Reward Function Complexity and Goals in Exploration-Exploitation Tasks

Brian Montambault (brian.montambault@tufts.edu)
Department Of Computer Science, Tufts University
Christopher Lucas
School of Informatics, University of Edinburgh

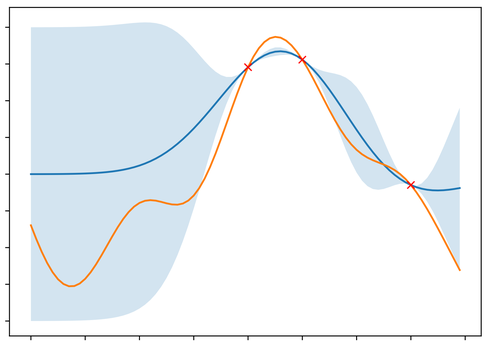
Contextual Multi-Armed Bandits

Choose one of *k* actions (Restaurants)

Each action is associated with a set of features (Menu items, location)

Features map to rewards via an unknown reward

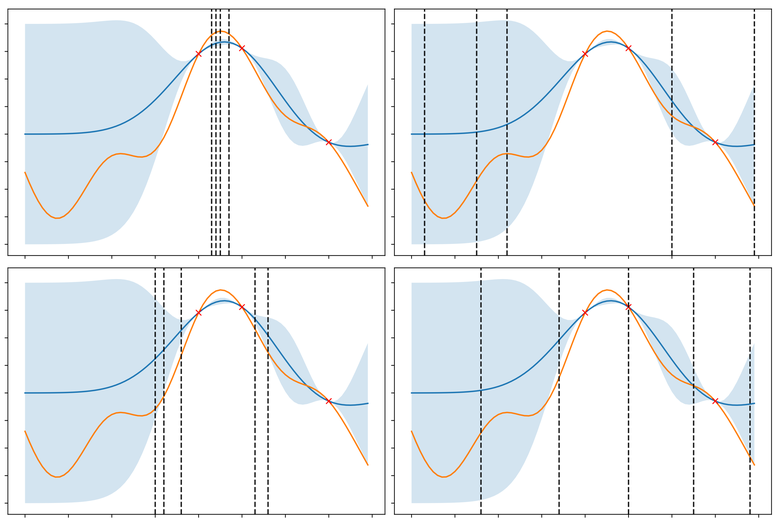
Explore-Exploit Dilemma: Should you choose an action with high expected reward or one that will give you more information about the reward function?



Let k actions be associated with one feature ranging from 1-k, with this feature determining the reward of an action via a reward function (orange).

After observing a set of actions and their associated rewards (red) the decision maker predicts the reward associated with each of the remaining actions and the uncertainty associated with each prediction (blue). How should can the decision maker use this information to choose their next action?

Exploration Strategies



Likely actions under four distinct strategies:

Mean Greedy (top left) - Choose actions with a high expected rewards

Variance Greedy (top right) - Choose actions with uncertain rewards

Entropy Search (bottom left) - Choose actions that give the most information about the maximum reward

Stochastic (bottom right) - Choose actions at random

What Determines How a Person Chooses a Strategy?

Goal?
- Maximize cumulative reward: balance exploration and exploitation on each trial (e.g. Upper confidence bound, expected improvement)
- Find the maximum reward: gain as much information about the location of the maximum as possible on each trial (Entropy search)

Reward Function Complexity?
- Simple reward function: Generalize to new actions
- Complex reward function: Explore randomly

Mixed Strategies: $u(k) = \underbrace{m_t(k)}_{\text{expected reward}} + \beta \underbrace{v_t(k)}_{\text{uncertainty}} + \lambda \underbrace{I(\{k, r\}; r^*)}_{\text{information gain about max}}$

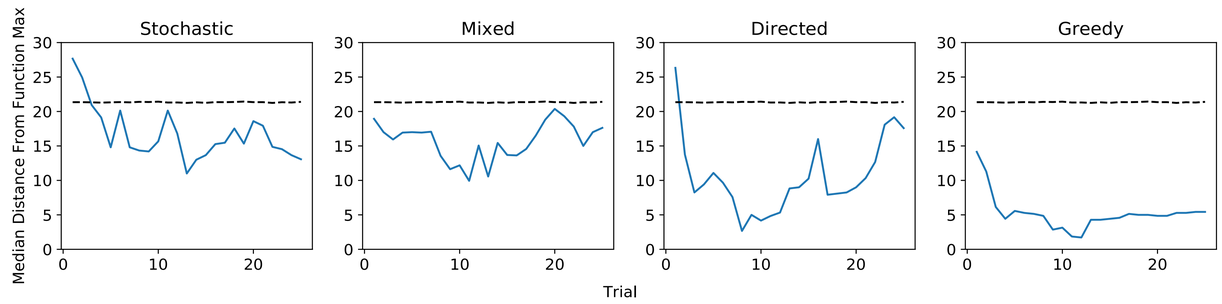
Infinite Possible Strategies: $p(a_t^i | g_i = z) = p(a_t^i | \beta_z, \lambda_z, \tau_z)$

Experiment
- 69 Participants
- 2 goal conditions (maximize reward, find maximum)
- 3 function conditions (linear, quadratic, sinusoidal
- *k* = 80 possible actions over 25 trials

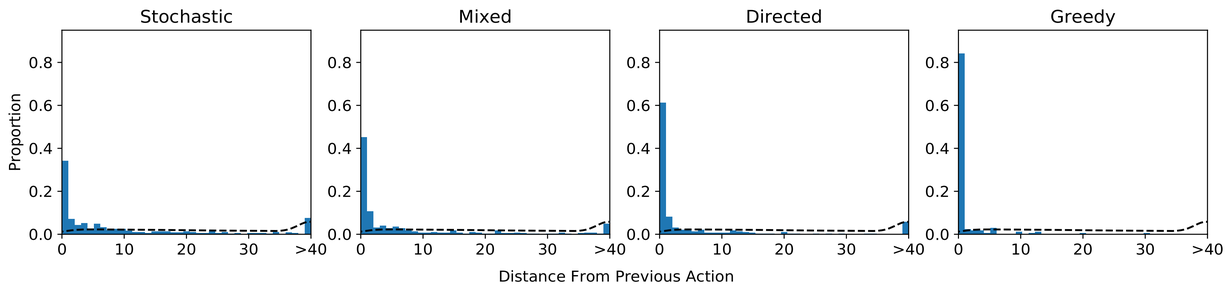
People use a mixture of strategies across goals and function complexities

	β	λ	τ	N Participants
Stochastic	0.29 ± 0.17	6.01 ± 8.87	5.23 ± 3.57	15
Mixed	1.58 ± 1.15	8.56 ± 8.32	1.77 ± 2.43	14
Directed	1.4 ± 1.29	11.19 ± 9.21	0.24 ± 0.22	11
Greedy	0.77 ± 0.44	0.61 ± 0.84	0.14 ± 0.19	8
	0.53 ± 0.27	4.06 ± 5.35	1.3 ± 1.14	7
	1.21 ± 0.55	6.38 ± 9.77	1.32 ± 0.61	6
	0.87 ± 0.33	8.77 ± 9.51	0.22 ± 0.24	4
	3.12 ± 2.33	0.99 ± 0.48	1.82 ± 1.14	2
	1.12 ± 0.39	6.98 ± 6.09	0.89 ± 0.63	2

Mean parameters for each strategy and the number of participants assigned to that strategy



Distance to the reward function maximum compared to random (dashed) for the top four strategies



Distance from previous actions compared to random (dashed) for the top four strategies