

# 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

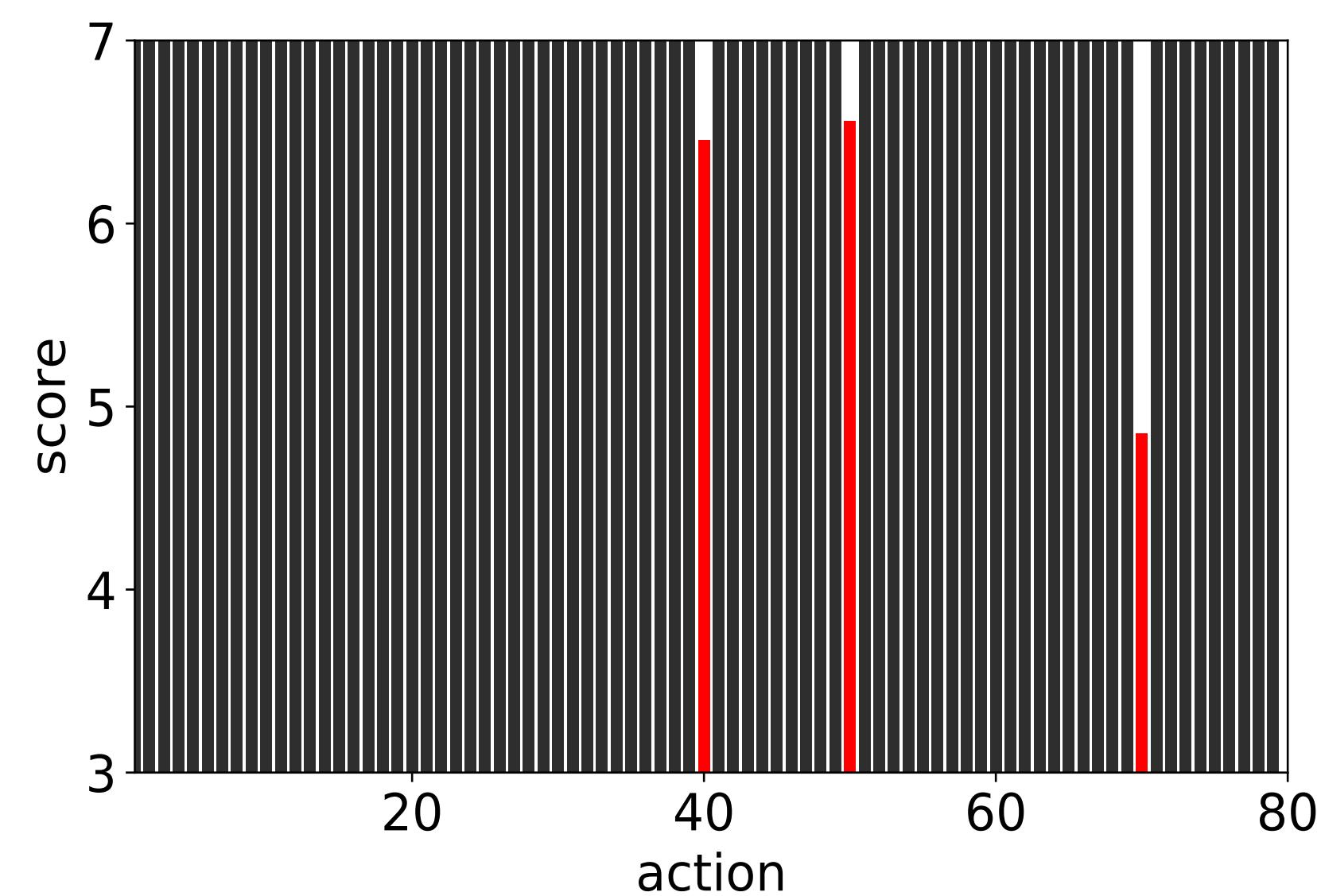
## Contextual Multi-Armed Bandits

Choose one of  $k$  actions (Restaurants)

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

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

**Task:** actions are associated with one feature ranging from 1-80, with this feature determining the reward of an action via an unknown reward function. Rewards are initially hidden (black) and revealed (red) after being selected.



