is irreducible, $J(q)$ (the Jacobian of F) is a non-negative irreducible matrix and the limiting o.d.e. is a cooperative o.d.e.

Proof (Outline):

- $u' > 0$ implies that $J(q)$ is a non-negative matrix
- $J(q) = P \times J_1(q)$
  - where $P_{(i,v),(j,w)} = p(j|i, v)$
  - and $J_1(q)$ is a positive diagonal matrix with $J_1(q)_{(j,w),(j,w)}$ being α times the integral in the Jacobian
- Since the Markov chain is irreducible, the matrix P is irreducible and hence, the matrix $J(q)$ will be irreducible

□

Boundedness

Lemma 3.2

When initiated in the set $\mathcal{S} := [k_{min}, K]^{sr}$, the limiting o.d.e. stays in the set $\mathcal{S}$.

Proof (Outline):

- The derivative of $q_t(i,v)$ can be bounded using:

$$k_{min} - q_t(i,v) \leq \frac{d}{dt} q_t(i,v) \leq k_{max} + \alpha u(K + c) - q_t(i,v)$$

- Discretization:

$$a_n k_{min} + (1 - a_n) q_n(i,v) \leq q_{n+1}(i,v) \leq a_n K + (1 - a_n) q_n(i,v)$$

- If initiated in the set $\mathcal{S} := [k_{min}, K]^{sr}$, q_n (and by its limit, the o.d.e.) stays in the set $\mathcal{S}$

$\square$

Monotone Dynamical Systems

- The Markov chain is irreducible and the iteration is initiated in the set $\mathcal{S}$.
- The o.d.e. is cooperative (Lemma 3.1) and it stays within the set $\mathcal{S}$ (Lemma 3.2)

Theorem 3.1

For initial conditions in an open dense set, the solutions of (1) converge to an equilibrium. [3]

- The same is true for the iterates of the discrete map $\Phi : \mathcal{S} \mapsto \mathcal{S}$ which maps q_0 to q_1
- Since the o.d.e. is cooperative, this map is monotone
- Also, order compact (maps each order interval to a bounded set)

[3] Hirsch, Smith. Competitive and cooperative systems: A mini-review, 2003

Monotone Dynamical Systems

Theorem 3.2

There exist maximal and minimal equilibria q^*, q_* *resp., such that any other equilibrium* $\hat{q}$ *satisfies* $q_* \leq \hat{q} \leq q^*$ *componentwise.* [4]

- $q_0 \geq q^* \implies q_t \to q^*$ and likewise, $q_0 \leq q_* \implies q_t \to q_*$
- If $q^* > q_*$, $q_* \leq q_0 \leq q^* \implies q_* \leq q_t \leq q^* \; \forall \; t \geq 0$ by monotonicity

[4] Hirsch, Smith. Monotone maps: a review, 2005

Theorem 3.3

At least one of the following holds: [5]

1. $\exists$ *a third equilibrium* $\hat{q}, q_* < \hat{q}, < q^*$,
2. $\exists$ *a trajectory* q_t *of (1) such that* $q_t \uparrow q^*$ *as* $t \uparrow \infty$ *and* $q_t \downarrow q_*$ *as* $t \downarrow -\infty$,
3. $\exists$ *a trajectory* q_t *of (1) such that* $q_t \downarrow q_*$ *as* $t \uparrow \infty$ *and* $q_t \uparrow q^*$ *as* $t \downarrow -\infty$.

Corollary 3.3.1

For stable q_* *and* q^*, *there is at least one more equilibrium* $\hat{q}$ *such that* $q_* < \hat{q} < q^*$.

[5] Hirsch, Smith. Monotone maps: a review, 2005

Equilibrium Points

Perron-Frobenius Theorem

- The stability of the equilibria of the Q-learning scheme, which are the same as equilibria of the differential equation can be analyzed by looking at the eigenvalues of its Jacobian matrix $J(q) - I$ evaluated at the equilibrium

Theorem 4.1

(Perron-Frobenius Theorem) Let A be a square non-negative irreducible matrix. Then

1. A *has a real positive eigenvalue* λ_A *and* λ_A *is strictly greater than the absolute value of any other eigenvalue of* A.
2. $r \leq \lambda_A \leq R$ *where* $r = \min_i r_i$ *and* $R = \max_i r_i$ *and* r_i *denotes the sum of the elements of row* i *of* A.

