Cultivating Desired Behaviour

Policy Teaching Via Environment-Dynamics Tweaks

by Zinovi Rabinovich, Lachlan Dufton, Kate Larson and Nicholas R. Jennings

IAM Group, ECS, University of Southampton
CS, University of Waterloo
Agenda

- Motivation and Intuition
- Behaviour Cultivating teacher-learner interaction
  - ... in Markovian environments
  - ... teaching costs
- visa vi Policy Iteration
- Experiments
- Conclusions
Riding a Bike - Task
Riding a Bike
Riding a Bike
Riding a Bike - Demo
Riding a Bike - Demo’s no good
Riding a Bike - Reward
Riding a Bike - Reward
Riding a Bike - Reward
Riding a Bike - Reward's no good
Riding a Bike - Dynamics Tweak
Riding a Bike - Dynamics Tweak
Behaviour Cultivation

Gradual (minimal) change of environment dynamics to engender the desired behaviour of an (adaptive) agent.
Formalities – Can do’s

Between two agents: a teacher and learner

Learner can:
- Sense its environment
- Act within it
- Obtain reward

Teacher can:
- Sense the learner’s behaviour
- Change the environment response
Formalities – Wants

- **Learner** wants
  - Behave to maximise the *reward*

- **Teacher** wants
  - Force a specific behaviour of the learner
Formalities – Wants

- **Learner** wants
  - Behave to maximise the *reward*

- **Teacher** wants
  - Force a specific behaviour of the learner

- **Teacher does not want**
  - To change the environment too much
Formalities – Wants

- **Learner** wants
  - Behave to maximise the *reward*

- **Teacher** want to minimise *costs* of
  - learner’s behaviour deviations
  - environment change
Formalities – Model

Environment $\langle S, A, c, \gamma, U, T \rangle$:
- $S$ – set of states
- $A$ – set of learner’s actions
- $c : S \times A \times S \rightarrow \mathbb{R}$ is the learner’s reward
- $\gamma \in (0, 1)$ is a learner’s discount factor
- $U$ is the set of teacher’s actions
- $T : S \times A \times U \rightarrow \Delta(S)$ are the environment dynamics
Formalities – Model

Environment $\langle S, A, c, \gamma, U, T \rangle$:

- $S$ – set of states
- $A$ – set of learner’s actions
- $c : S \times A \times S \to \mathbb{R}$ is the learner’s reward
- $\gamma \in (0, 1)$ is a learner’s discount factor
- $U$ is the set of teacher’s actions
- Nominal dynamics $T^0 = T_{u_{\text{zero}}}$ for some $u_{\text{zero}} \in U$
- $T : S \times A \times U \to \Delta(S)$ are the environment dynamics
Formalities – Model

Environment $\langle S, A, c, \gamma, U, T \rangle$:

- $S$ – set of states
- $A$ – set of learner’s actions
- $c : S \times A \times S \rightarrow \mathbb{R}$ is the learner’s reward
- $\gamma \in (0, 1)$ is a learner’s discount factor
- $U$ is the set of teacher’s actions
- $T : S \times A \times U \rightarrow \Delta(S)$ are the environment dynamics

At every time step $t$ the learner faces $\langle S, A, T_{u_t}, c \rangle$
Formalities – Model

- Environment $\langle S, A, c, \gamma, U, T \rangle$:
  - $S$ – set of states
  - $A$ – set of learner’s actions
  - $c : S \times A \times S \rightarrow \mathbb{R}$ is the learner’s reward
  - $\gamma \in (0, 1)$ is a learner’s discount factor
  - $U$ is the set of teacher’s actions
  - $T : S \times A \times U \rightarrow \Delta(S)$ are the environment dynamics

- At every time step $t$ the learner faces $< S, A, T_{ut}, c >$

- Learner seeks $\pi : S \rightarrow \Delta(A)$, and assumes the dynamics $T_{ut}$ hold indefinitely
Formalities – Model

Environment $\langle S, A, c, \gamma, U, T \rangle$:

- $S$ – set of states
- $A$ – set of learner’s actions
- $c : S \times A \times S \rightarrow \mathbb{R}$ is the learner’s reward
- $\gamma \in (0, 1)$ is a learner’s discount factor
- $U$ is the set of teacher’s actions
- $T : S \times A \times U \rightarrow \Delta(S)$ are the environment dynamics

At every time step $t$ the learner faces $\langle S, A, T_{ut}, c \rangle$

Learner seeks $\pi : S \rightarrow \Delta(A)$, and assumes the dynamics $T_{ut}$ hold indefinitely