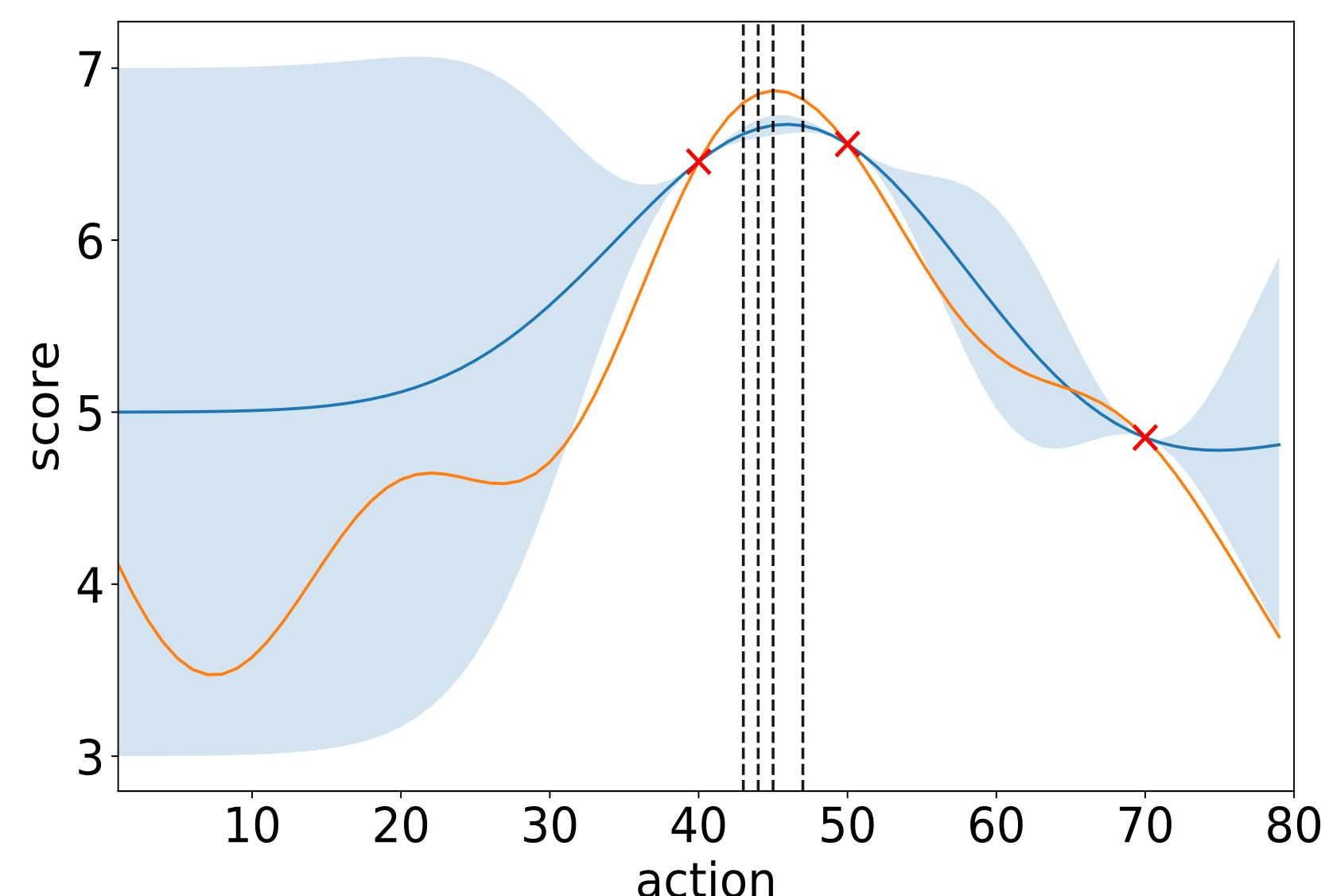
### Goal Conditions:

