DRA vs Max Ent RL

Max-entropy RL refresher

Quick refresher on max-entropy RL since we use the intuitions for the some of the results below.

$J_{\text{MaxEnt}}(\pi; p, r) = \mathbb{E}_{a_t \sim \pi(a_t|s_t),\, s_{t+1} \sim p(s_{t+1}|s_t, a_t)} \left[\sum_t r(s_t, a_t) + \alpha \mathcal{H}_\pi \left(a_t|s_t\right)\right]$

$q_{\text{soft}}(s, a) = r(s, a) + \gamma V_{\text{soft}}(s')$

$V_{\text{soft}}(s) = \alpha \log \mathbb{E}_\pi \left[\exp\left(\frac{q_{\text{soft}}(s,\cdot)}{\alpha}\right)\right]$

$\pi(a|s) = \frac{1}{Z} \exp\left(\frac{q_{\text{soft}}(s,a) - V_{\text{soft}}(s)}{\alpha}\right)$

> ⚠ **Caution**
>
> Ensure that the exploration problem of q-learning (positive bias) hasn't crept into the results. To do so, re-run the simulations with

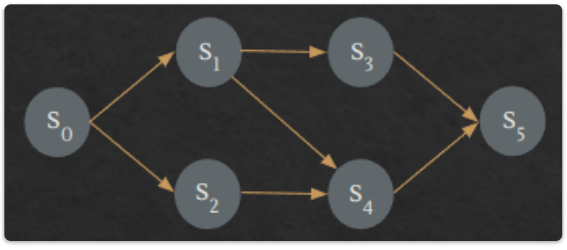
Simulation results

- [x] Ziebart's task
- [x] Mattar & Daw's maze
- [ ] Memory 2AFC task
- [ ] Bottleneck task
- [ ] Slot machines task
- [ ] Huys task

Ziebart's task

Ziebart's proposition for inverse RL was to penalize the entropy of trajectories sampled from the policy, since that would be intractable for large MDPs. We want to see if this differs from the approach by Haarnoja et al., who penalize the entropy of the policy $\pi(a_t|s_t)$ at time t.

Task schematic



We will refer to $s_0 \to s_1 \to \ \ldots \to s_5$ as the top-branch, and $s_0 \to s_2 \to s_4 \to s_5$ as the bottom branch. Notably, only the final state has a reward, i.e.

