Value iteration is optic composition

Jules Hedges, Riu Rodríguez Sakamoto

Mathematically Structured Programming group,
Department of Computer and Information Sciences,
University of Strathclyde

riu.rodriguez-sakamoto@strath.ac.uk

5th International Conference on Applied Category Theory,
Glasgow (Scotland)

July 21st, 2022

Towards Reinforcement Learning algorithms

- n-armed bandit problem: Exploration vs exploitation (greedy, ε-greedy)
- Contextual bandits: Many states, immediate rewards
- Reinforcement Learning problem: Many states, delayed rewards

Decision Processes

state space X — transition $f : X \times A \to X$

action space A — utility $U : X \times A \to \mathbb{R}$

Objective: Which actions to choose where, to optimize long-run reward?

$$\text{policy } \pi : X \to A$$

Long-run reward: Discounted[1] sum with infinite horizon

$$x_0 \xrightarrow{f(x_0,\pi(x_0))} x_1 \xrightarrow{f(x_1,\pi(x_1))} x_2 \longrightarrow \cdots$$

$$V_\pi(x_0) = \sum_{k=0}^{\infty} \beta^k U(x_k, \pi(x_k))$$

[1] Discount factor $\beta \in (0, 1)$

Dynamic Programming

Assign value functions to policies.
Start with some policy $\pi : X \to A$, and some value function $V : X \to \mathbb{R}$.
Bellman update steps:

- Value improvement: $V'(x) = U(x, \pi(x)) + \beta V(f(x, \pi(x)))$ (policy evaluation)
- Policy improvement: $\pi'(x) = \arg\max_{a \in A} U(x, a) + \beta V(f(x, a))$

Bellman optimality condition: Fixpoint of the update function.

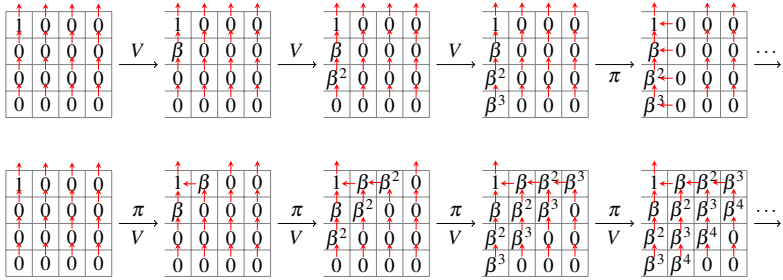
- Value improvement: $V(x) = U(x, \pi(x)) + \beta V(f(x, \pi(x)))$
- Policy improvement: $\pi(x) = \arg\max_{a \in A} U(x, a) + \beta V(f(x, a))$

Dynamic Programming

Classical DP algorithms:

- Policy Iteration: $((V \text{ impr.})^*; \pi \text{ impr.})^*$
- Value Iteration: $(V \text{ impr.}; \pi \text{ impr.})^*$
  - Fusion of both steps:

$$V(x) = \max_{a \in A} U(x, a) + \beta V(f(x, a))$$



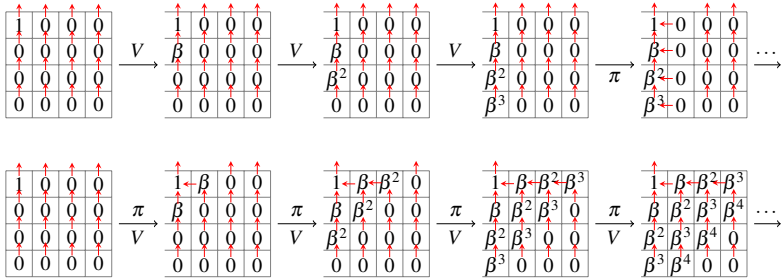
Markov Decision Processes

MDP: Some function is stochastic: Kleisli morphism of Δ (e.g. Gridworld with uncertainty):

state space X — transition $f : X \times A \to \Delta X$

action space A — utility $U : X \times A \to \Delta\mathbb{R}$

policy $\pi : X \to \Delta A$



Continuous spaces w/o stochasticity

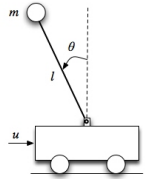
Continuous state or action space, e.g. savings model in economics:

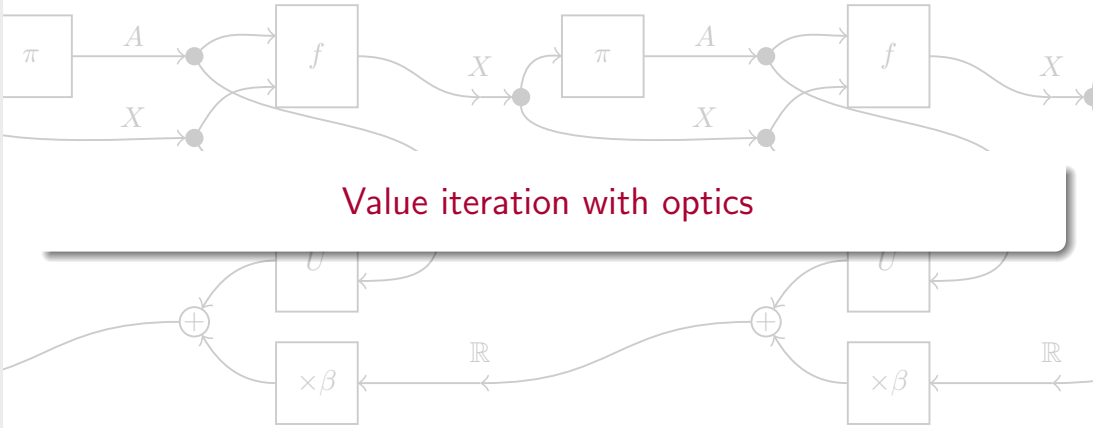
- $X = \mathbb{R}$: Assets (savings)
- $A = \mathbb{R}$: Consumption
- $(1 + \gamma)$: Gross rate of return
- i: Interest (constant)

- Policy: $\pi : x \mapsto c$
- Transition: $f : (x, c) \mapsto (1 + \gamma)(x - c + i)$
- Utility: $U : (\_, c) \mapsto U(c)$
  (w/ Inada or transversality condition)

$$V'(x) = \max_{0 \leq c \leq x+i} U(c) + \beta V((1 + \gamma)(x - c + i))$$

Also, the inverted pendulum control problem.
Continuous spaces *and* stochastic maps: Stoch $= \mathrm{Kl}(\mathrm{Giry})$ or some subcategory.





Value iteration with optics

Optics

$\mathcal{M}$ monoidal category

$\downarrow \bullet : \mathcal{M} \times \mathcal{C} \to \mathcal{C}$ action of $\mathcal{M}$ on $\mathcal{C}$

$\mathcal{C}$ category

Then $\mathcal{C}$ is a $\mathcal{M}$-actegory. Given $\mathcal{M}$ monoidal, and two $\mathcal{M}$-actegories $\mathcal{C}$, $\mathcal{D}$, the category of mixed optics $\mathrm{Optic}_{\mathcal{C},\mathcal{D}}$ has

- objects $\begin{pmatrix} X \\ X' \end{pmatrix} \in \mathcal{C} \\ \in \mathcal{D}$
- morphisms are coends

$$\mathrm{Optic}_{\mathcal{C},\mathcal{D}}\left(\begin{pmatrix} X \\ X' \end{pmatrix}, \begin{pmatrix} Y \\ Y' \end{pmatrix}\right) = \int^{M:\mathcal{M}} \underbrace{\mathcal{C}(X, M \bullet Y)}_{f} \times \underbrace{\mathcal{D}(M \bullet Y', X')}_{f'}$$

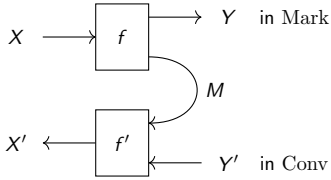
i.e. equivalence classes of (M, f, f').

