ulm university universität ulm

An Information-theoretic On-line Learning Principle for Specialization in Hierarchical Decision-Making Systems

Heinke Hihn, Sebastian Gottwald, and Daniel A. Braun

Ulm University
Institute for Neural Information Processing

December 12, 2019
58^{th} IEEE Conference on Decision and Control, Nice

Emergence of Specialized Decision-Makers

Decision-Maker: optimizes a utility U in state s:

$$a_s^* = \arg\max_a U(s, a)$$

Central Idea:
Limited resources such as

- Linear Decision-Makers
- Limited information processing

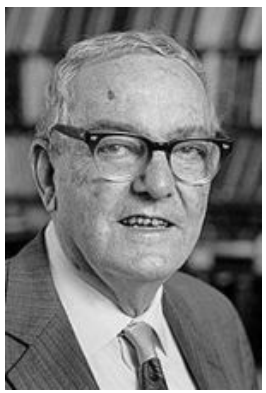
drive specialization. [1] [2]

Motivation: Linear decision-makers are easy to analyze.

[1] Genewein, T., Leibfried, F., Grau-Moya, J., and Braun, D.A. *Bounded rationality, abstraction, and hierarchical decision-making: An information-theoretic optimality principle.* Frontiers in Robotics and AI, 2:27, 2015.

[2] Hihn, H., Gottwald, S., and Braun, D. A. (2018). Bounded rational decision-making with adaptive neural network priors. *In IAPR Workshop on Artificial Neural Networks in Pattern Recognition*.

Bounded Rationality and Specialization



Herbert A. Simon coined the term *Bounded Rationality*

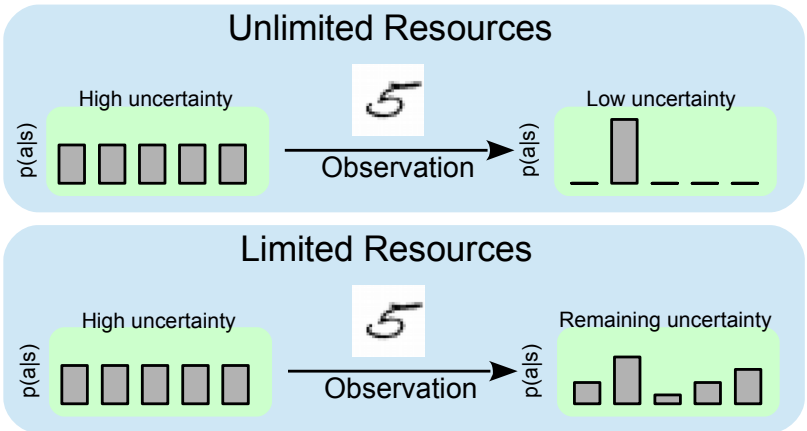
Intelligent agents must invest their resources such that they optimally trade off utility versus processing costs [3] [4]

Consequence: Specialization

[3] Simon, H. A. *A behavioral model of rational choice.* The Quarterly Journal of Economics, 69(1):99–118, 1955.

[4] Gershman, S. J., et al. *Computational rationality: A converging paradigm for intelligence in brains, minds, and machines.* Science (2015)

Information-theoretic Bounded Rationality [5]



$$\max_{p(a|s)} \mathbb{E}_{p(s),p(a|s)}\left[U(s,a)\right] \text{ s.t. } I(S;A) \leq C \tag{1}$$

$$p^*(a|s) = \arg\max_{p(a|s)} \mathbb{E}\left[U(s,a)\right] - \frac{1}{\beta} I(S;A) \tag{2}$$

Mutual Information: $I(S;A) = \mathbb{E}_{p(a|s)}\left[\mathrm{D}_{\mathrm{KL}}(p(a|s)||p(a))\right]$

[5] Ortega, P. A., and Braun, D.A.. *Thermodynamics as a theory of decision-making with information-processing costs.* Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 469(2153), 2013.

Hierarchical Decision-Making

Extend to two-level hierarchy with experts $x \in X$ [6]

$$S \to X \to A \tag{3}$$

Extended objective:

$$\max_{p(a|s,x),p(x|s)} \mathbb{E}[U(s,a)] - \frac{1}{\beta_1} I(S;X) - \frac{1}{\beta_2} I(S;A|X). \tag{4}$$

[6] Genewein, T., Leibfried, F., Grau-Moya, J., and Braun, D.A. *Bounded rationality, abstraction, and hierarchical decision-making: An information- theoretic optimality principle.* Frontiers in Robotics and AI, 2:27, 2015.