- $r(s_i) = 0 \quad \forall\, i \in \{1, \ldots, 4\}$

- $r(s_5) = 1$

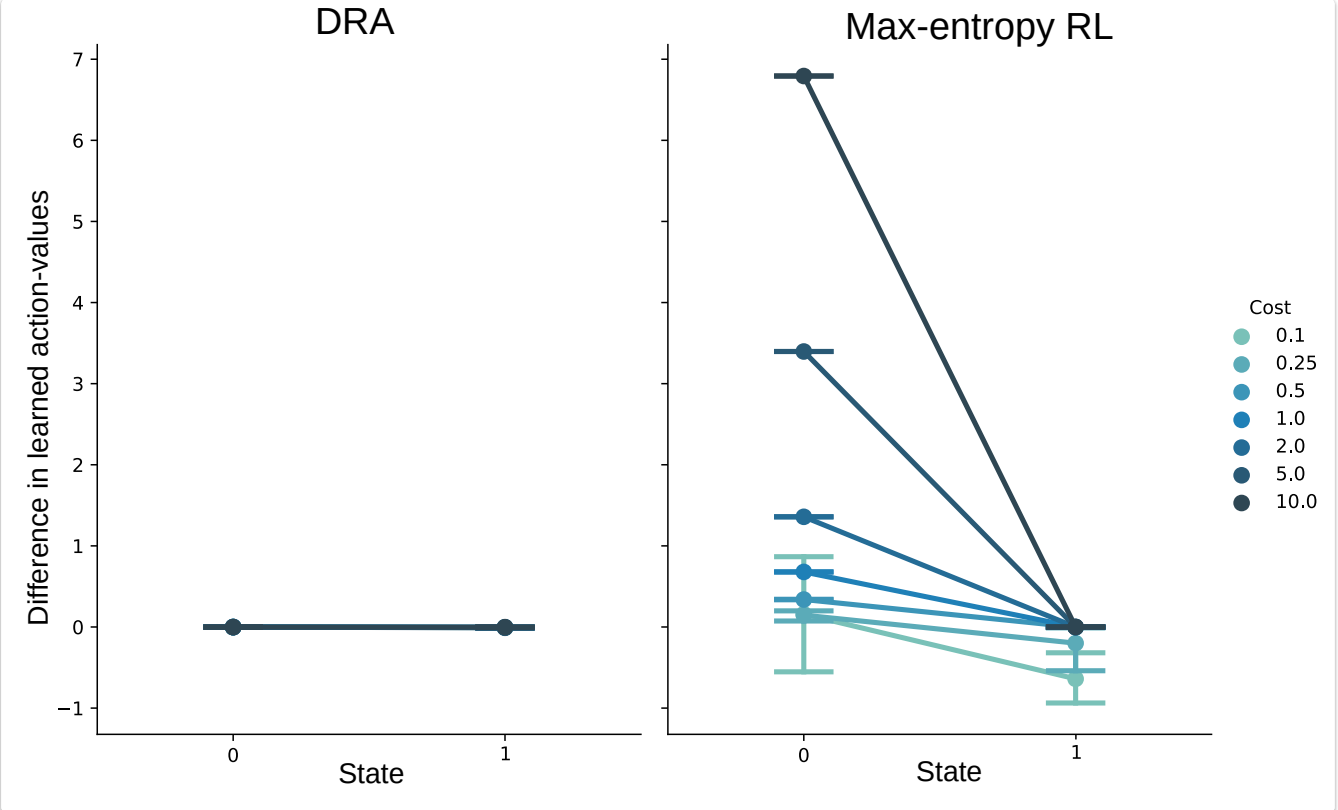
Results

Both models were simulated with $\gamma = 0.98$. For both models, we varied the noise parameters: $\lambda > 0$, and $\sigma_{\text{base}} > 0$ for DRA, and $\alpha > 0$ for max-entropy RL. Note that if a model does not show any preference between two choices, it's entropy should be $\mathcal{H}(s) = \ln(2) \approx 0.693$.

Model	$\mathcal{H}(s_0)$	$\mathcal{H}(s_1)$	Noise
DRA	0.693	0.693	$\forall\ (\lambda, \sigma_{\text{base}})$
maxEnt	0.64	0.69	$\alpha = 10$
maxEnt	0.64	0.58	$\alpha = 1$
maxEnt	0.5	0.24	$\alpha = 0.5$
maxEnt	0	0	$\alpha = 0.1$

As expected, DRA doesn't show a preference for either branch. This is because it encodes the true action-values which are equivalent for both branches. Max-entropy RL, on the other hand, shows a distinct preference for the top branch because it doesn't encode the true action-values, but instead "soft" action-values, which are different for the top and bottom branches. The intuition of why the soft action-values for the top branch is higher than that of the bottom branch is hard to get just from the equations I have listed at the top of this page, but if you work out the terms of the value, it is the normal RL value + a term that corresponds to the entropy of the agent's choices in that state.

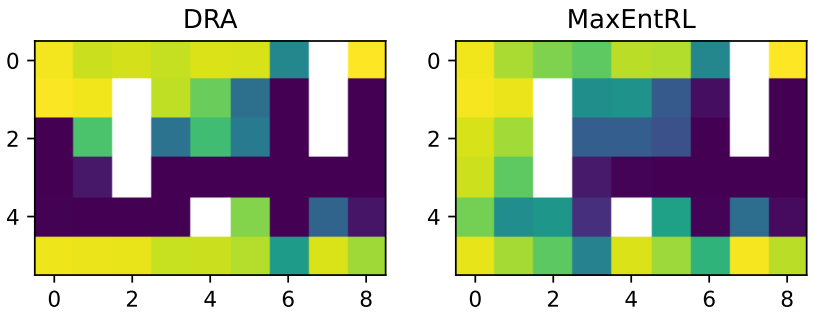
We can also see the preference in this figure showing the difference in the learned action-values for each of the two choice states (s_0, s_1) for both models:

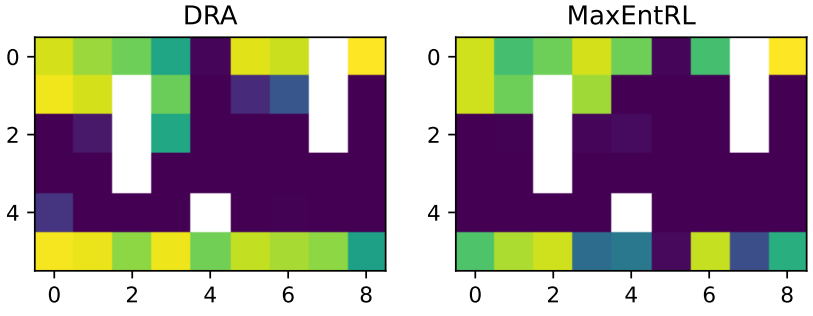


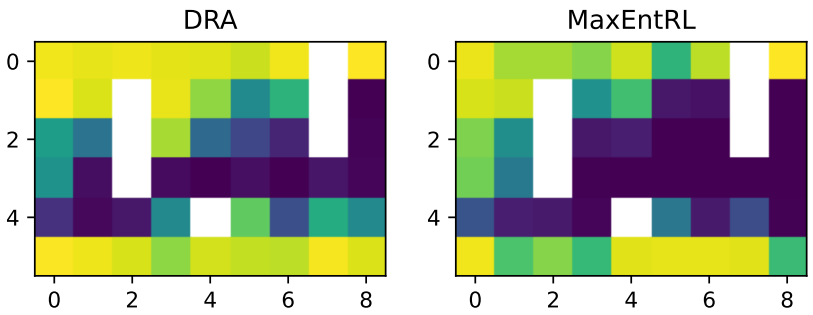
Mattar & Daw's maze

Not entirely sure how to quantify the differences/similarities, but here they are. For both models, varying the cost parameter gives a whole range of behavior. The color scale is missing but dark purple indicates low entropy or greedy behavior and yellow indicates high entropy or random behavior.

