The Effects on Adaptive Behaviour of Negatively Valenced Signals in Reinforcement Learning

Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics (ICDL-EpiRob)

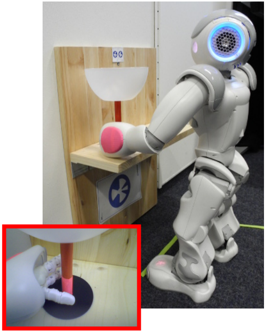
Nicolás Navarro Guerrero
Robert Lowe & Stefan Wermter
Department of Informatics, University of Hamburg

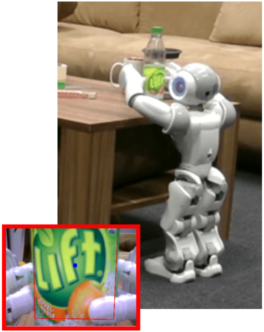
https://nicolas-navarro-guerrero.github.io/

20th September 2017 – Lisbon, Portugal

Motivation

- Learn from scratch
- Avoid both collisions and self-collisions





TD-learning

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$$

$$r_t = Reward + Punishment$$

- TD-learning algorithms cannot tell apart
  - *high-gain/high-risk* from *low-gain/no-risk* options

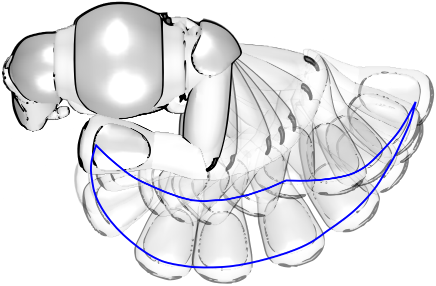
(Palminteri and Pessiglione, 2017; Seymour et al., 2015, 2005)

- An embodied solution is to use nociception, motivated by the Somatic Marker Hypothesis (Damasio, 1996)

Nociception: perception of (potentially) harmful stimuli

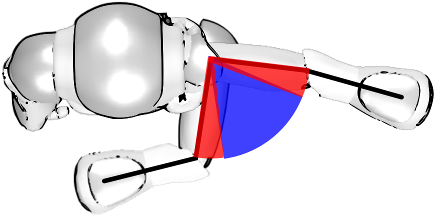
Task description

- Focus on punishment and nociception on robot learning



2D workspace of NAO's left arm

Nociception and punishment

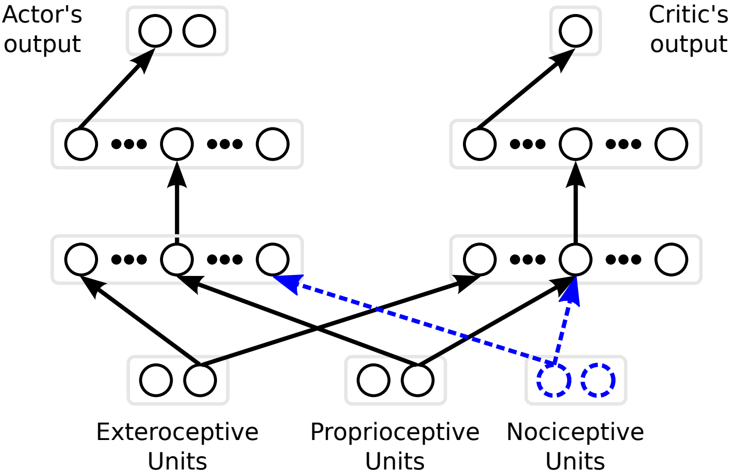


Range of movement of shoulder joint.

- Punishment/nociception in upper/lower 10% of the range
- Different activation of punishment/nociception
- Nociception differentiates between upper/lower *pain*
- Nociception differentiates between elbow and shoulder

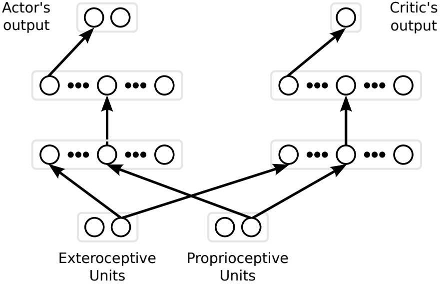
Neural architecture

$\delta_t \propto r_t = R + P$

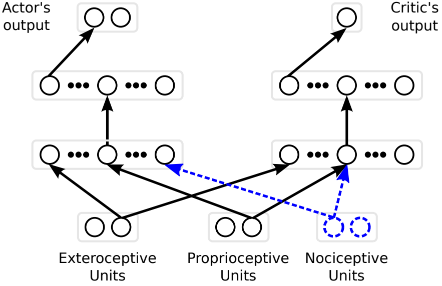


Four conditions

No punishment	$\delta_t \propto r_t = R$	$\delta_t \propto r_t = R$
Punishment	$\delta_t \propto r_t = R + P$	$\delta_t \propto r_t = R + P$