Learning via Gradient Descent

Parametrize distributions with parameters θ and ϑ:

$$J(s,x,a) = U(s,a) - \frac{1}{\beta_1}\log\frac{p_\theta(x|s)}{p(x)} - \frac{1}{\beta_2}\log\frac{p_\vartheta(a|s,x)}{p(a|x)} \tag{5}$$

$$\max_\theta \quad \mathbb{E}_{p_\theta(x|s)}\left[\hat{f}(x,s) - \tfrac{1}{\beta_1}\log\tfrac{p_\theta(x|s)}{p(x)}\right] \tag{6}$$

$$\hat{f}(x,s) = \underbrace{\mathbb{E}_{p_\vartheta(a|x,s)}\left[U(s,a) - \frac{1}{\beta_2}\log\frac{p_\vartheta(a|s,x)}{p(a|x)}\right]}_{\text{Expert Objective}} \tag{7}$$

Approximate the prior distributions $p(x)$ and $p(a|x)$ by running means.

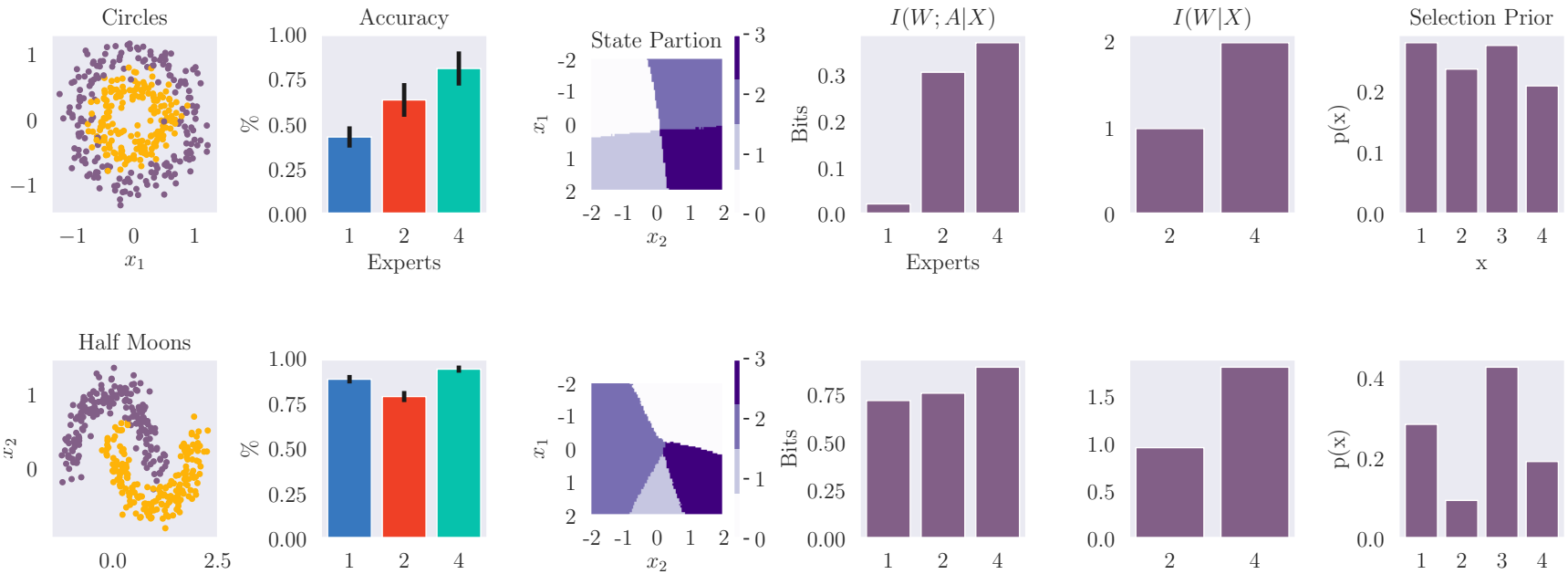
Utilities for Classification and Regression

1. cross-entropy loss
   $\mathcal{L}(y, \hat{y}) = \sum_i y_i \log \frac{1}{\hat{y}_i} = -\sum_i y_i \log \hat{y}_i$
2. mean squared error $\mathcal{L}(y, \hat{y}) = \sum_i (\hat{y}_i - y_i)^2$

$$\max_{\theta} \quad \mathbb{E}_{p_\theta(x|s)} \left[\hat{f}(x, s) - \frac{1}{\beta_1} \log \frac{p_\theta(x|s)}{p(x)} \right] \tag{8}$$

$$\hat{f}(x, s) = \underbrace{\mathbb{E}_{p_\vartheta(\hat{y}|x,s)} \left[-\mathcal{L}(\hat{y}, y) - \frac{1}{\beta_2} \log \frac{p_\vartheta(\hat{y}|s, x)}{p(\hat{y}|x)} \right]}_{\text{Expert Objective}} \tag{9}$$