Optics

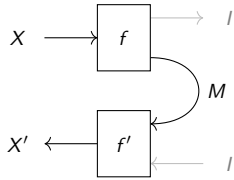
Let $\mathcal{C} = \mathcal{M} = \mathsf{Mark} = \mathrm{Kl}(\Delta)$, and $\mathcal{D} = (\mathsf{Conv}, \boxtimes, \{*\}) = \mathrm{EM}(\Delta)$, where $\boxtimes$ is similar to tensor product in R-Mod. Conv is monoidal closed under $\boxtimes$, where the function set $[X, Y]$ has pointwise convex structure.

$$\mathsf{Optic}_{\mathsf{Mark},\mathsf{Conv}}\left(\begin{pmatrix} X \\ X' \end{pmatrix}, \begin{pmatrix} Y \\ Y' \end{pmatrix}\right) = \int^{M:\mathsf{Mark}} \mathsf{Mark}(X, M \otimes Y) \times \mathsf{Conv}(\Delta M, [Y', X'])$$

String diagram syntax (informal for mixed case):



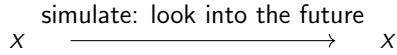
(a) Morphism

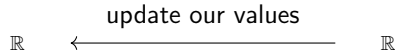


(b) Costate

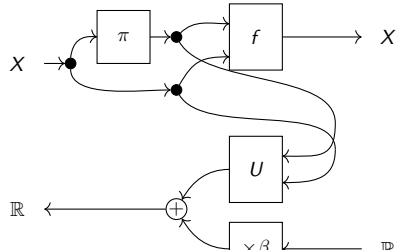
Optics for Value Iteration

Deterministic Decision Processes





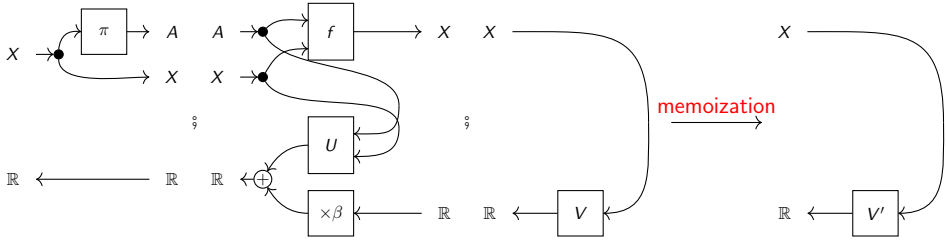
(a) Lenses are iteration steps



Optics for Value Iteration

Bellman update steps on a dynamical system with state space $A^X \times \mathrm{Optic}\left(\binom{X}{\mathbb{R}}, \binom{I}{I}\right)$:

- V improv: $(\pi, V) \mapsto (\pi, \bar{\pi} \fatsemi \lambda \fatsemi V)$



- π improv: $(\pi, V) \mapsto (x \mapsto \arg\max_{a \in A}(\lambda \fatsemi V)(x, a), V)$

Optics for Value Iteration

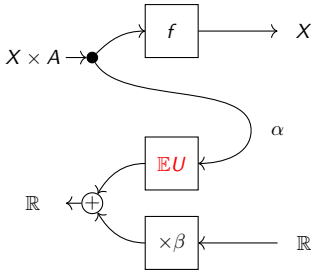
Mixed optics for MDPs

$$\lambda : \begin{pmatrix} X \times A \\ \mathbb{R} \end{pmatrix} \to \begin{pmatrix} X \\ \mathbb{R} \end{pmatrix} \qquad \text{in } \mathsf{Optic}_{\mathsf{Mark},\mathsf{Conv}}$$

where $\lambda = (X \otimes A, g, g')$

$$g : X \otimes A \to X \otimes A \otimes A \qquad \text{in Mark}$$

$$g' : \Delta(X \otimes A) \to [\mathbb{R}, \mathbb{R}] \qquad \text{in Conv}$$
$$g'(\alpha)(r) = \mathbb{E}U(\alpha) + \beta r$$



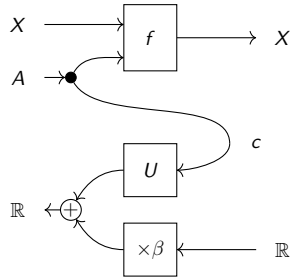
$\alpha : \Delta(X \otimes A)$, joint distribution on states and actions.