No Nociception



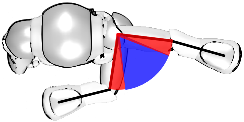
Nociception

Functions for Punishment and Nociception

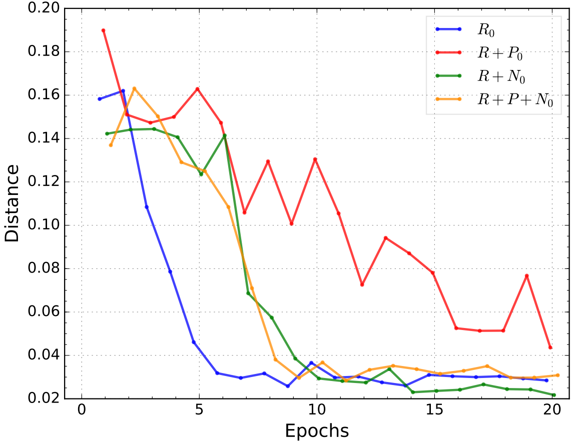
Binary

$$r_t^- = \begin{cases} -P & : \xi_i = \xi_i^{min} \vee \xi_i = \xi_i^{max} \\ 0 & \text{otherwise.} \end{cases}$$

$$n_t = \begin{cases} -1 & : \xi_i = \xi_i^{min} \\ 1 & : \xi_i = \xi_i^{max} \\ 0 & \text{otherwise.} \end{cases}$$



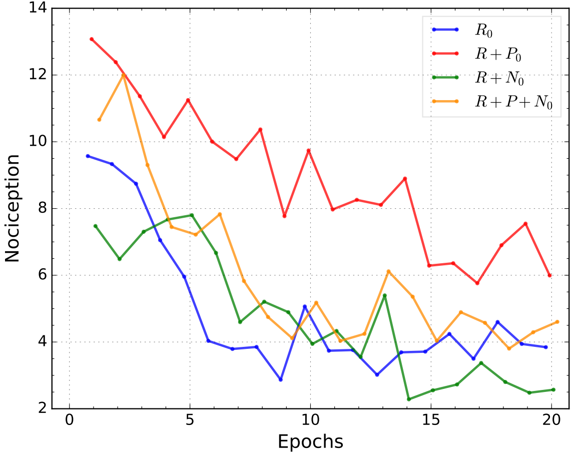
Results: Positioning error



Change of mean distance during learning
(Abrupt Exponential)

Navarro-Guerrero et al., 2017, *Frontiers in Neurorobotics*

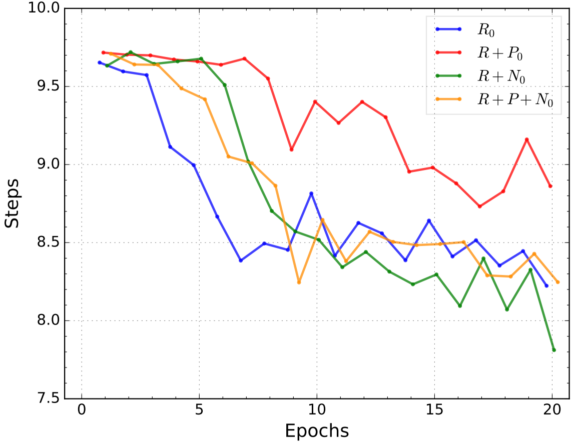
Results: Potential for damage



Change of mean nociception during learning
(Abrupt Exponential)

Navarro-Guerrero et al., 2017, *Frontiers in Neurorobotics*

Results: Length of action sequences



Change of mean action sequence length during learning
(Abrupt Exponential)

Navarro-Guerrero et al., 2017, *Frontiers in Neurorobotics*

Results: Significance mean positioning error

			Binary	Step	Linear	$e \propto \sigma$	$e \propto 3\sigma$
R+P	R+N	After Learning	0.1401 −30.49 %	0.3143 −17.00 %	0.9860 1.20 %	0.0742 25.97 %	0.4958 −33.61 %
R+P	R+P+N	After Learning	0.9949 1.37 %	0.0004 −45.75 %	0.8151 5.09 %	0.9829 2.21 %	0.0011 −108.30 %
R+N	R+P+N	After Learning	0.1150 24.42 %	0.0372 −24.57 %	0.8945 3.94 %	0.1095 −32.09 %	0.0345 −55.90 %
R+P	R+N	Cumulative	0.7141 −4.04 %	0.0004 −17.29 %	0.0921 8.72 %	0.0000 31.72 %	0.0000 44.09 %
R+P	R+P+N	Cumulative	0.0065 −16.09 %	0.1856 −7.65 %	0.0002 17.36 %	0.0016 17.46 %	0.5804 4.27 %
R+N	R+P+N	Cumulative	0.0550 −11.58 %	0.0716 8.22 %	0.0961 9.46 %	0.0124 −20.89 %	0.0000 −71.23 %