Classification



Reinforcement Learning: Setup

Markov Decision Process as a tuple $(\mathcal{S}, \mathcal{A}, P, r)$, where

- $\mathcal{S}$ is the set of states
- $\mathcal{A}$ the set of actions
- $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ is the transition probability
- $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is a reward function

Find policy π_θ maximizing expected reward:

$$\theta^* = \arg\max_\theta \underbrace{\mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^{\infty} r(s_t, a_t) \right]}_{J(\pi_\theta)}. \tag{10}$$

RL Objective

Penalize deviation from a prior policy:

$$\arg\max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t \left(r(s_t, a_t) - \frac{1}{\beta} \log \frac{\pi(a_t|s_t)}{\pi(a)} \right) \right]. \quad (11)$$

Similar to MaxEnt RL [7], Trust Region Policy Optimization [8], Mutual Information Regularized RL [9]

[7] Eysenbach, B. and Levine, S. *If MaxEnt RL is the Answer, What is the Question?*. arXiv preprint (2019).

[8] Schulman, J., et al. *Trust region policy optimization*. In International Conference on Machine Learning (2015)

[9] Leibfried, F., and Grau-Moya, J. *Mutual-information regularization in markov decision processes and actor-critic learning.* Conference on Robot Learning (2019).

RL Objectives

Advantage-Actor-Critic [10] Selection Stage Objective:

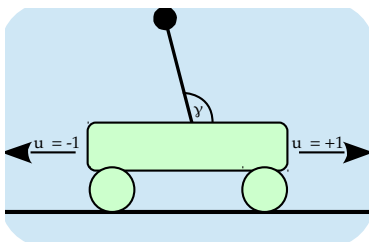
$$\max_{\theta} \mathbb{E}_{\pi_\theta(x|s)} \left[\hat{f}(s,x) - \frac{1}{\beta_1} \log \frac{\pi_\theta(x|s)}{\pi(x)} \right], \tag{12}$$