Bounds on Eigenvalues

- $\Gamma(q)_{i,v}$: Sum of the $(i,v)^{\text{th}}$ row of $J(q)$
- $\Gamma(q)^* = \max_{i,v} \Gamma(q)_{i,v}$ and similarly $\Gamma(q)_* = \min_{i,v} \Gamma(q)_{i,v}$
- Let λ^* be the Frobenius eigenvalue of $J(q)$, then $\Gamma(q)_* \leq \lambda^* \leq \Gamma(q)^*$
- For any eigenvalue λ of $J(q)$, $\lambda - 1$ is an eigenvalue of the Jacobian $J(q) - I$
- Real part of all eigenvalues of $J(q) - I$ are less than $\lambda^* - 1$

Comments

- $u(\cdot)$ is a s-shaped function
- $u'(x) < 1 < \frac{1}{\alpha}$ for low and high values of x and can exceed $\frac{1}{\alpha}$ in the mid-range
- If $u'(x) < \frac{1}{\alpha} \forall x \in [0, K + c]$, then we can use the results by Shen et al., which show that there will exist only one equilibrium point in the set and will be stable
- We consider the case where $u'(x)$ exceeds $\frac{1}{\alpha}$ in the middle region
- Define points a, b in $[0, K]$ as the largest and smallest points in $[0, K]$ such that $u'(x) < \frac{1}{\alpha} \forall x \in [0, a) \cup (b, K + c]$

Example

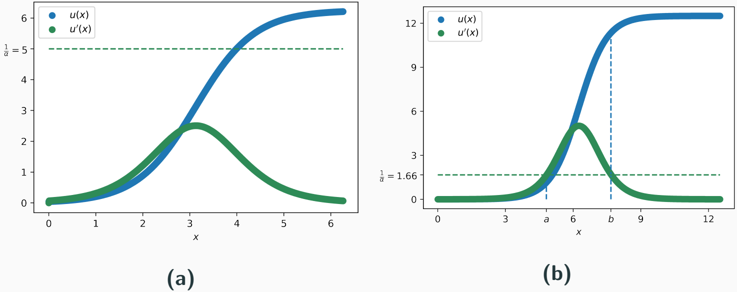


Figure 2: Examples of s-shaped valuation maps: (a) shows the case where $u'(x) < \frac{1}{\alpha} \forall x \in [0, K + c]$ and (b) depicts a and b in a case where $u'(x)$ exceeds $\frac{1}{\alpha}$ in the middle region

Stable Regions

Theorem 4.2

There can be at most one equilibrium point in the set $(b + c, K]^{sr}$ *and if such an equilibrium point exists, it will be a stable equilibrium and the maximal equilibrium point. Similarly, there can be at most one equilibrium point in the set* $[k_{min}, a - c)^{sr}$ *and if such an equilibrium point exists, it will be a stable equilibrium and the minimal equilibrium point.*

Proof (Outline):

- Stability:
  - For any point in these sets, sum of elements in each row is less than 1
  - Hence, $\lambda^* < 1$ and hence, real part of all eigenvalues of the Jacobian $J(q) - I$ are negative
  - Any equilibrium point lying in this region will be stable.

Stable Regions (cont.)

Proof (cont.):

- Suppose that there are two equilibria q_1, q_2 in $(b+c, K]^{sr}$
- They can be ordered or unordered
- First consider the case where they are ordered and $q_1 < q_2$:
  - There exists another equilibrium point between any two stable equilibria so $\exists\ q_3$, another equilibrium point such that $q_1 < q_3 < q_2$ (Corollary 3.3.1)
  - q_3 will also be a stable equilibrium and hence there will be more stable equilibrium points between q_1, q_3, and between q_3, q_2
  - Repeated application of this argument implies that we will have a curve of non-isolated equilibria
  - Real part of all eigenvalues of the Jacobian $J(q) - I$ are negative in this region implying all equilibria are isolated giving us a contradiction

Stable Regions (cont.)

Proof (cont.):

- Now consider the case where they are unordered:
  - There exists q^* such that all equilibrium points q satisfy $q \leq q^*$ (Theorem 3.2)
  - Since, no ordering exists between q_1 and q_2, they can't be equal to q^*
  - So, $q_1 < q^*$ where both q_1 and q^* lie in this region. But we have shown earlier that there cannot exist ordered equilibria in the region.

□

We subsequently refer to the sets $[k_{min}, a - c)^{sr}$ and $(b + c, K]^{sr}$ as the **lower** and **upper stable regions** respectively.