Optics for Value Iteration

Mixed optics for continuous-space MDPs

Savings problem: $V'(x) = \max_{0 \le c \le x+i} U(c) + \beta V((1+\gamma)(x-c+i))$

- V improv:
  $V'(x) = U(\pi(x)) + \beta V(f(x, \pi(x)))$
- π improv:
  $\pi'(x) = \operatorname{argmax}_{0 \le c \le x+i} U(c) + \beta V(f(x, c))$



Recall:

- Policy: $\pi : x \mapsto c$
- Transition: $f : (x, c) \mapsto (1+\gamma)(x-c+i)$
- Utility: $U : (\_, c) \mapsto U(c)$

Stoch: Markov category of measurable spaces and stochastic maps
(Gauss: Euclidean spaces and affine functions with noise $Y = MX + \xi$)

From Value Iteration to Q-learning

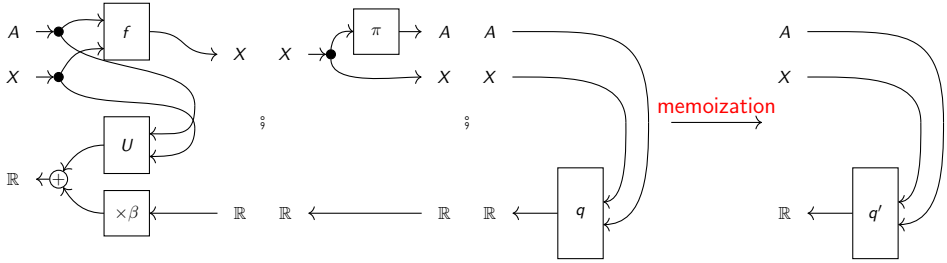
Shift from learning state values $V : X \to \mathbb{R}$ to state-action values $q : X \times A \to \mathbb{R}$.

- Value Iteration, in $A^X \times \text{Optic}\left(\binom{X}{\mathbb{R}}, \binom{I}{I}\right)$:

$$(\pi, V) \mapsto (x \mapsto \arg\max_a(\lambda \fatsemi V)(x, a), \bar{\pi} \fatsemi \lambda \fatsemi V)$$

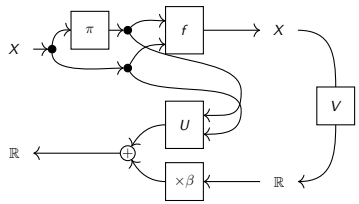
- State-Action Value Iteration, in $A^X \times \text{Optic}\left(\binom{X \times A}{\mathbb{R}}, \binom{I}{I}\right)$:

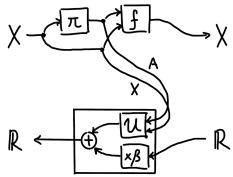
$$(\pi, q) \mapsto (x \mapsto \arg\max_a q(x, a), \lambda \fatsemi \bar{\pi} \fatsemi q)$$



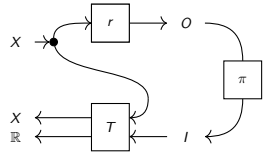
Related work

MDPs in Myers' open dyn. systems[2]: Functor lenses of $\mathrm{BiKl}(C \times -, \Delta(\mathbb{R} \times -))^{\mathrm{op}}$
Different interfaces:





(a) Interface is a value function (ours)



(b) Interface is a policy (Myers')

$$\begin{pmatrix} X \\ \mathbb{R} \end{pmatrix} \to \begin{pmatrix} X \\ \mathbb{R} \end{pmatrix}$$

$$\begin{pmatrix} X \\ \Delta(X \times \mathbb{R}) \end{pmatrix} \to \begin{pmatrix} O \\ I \end{pmatrix}$$

[2] Jaz Myers, "Double Categories of Open Dynamical Systems (Extended Abstract)".