Results: Significance mean potential for damage

			Binary	Step	Linear	$e \propto \sigma$	$e \propto 3\sigma$
R+P	R+N	After Learning	0.3259 −28.16 %	0.0900 7.02 %	0.9745 −3.19 %	0.0113 36.42 %	0.0009 −111.74 %
R+P	R+P+N	After Learning	0.5912 −19.20 %	0.0809 7.18 %	0.6319 13.48 %	0.9461 3.93 %	0.0032 −100.83 %
R+N	R+P+N	After Learning	0.8914 6.99 %	0.9987 0.17 %	0.4968 16.16 %	0.0271 −51.08 %	0.9301 5.15 %
R+P	R+N	Cumulative	0.3618 −6.63 %	0.0000 13.44 %	0.0001 13.58 %	0.0000 35.34 %	0.0000 35.17 %
R+P	R+P+N	Cumulative	0.0000 −27.42 %	0.0000 9.82 %	0.0000 22.36 %	0.0000 15.37 %	0.0665 8.28 %
R+N	R+P+N	Cumulative	0.0001 −19.49 %	0.0031 −4.18 %	0.0201 10.16 %	0.0000 −30.88 %	0.0000 −41.49 %

Results: Significance mean positioning speed

			Binary	Step	Linear	$e \propto \sigma$	$e \propto 3\sigma$
R+P	R+N	After Learning	0.8685 −0.79 %	0.5035 −1.63 %	0.9690 0.32 %	0.0001 6.31 %	0.1335 −2.94 %
R+P	R+P+N	After Learning	0.5025 −1.75 %	0.0000 −6.75 %	0.9988 0.06 %	0.2311 2.34 %	0.0002 −6.27 %
R+N	R+P+N	After Learning	0.8117 −0.95 %	0.0018 −5.04 %	0.9798 −0.26 %	0.0163 −4.25 %	0.0773 −3.23 %
R+P	R+N	Cumulative	0.8461 0.24 %	0.0008 −1.51 %	0.0016 1.46 %	0.0000 4.74 %	0.0000 5.14 %
R+P	R+P+N	Cumulative	0.0046 −1.38 %	0.0000 −2.60 %	0.0000 2.33 %	0.0000 2.35 %	0.1490 0.89 %
R+N	R+P+N	Cumulative	0.0007 −1.62 %	0.0206 −1.07 %	0.0867 0.89 %	0.0000 −2.50 %	0.0000 −4.49 %

Contributions: Damage minimization

Nociception can improve:

- behavioural performance,
- reduce potential to damage, and
- reduce action sequences.

Future work:

- Underlying mechanism leading to improvements
- Test effect of nociception on human-like poses
- Alternative way to use negatively valenced signals

Reference List

Damasio, A. R. (1996). « The Somatic Marker Hypothesis and the Possible Functions of the Prefrontal Cortex [and Discussion] ». *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 351(1346), pp. 1413–1420 (cit. on pp. 3–5).

Navarro-Guerrero, N., Lowe, R., and Wermter, S. (2017). « Improving Robot Motor Learning with Negatively Valenced Reinforcement Signals ». *Frontiers in Neurorobotics* 11(10) (cit. on pp. 19–21).

Palminteri, S. and Pessiglione, M. (2017). « Opponent Brain Systems for Reward and Punishment Learning: Causal Evidence from Drug and Lesion Studies in Humans ». In: *Decision Neuroscience: An Integrative Approach.* Ed. by J.-C. Dreher and L. Tremblay. Chapter: 23. San Diego: Academic Press, pp. 291–303 (cit. on pp. 3–5).

Seymour, B., Maruyama, M., and De Martino, B. (2015). « When Is a Loss a Loss? Excitatory and Inhibitory Processes in Loss-Related Decision-Making ». *Current Opinion in Behavioral Sciences.* Neuroeconomics 5, pp. 122–127 (cit. on pp. 3–5).

Seymour, B., O'Doherty, J. P., Koltzenburg, M., Wiech, K., Frackowiak, R., Friston, K., and Dolan, R. (2005). « Opponent Appetitive-Aversive Neural Processes Underlie Predictive Learning of Pain Relief ». *Nature Neuroscience* 8(9), pp. 1234–1240 (cit. on pp. 3–5).