Teacher wants $\pi^* : S \rightarrow \Delta(A)$
Teaching problem

Find a sequence $\{u_t\}_{t=1}^{t_{max}}$ to minimise teaching costs
Teaching problem

Find a sequence $\{u_t\}_{t=1}^{t_{\text{max}}}$ to minimise teaching costs

$$\min_{u_t} \sum_{t=1}^{t_{\text{max}}} \text{Cost}(\pi_t, u_t)$$

s.t.

$$\pi_t = \pi(x_t)$$

$$x_t = F(x_{t-1}, u_t).$$
Teaching problem

- Find a sequence \( \{u_t\}_{t=1}^{t_{max}} \) to minimise teaching costs

\[
\min_{u_t} \sum_{t=1}^{t_{max}} \text{Cost}(\pi_t, u_t)
\]

s.t.
\[
\pi_t = \pi(x_t)
\]
\[
x_t = F(x_{t-1}, u_t).
\]

- \( x_t \) is all the learner knows at time \( t \)
Teaching problem

- Find a sequence \( \{u_t\}_{t=1}^{t_{\text{max}}} \) to minimise teaching costs

\[
\min_{u_t} \sum_{t=1}^{t_{\text{max}}} \text{Cost}(\pi_t, u_t)
\]

s.t.

\[
\pi_t = \pi(x_t)
\]

\[
x_t = F(x_{t-1}, u_t).
\]

- \( x_t \) is all the learner knows at time \( t \)
- \( F \) is the learner’s algorithm
Teaching problem

- Find a sequence \( \{u_t\}_{t=1}^{t_{\text{max}}} \) to minimise teaching costs

\[
\min_{u_t} \sum_{t=1}^{t_{\text{max}}} \text{Cost}(\pi_t, u_t)
\]

s.t.

\[
\pi_t = \pi(x_t)
\]

\[
x_t = F(x_{t-1}, u_t).
\]

- \( x_t \) is all the learner knows at time \( t \)
- \( F \) is the learner’s algorithm
- How do we measure the \( \text{Cost} \)?
Teaching Costs

- Composition of
  - Teaching success
  - Teaching effort
Teaching Costs

- Composition of
  - Difference between $\pi_t$ and $\pi^*$
  - Teaching effort
Teaching Costs

- Composition of
  - Difference between $\pi_t$ and $\pi^*$
  - Difference between $T_{ut}$ and $T^0$
Teaching Costs

- Composition of
  - Difference between $\pi_t$ and $\pi^*$
  - Difference between $T_{ut}$ and $T^0$

- Fused environment dynamics:
  - Applied $P_t(s', a'|s, a) = T_{ut}(s'|s, a)\pi_t(a'|s')$
  - Ideal $P^*(s', a'|s, a) = T^0(s'|s, a)\pi^*(a'|s')$
Teaching Costs

- Composition of
  - Difference between $\pi_t$ and $\pi^*$
  - Difference between $T_{ut}$ and $T^0$

- Fused environment dynamics:
  - Applied $P_t(s', a'|s, a) = T_{ut}(s'|s, a)\pi_t(a'|s')$
  - Ideal $P^*(s', a'|s, a) = T^0(s'|s, a)\pi^*(a'|s')$

- Cost is then the difference between $P_t$ and $P^*$
Teaching Costs

- Composition of
  - Difference between $\pi_t$ and $\pi^*$
  - Difference between $T_{ut}$ and $T^0$

- Fused environment dynamics:
  - Applied $P_t(s', a'|s, a) = T_{ut}(s'|s, a)\pi_t(a'|s')$
  - Ideal $P^*(s', a'|s, a) = T^0(s'|s, a)\pi^*(a'|s')$

- Cost is then the difference between $P_t$ and $P^*$
  - What difference?
Teaching Costs – KL Rate

Both $P_t$ and $P^*$ are Markovian processes over $S \times A$. 
Teaching Costs – KL Rate

- Both $P_t$ and $P^*$ are Markovian processes over $S \times A$
- Kullback-Leibler Rate for Markovian processes
Teaching Costs – KL Rate

- Both $P_t$ and $P^*$ are Markovian processes over $S \times A$
- Kullback-Leibler Rate for Markovian processes
  - Kullback-Leibler Divergence for $p(\cdot)$ and $q(\cdot)$:

$$D_{KL}(p||q) = \sum_i p(i) \log \frac{p(i)}{q(i)}$$
Both $P_t$ and $P^*$ are Markovian processes over $S \times A$

Kullback-Leibler Rate for Markovian processes

$$D^{KL}(p||q) = \sum_i p(i) \log \frac{p(i)}{q(i)}$$
Both $P_t$ and $P^*$ are Markovian processes over $S \times A$

Kullback-Leibler Rate for Markovian processes

$$D_{KL}(p||q) = \sum_i p(i) \log \frac{p(i)}{q(i)}$$

KL Rate for processes $P(\cdot|\cdot)$ and $Q(\cdot|\cdot)$:

$$KLR(P||Q) = \sum_x D_{KL}(P(\cdot|x)||Q(\cdot|x))p_{stat}(x)$$
Both $P_t$ and $P^*$ are Markovian processes over $S \times A$.