Summary and future work

- Optics can be tailored to study different MDPs (deterministic, stochastic, discrete-space, continuous-space).
- Q-learning feels ad hoc: use categorical cybernetics to express it with optics

	One state ($X = 1$)	Many states	Imperfect states
Without agency (fixed π)	-	Markov chain	Hidden Markov Model (HMM)
With agency (learned π)	"Bandit problem" w/ delayed rewards	MDP	Partially Observable MDP (POMDP)

- Known/unknown environment
- Problem structure: Adaptive Control, Dynamic Programming, Reinforcement Learning.

Value iteration is optic composition

Jules Hedges, Riu Rodríguez Sakamoto

Mathematically Structured Programming group,
Department of Computer and Information Sciences,
University of Strathclyde

riu.rodriguez-sakamoto@strath.ac.uk

5th International Conference on Applied Category Theory, Glasgow (Scotland)

July 21st, 2022

Dynamic Programming theorems

- Policy improvement theorem: Given deterministic policies $\pi, \pi' : X \to A$,

$$\begin{aligned} & q_\pi(x, \pi'(x)) \geq V_\pi(x) \ \forall x \in X \\ \text{implies} \quad & V_{\pi'}(x) \geq V_\pi(x) \ \forall x \in X \end{aligned}$$

  Also for stochastic policies $\pi, \pi' : X \to \Delta A$, by defining

$$q_\pi(s, \pi'(s)) = \sum_a \pi'(a \mid s) q_\pi(s, a)$$

- Convergence proofs require metric. Optics enriched in metric spaces?

MDP example: Inverted pendulum

- System of non-linear differential equations[3]:

$$(M+m)\ddot{y} + mL\ddot{\theta}\cos\theta - mL\dot{\theta}^2\sin\theta = a$$
$$mL\ddot{y}\cos\theta + mL^2\ddot{\theta} - mLg\sin\theta = 0$$

- RL approach: Linearization near equilibrium: $\dot{x} = Ax(t) + Ba(t)$ and time-discretization via Euler approximation $x_{k+1} = x_k + \Delta t(Ax_k + Ba_k)$

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & -\frac{mg}{M} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & \frac{(M+m)g}{ML} & 0 \end{pmatrix} \qquad B = \begin{pmatrix} 0 \\ \frac{1}{M} \\ 0 \\ -\frac{1}{ML} \end{pmatrix} \qquad x = \begin{pmatrix} y \\ \dot{y} \\ \theta \\ \dot{\theta} \end{pmatrix}$$

- Optimal control approach:

$$J(x,a) = \int_0^\infty C(x(t), a(t))\mathrm{d}t$$

[3] Friedland, *Control Systems Design*, Example 2E.

Conv

- $(\mathrm{Conv}, \boxtimes, \{*\})$ monoidal product[4] of two convex spaces $A \boxtimes B$:
  - Free convex structure on $A \times B$, then take the smallest congruence relation on the set s.t.

$$\sum_i \alpha_i (a_i, b) \equiv (\sum_i \alpha_i a_i, b) \quad \text{and} \quad \sum_i \alpha_i (a, b_i) \equiv (a, \sum_i \alpha_i b_i)$$

  - Classifies bi-affine functions $f : A \times B \to C$, like how $A \otimes B$ classifies bilinear functions $f : A \times B \to C$ in R-Mod.
  - Monoidal closure: Convex structure on $B^A = \mathrm{Conv}(A, B)$ is pointwise: Given $f, g \in \mathrm{Conv}(A, B)$, $(f +_\alpha g)(a) = f(a) +_\alpha g(a)$.

[4] Stirtz, "Categorical probability theory", Sec 2.2.

- Friedland, Bernard. *Control Systems Design. An Introduction To State-Space Methods*. McGraw-Hill Companies, 1985, p. 513. ISBN: 9780070224414.
- Jaz Myers, David. "Double Categories of Open Dynamical Systems (Extended Abstract)". en. In: *Electronic Proceedings in Theoretical Computer Science* 333 (Feb. 2021), pp. 154–167. ISSN: 2075-2180. DOI: 10.4204/EPTCS.333.11. URL: http://arxiv.org/abs/2005.05956v2 (visited on 12/12/2021).
- Stirtz, Kirk. "Categorical probability theory". arXiv:1406.6030. 2015.