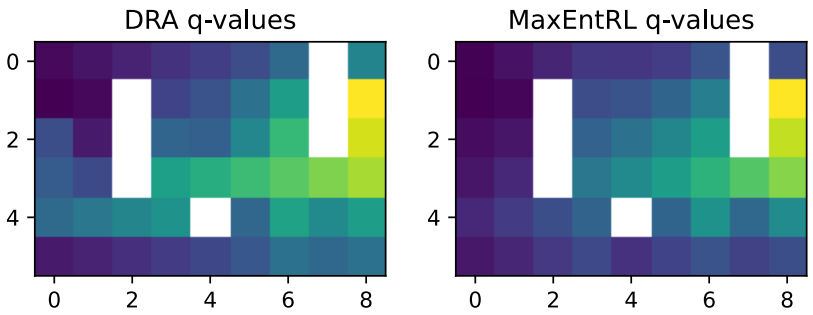
Entropy

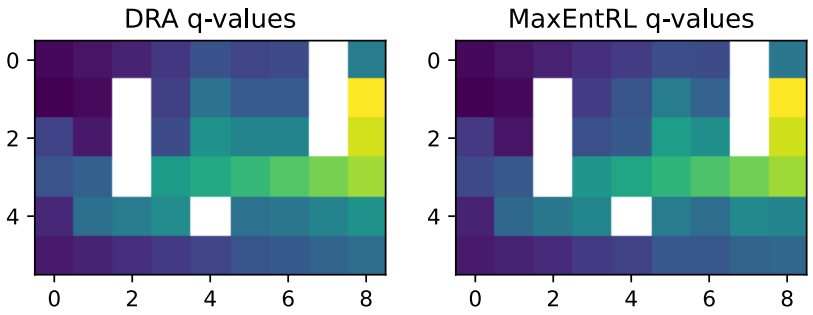


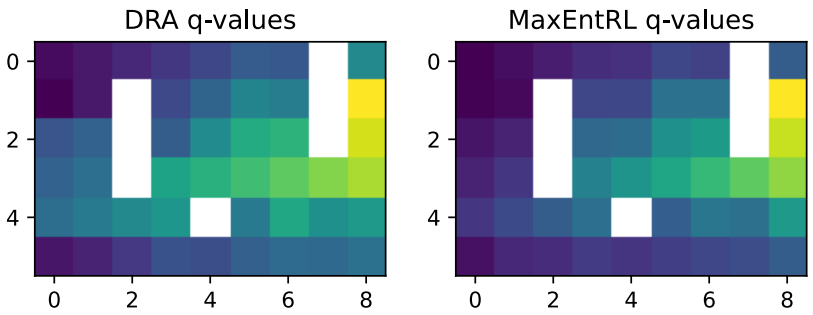




Action values


DRA q-values
MaxEntRL q-values
0
2
4
0
2
4
6
8


DRA q-values
MaxEntRL q-values
0
2
4
0
2
4
6
8



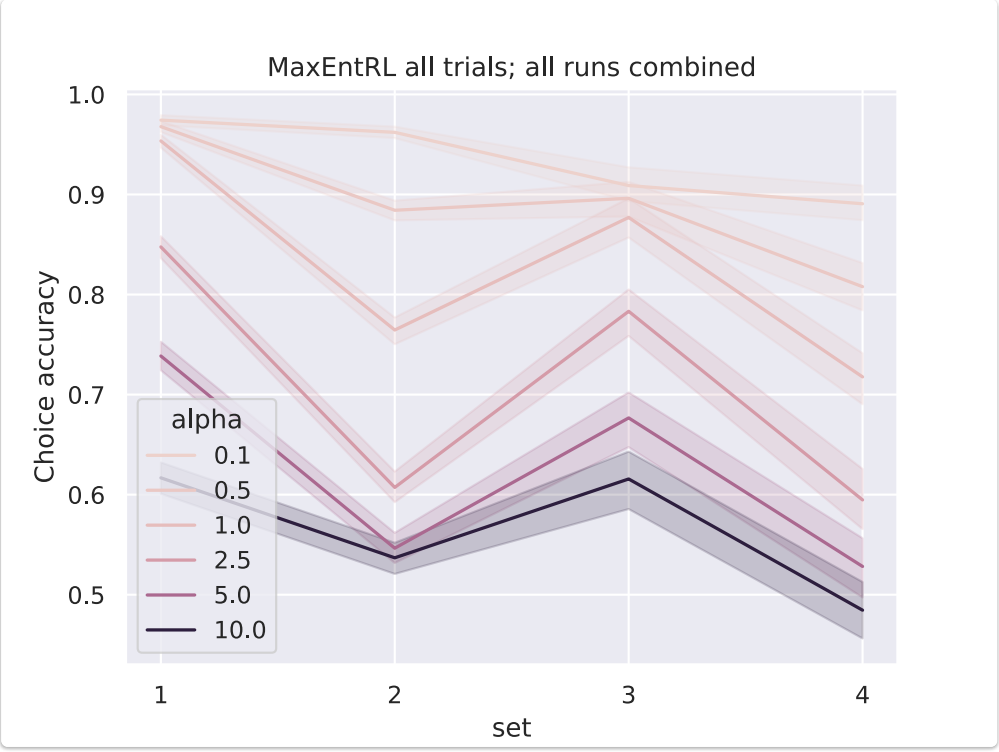
Memory 2AFC task

Task schematic

	High stakes	Low stakes
High frequency	set 1	set 2
Low frequency	set 3	set 4

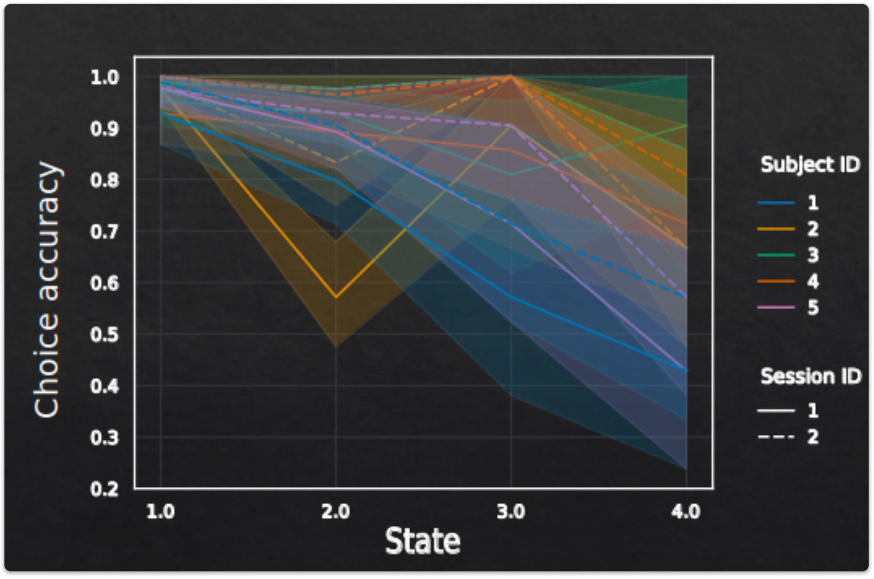
Each set has 3 options. Let's call them A, B, and C.