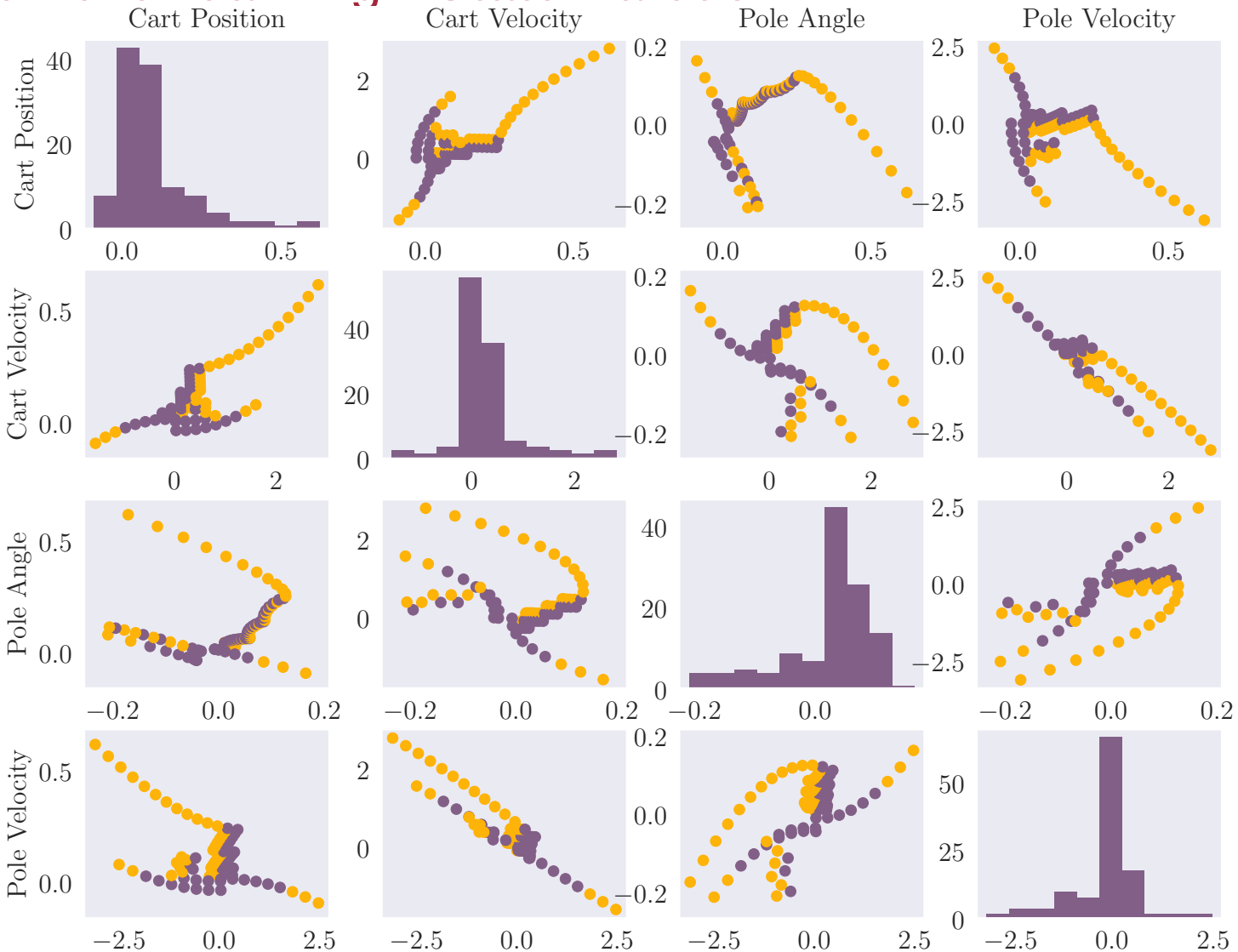
where

$$\hat{f}(s,x) = \underbrace{\mathbb{E}_{\pi_\vartheta(a|s,x)} \left[r(s,a) - \frac{1}{\beta_2} \log \frac{\pi_\vartheta(a|s,x)}{\pi(a|x)} \right]}_{\text{Expert Objective}} \tag{13}$$

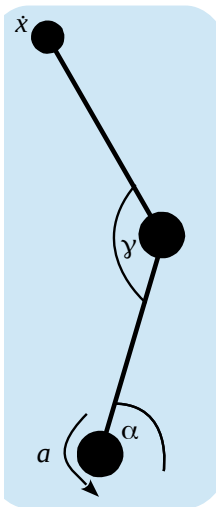
[10] Schulman, J., et al. *High-dimensional continuous control using generalized advantage estimation.* International Conference on Learning Representations (2015)

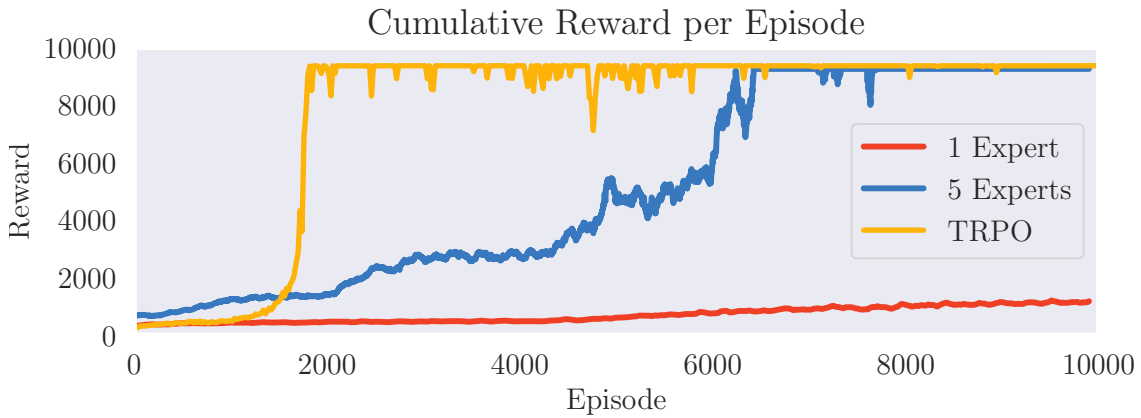
Reinforcement Learning - State Partition