- Let points d, e in $[0, K]$ be the smallest and largest points in $[0, K]$ such that $u'(x) > \frac{1}{\alpha} \forall x \in (d, e)$.

Theorem 4.3

Any equilibrium point in the region $(d + c, e - c)^{sr}$ is an unstable equilibrium point.

Proof (Outline):

- $\lambda^* > 1$
- At least one eigenvalue has a poisitive real part and hence, any equilibrium point in this region will be unstable

□

Theorem 4.4

If all equilibrium points are hyperbolic and $u(x)$ is convex and concave in the regions $x < m_1$ and $x > m_1$ respectively, then there can exist at most one stable equilibrium point in the region $[k_{min}, m_1 - c)^{sr}$. Similarly in the region $(m_1 + c, K]^{sr}$, there can exist at most one stable equilibrium. If these exist then they will be the minimal and maximal equilibrium points respectively.

- This theorem can also be applied where the valuation map is a traditional utility function
- In our case, there can exist many other stable equilibrium points with some components below and some above m_1

Numerical Experiments

Parameters

- $u(\cdot)$:

$$u(x) = \frac{L}{1 + e^{-\gamma(x - x_0)}}$$

- State and Action Space: Values of s and r ranged from 2 to 100
- $a(n)$:

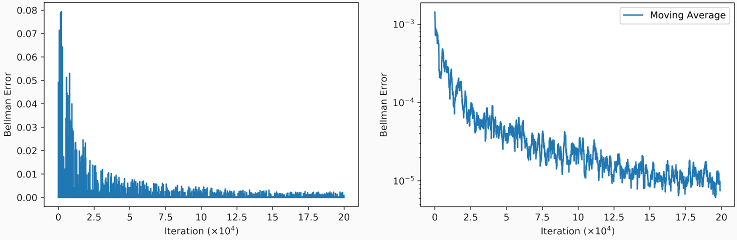
$$a(n) = \frac{1}{\lceil \frac{n}{100} \rceil}$$

- k :Randomly generated in a given range set by fixing k_{min} & k_{max}
- Noise: Cosine distribution with $c \approx 0.01$
- α : Varied from 0.01 to 0.99
- Transition matrix: Randomly generated
- $\epsilon = 0.05$

Convergence

- Q-learning iteration and the o.d.e. converged to an equilibrium point and to the same point when initiated at the same point
- Values of s and r (size of state and action space), $a(n)$ and ϵ have an impact on the rate of convergence but do not observably affect the equilibrium points
- Plots of Bellman error $(|Q_{n+1}(X_n, Z_n) - Q_n(X_n, Z_n)|)$ on next slide

Convergence Plots: Bellman Error



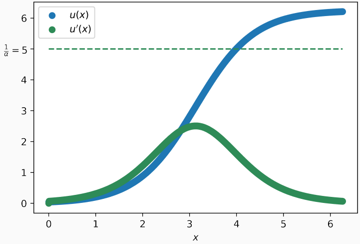
(a) **(b)**

Figure 3: Convergence plots: (a) shows the Bellman error plot for modified Q-learning scheme and (b) shows the moving average of the same over 1000 iterations

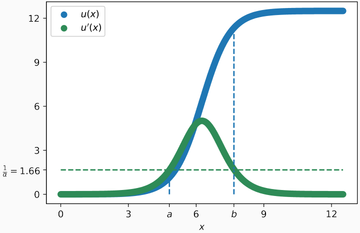
Observations

- As expected, when either α is too small or the function $u(\cdot)$ rises very gradually (i.e. $u'(x) < \frac{1}{\alpha}$ in the whole region), then there exists only one equilibrium point
- For very steep $u(\cdot)$, the iteration usually converges to one of the two equilibria, one each in the upper and lower zones, depending on the initiation

Observations



(a)



(b)

Figure 4: Only one equilibrium point exists in the case of (a), while we observe two equilibrium points for (b), one each in the upper and lower stable regions

Third Equilibrium Point

- In our initial experiments, we noticed that the iteration converged either to the maximal or to the minimal equilibrium point only
- To confirm the possibility of existence of a third equilibrium point:
    - Manually constructed and computed the equilibrium points for a small system $(s = 4, r = 2)$
    - Assigned the value 2 to all rewards (i.e., $k(i, u) = 2, \forall i, u$) for simplicity
    - The two actions were kept identical (i.e. $p(j|i, u) = p(j|i, v), \forall i, j$ where u, v are the two actions for state i)

