A unifying theory of explore-exploit decisions

Robert C. Wilson and Jonathan D. Cohen

**Summary**

How do we solve the explore-exploit dilemma?

Two strategies [Wilson et al. JEP:G 2014]:
- Directed exploration – choose unknown options
- Random exploration – choose randomly

Here we show how directed and random exploration arise naturally from **deep exploration** [van Roy 2015].

**Task**

Choose between two slot machines to maximize reward

Rewards come from Gaussian distributions with **unknown means**

Before first choice, participants see **example plays**.

Precisely controls what they know

They then play out to end of time **horizon**

First **infer** distribution over mean of each option \( p(m_i | r_{1,1}) \)

Next, **sample** mean of each option: \( s_i \sim p(m_i | r_{1,1}) \)

Then **simulate** possible futures assuming means are \( s_i \)

e.g. simulation after choosing left first ...

**Expt 1: Uncertainty manipulation**

Horizon = 5; Uncertainty = [1 1], [1 2], [1 3], [2 2], [2 3], [3 3]

Total value of choosing option \( i \) is average over multiple simulations, \( Q_i = \langle q_i \rangle \)

**Expt 2: Horizon manipulation**

Horizon = 1, 2, 4, 8, 12; Uncertainty = [1 3]

**Theory**

Simulate future **rewards** based on simulated choice \( r \sim p(r | s_i) \)

Simulate future inference based on simulated rewards

Simulate future decisions assuming greedy choice

For each simulation, **value** of choosing option \( i \) on first choice is sum of simulated rewards, \( q_i = \sum_i r_i \)

Focus on **first choice**

Two parameters:
- Information bonus, \( \theta \)
- Decision noise, \( \sigma \)

Directed exploration scales with **difference in uncertainty**

Random exploration scales with **total uncertainty**

Directed exploration **asymptotes** with horizon

Random exploration does not asymptote with horizon