**Kullback-Leibler Rate** for Markovian processes

\[
D^{KL}(p||q) = \sum_i p(i) \log \frac{p(i)}{q(i)}
\]

\[
KLR(P||Q) = \sum_x D^{KL}(P(\cdot|x)||Q(\cdot|x))p_{stat}(x)
\]
Both $P_t$ and $P^*$ are Markovian processes over $S \times A$

**Kullback-Leibler Rate** for Markovian processes

\[
D^{KL}(p||q) = \sum_i p(i) \log \frac{p(i)}{q(i)}
\]

\[
KLR(P||Q) = \sum_x D^{KL}(P(\cdot|x)||Q(\cdot|x))p_{stat}(x)
\]

\[
Cost(u_t, \pi_t) = KLR(P_t||P^*)
\]
Who’s the learner

Find a sequence \( \{u_t\}_{t=1}^{t_{max}} \) to minimise teaching costs

\[
\min_{u_t} \sum_{t=1}^{t_{max}} \text{Cost}(\pi_t, u_t)
\]

s.t.

\[
\pi_t = \pi(x_t)
\]

\[
x_t = F(x_{t-1}, u_t)
\]
Policy Iteration

- Given $< S, A, T_{ut}, c >$
- Iterate until convergence
  - Evaluate current policy $\pi_t$
  - Improve the policy $\pi_{t+1}$
Policy Iteration

Given $< S, A, T_{Ut}, c >$

Iterate until convergence

Evaluate current policy $\pi_t$

$V_t = V^{\pi_t}(s)$ reward accumulated starting in $s$ following $\pi_t$

Improve the policy $\pi_{t+1}$
Policy Iteration

- Given $< S, A, T_{ut}, c >$
- Iterate until convergence
  - Evaluate current policy $\pi_t$
    - $V_t = V^{\pi_t}(s)$ reward accumulated starting in $s$ following $\pi_t$
  - Improve the policy $\pi_{t+1}$
    - The best first step, followed by $\pi_t$
Policy Iteration

- Given $< S, A, T_u, c >$
- Iterate until convergence
  - Evaluate current policy $\pi_t$
    - $V_t = V^{\pi_t}(s)$ reward accumulated starting in $s$ following $\pi_t$
  - Improve the policy $\pi_{t+1}$
    - The best first step, followed by $\pi_t$
  - $x_t$ is composed of $\pi_t$ and $V_t$
Experiments – Bike Ride
Experiments – Bike Ride
Experiments – Bike Ride
Experiments – Transportation

- Move to the target cell
- Punished for delays
- Motion success follows topography
Experiments – Teacher vs Learner

<table>
<thead>
<tr>
<th>Learner’s best</th>
<th>Teacher’s preferred</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Diagram" /></td>
<td><img src="image" alt="Diagram" /></td>
</tr>
</tbody>
</table>
Experiments – Terrain Shift

Neutral

A modification
Experiments – Necessity
Experiments – Necessity

Demonstration does not work
Experiments – Necessity

Demonstration does not work

No pressure on the learner
Demonstration does not work
- No pressure on the learner
- Rewarding does not work
Experiments – Necessity

- Demonstration does not work
  - No pressure on the learner
- Rewarding does not work
  - No effective reward pattern exists
Experiments – Necessity

- Demonstration **does not work**
  - No pressure on the learner
- Rewarding **does not work**
  - No effective reward pattern exists
- Behaviour Cultivation **works!**
Conclusions

New teacher-learner interaction: Behaviour Cultivation
Conclusions

- New teacher-learner interaction: Behaviour Cultivation
- Behaviour Cultivation is effective
Conclusions

New teacher-learner interaction: *Behaviour Cultivation*
- Behaviour Cultivation is *effective*
- Behaviour Cultivation is *necessary*
Future and Current Work

- Alternative teacher’s cost functions
- Partial observability
  - of the environment by the learner
  - of the learner by the teacher
- Alternative learners
  - Scaling up to multi-agent systems
  - Partially or arbitrary learner
Thank you
Thank you