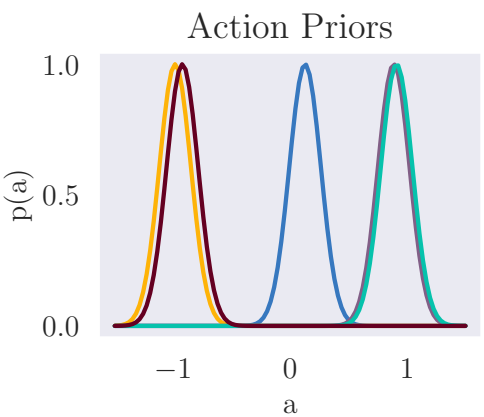


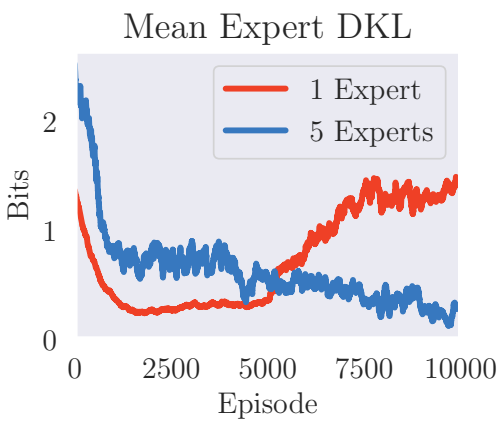


Reinforcement Learning - Continuous Control Problems









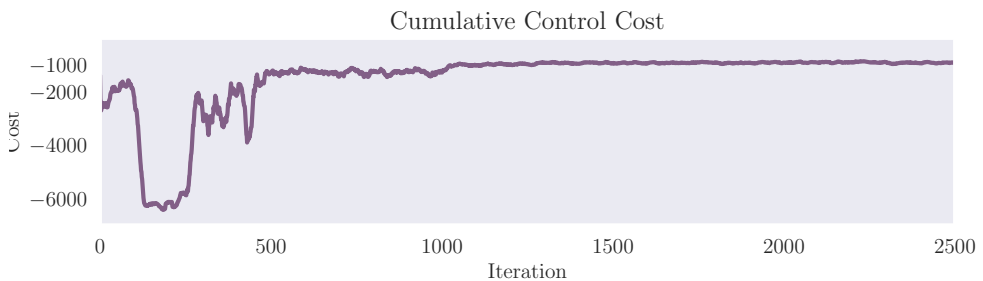
see [11]

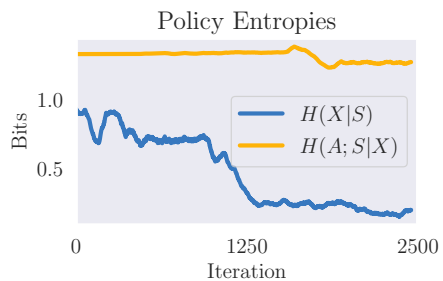
[11] Schulman, J., et al. *Trust region policy optimization*. In International Conference on Machine Learning (2015)

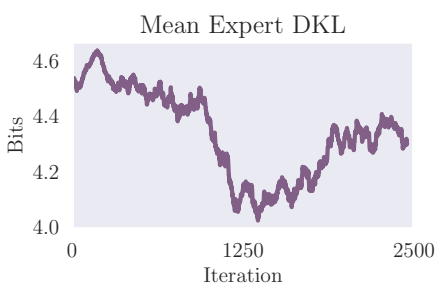
Gain Scheduling

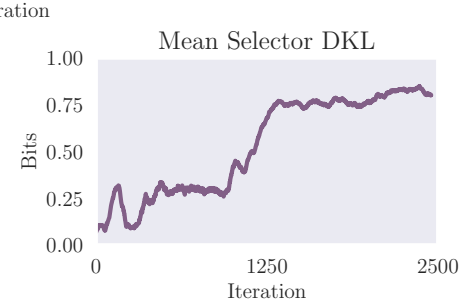
$$\dot{x} = A_i x + B_i u + \epsilon, \qquad \text{for } x \in X_i$$

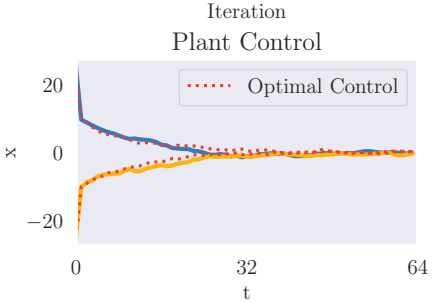
$$B_i = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0 \end{cases} \tag{14}$$











Conclusion

- Principled method applicable to a variety of tasks
- Resource limitation drives specialization
- No prior task information required: utility driven partitioning
- Normative framework to analyze hierarchical structures
- System build only by linear decision-makers
- Open Questions
  - High dimensional tasks
  - Sample efficiency in RL