Max entropy RL shows a similar pattern of behavior as DRA. For some reason, I do not have the same plot for DRA, and I need to debug the code. But I know that varying $\lambda, \sigma_{\text{base}}$ can give all of these curves. DRA, however, has more independent control over the choice accuracy in the different sets owing to the extra parameter.

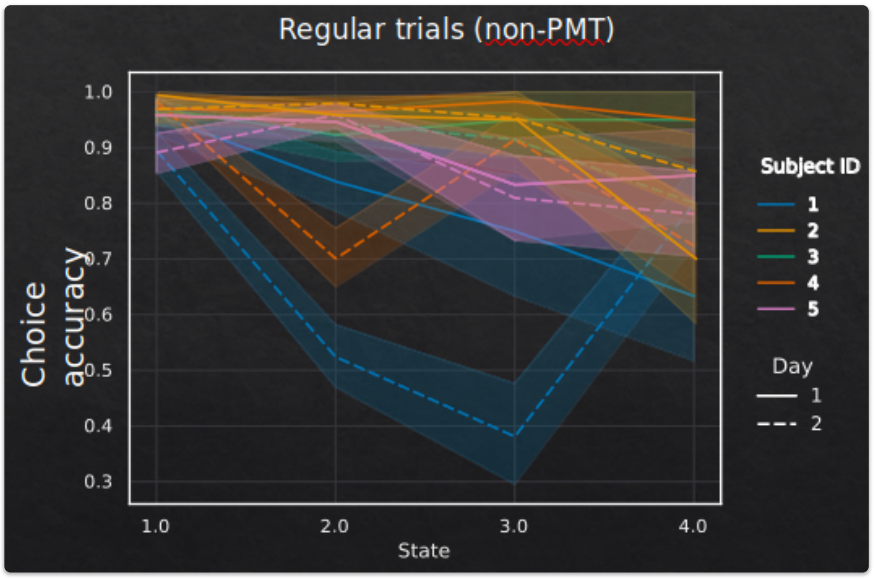


This also looks qualitatively similar to the pilot data we recorded, although the models are not really distinguishable from frequency/stakes if we were to just look at the regular trials.

Data from training sessions:

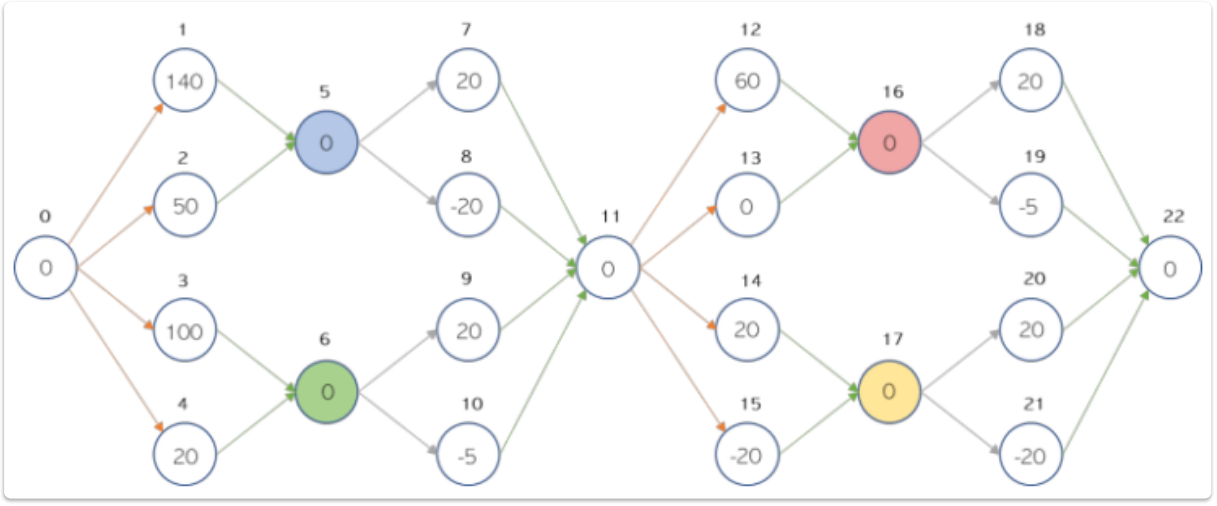


Data from test sessions:



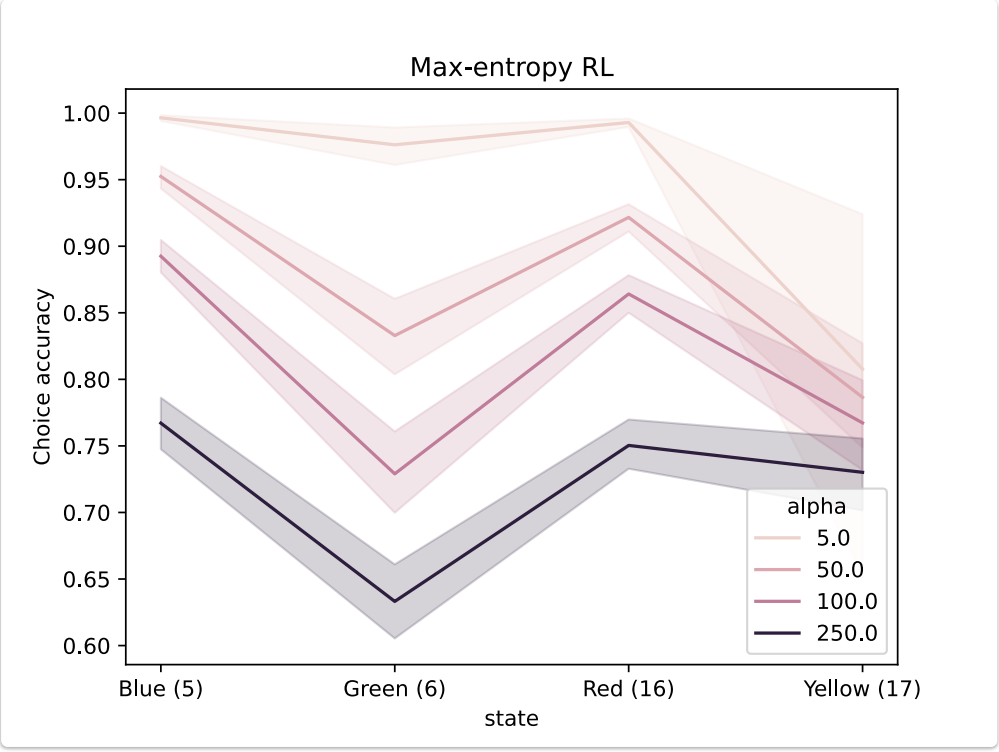
Bottleneck task

Task schematic



The orange transitions are stochastic. On any given trial, only two of the four choice options are available with their probabilities carefully chosen to adapt the experienced frequency of the states in color. If agents are behaving optimally, the top branches are visited 80% of the time and the bottom branches are visited 20% of the time.

Max-entropy RL



DRA

DRA has an extra parameter, so it can produce a wider range of curves than max-entropy RL. However, it doesn't stick to the rigid structure of the curve like max-entropy RL. It's more like the blue state (5) is the most accurate, and the rest could be anything. In this task, the green state (6) is *not* the least accurate, unlike the 2AFC task. Probably because it is less controlled? But in any case, the fact that it can fit pretty much anything makes me wonder if this is even a useful model to provide a normative explanation of human behavior.