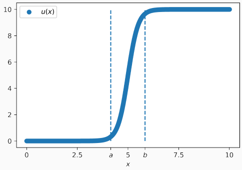


Figure 5

Third Equilibrium Point

- Observations for our constructed system:
  - Observed that there are 4 stable equilibrium points
  - Iteration converges to these additional equilibrium points when initiated in close vicinity to them
- Apart from this above constructed case, we never observed the Q-learning iteration to converge to these middle stable equilibrium points
- While 3 or more stable equilibria can exist for many systems, convergence to these points seems very infrequent

Alternate Formulation

Alternate Formulation

- In our original formulation, only the future returns are distorted using the prospect theoretic valuation map
- Now, the s-shaped curve $u(\cdot)$ is applied to the total returns i.e., both the current rewards and the future returns are distorted
- Q-learning iteration:

$$\begin{aligned} Q_{n+1}(i,v) &= Q_n(i,v) + a(n) I\{X_n = i, Z_n = v\} \Bigg(u\Big(k(i,v) + \\ &\quad \alpha (Q_n(X_{n+1}, Z_{n+1}) - \xi_n(X_{n+1}, Z_{n+1})) \Big) - Q_n(i,v) \Bigg) \end{aligned}$$

- $u : [0, K + \alpha c] \mapsto [0, K]$

Convergence

- When the Markov chain is irreducible and the iteration is initiated in $\mathcal{S}_1 := [0, K]^{sr}$, this formulation of Q-learning also converges

Stable Regions

- Upper stable region: $(b' + c, K]^{sr}$ where $b' = \frac{b-k_{min}}{\alpha}$. Exists if the following holds:

$$b' + c < K \Leftrightarrow \frac{b - k_{min}}{\alpha} + c < K \Leftrightarrow b < k_{min} + \alpha(K - c).$$

- Lower stable region: $[0, a' - c)^{sr}$ where $a' = \frac{a-k_{max}}{\alpha}$. Exists if the following holds:

$$a' - c > 0 \Leftrightarrow \frac{a - k_{max}}{\alpha} - c > 0 \Leftrightarrow a > k_{max} + \alpha c.$$

- They are more likely to exist for high values of α

Numerical Experiments

- Converges and exhibits trends similar to the original scheme
- An important difference:
  - Maximal equilibrium point of the alternate formulation is higher than the maximal equilibrium for the original formulation
  - Similarly, minimal equilibrium point of the alternate formulation is lower than the minimal equilibrium for the original formulation

Thank You!

Additional Results
(if time permits)

Existence of Equilibrium in Stable Regions

- $u_1(x) := k_{min} + \alpha u(x - c)$

Theorem 7.1

If $u_1(b + c) \geq b + c$, then there exists a stable maximal equilibrium point in the region $[b + c, K]^{sr}$ and any iteration initiated in this set will converge to this equilibrium point.

Theorem 7.2

If $u_1(a + c) > a + c$, then there exists only one equilibrium point in the set $[k_{min}, K]^{sr}$ and it will lie in the region $(b + c, K]^{sr}$.

Existence of Equilibrium in Stable Regions

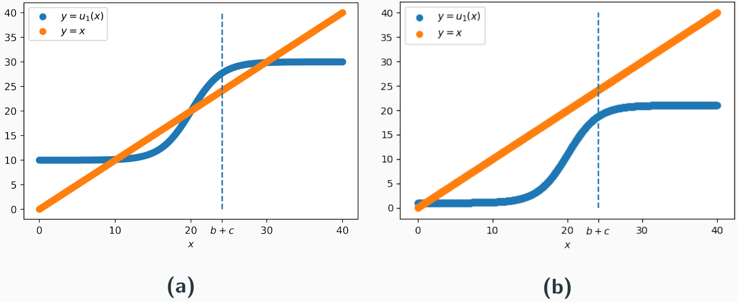


Figure 6: Theorem 7.1 only gives a sufficient condition: An equilibrium point exists in the upper stable region both (a) and (b)

Existence of Equilibrium in Stable Regions

- $u_2(x) := k_{max} + \alpha u(x + c)$

Theorem 7.3

If $u_2(a - c) \leq a - c$, then there exists a stable maximal equilibrium point in the region $[k_{min}, a - c]^{sr}$ and any iteration initiated in this set will converge to this equilibrium point.