- Maximize cumulative reward over  $N$  trials
- Find the maximum reward within  $N$  trials

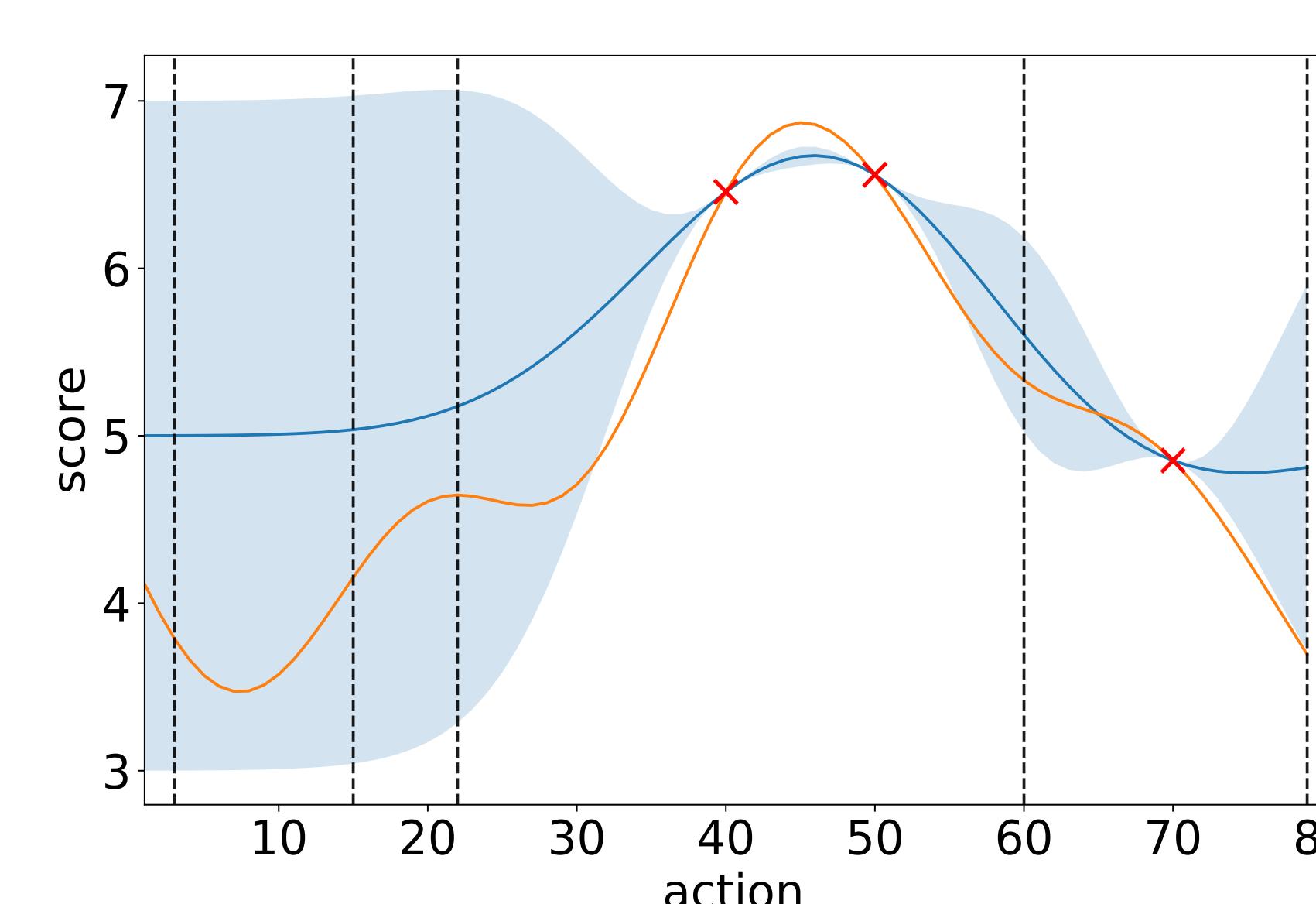
### Reward Function Conditions:

- Linear
- Quadratic
- Sincoidal (sinc function)

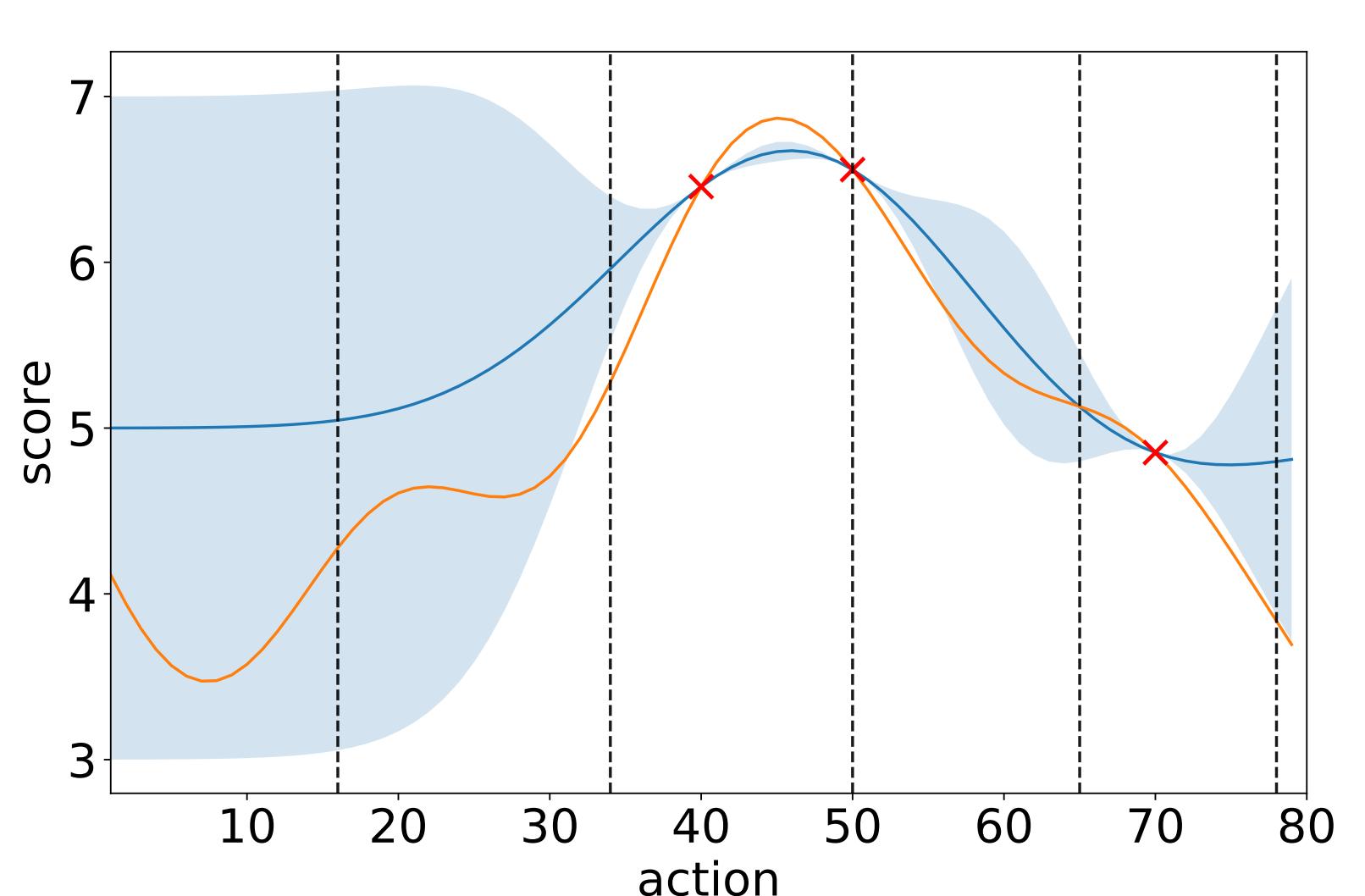
## Explore-Exploit Strategies



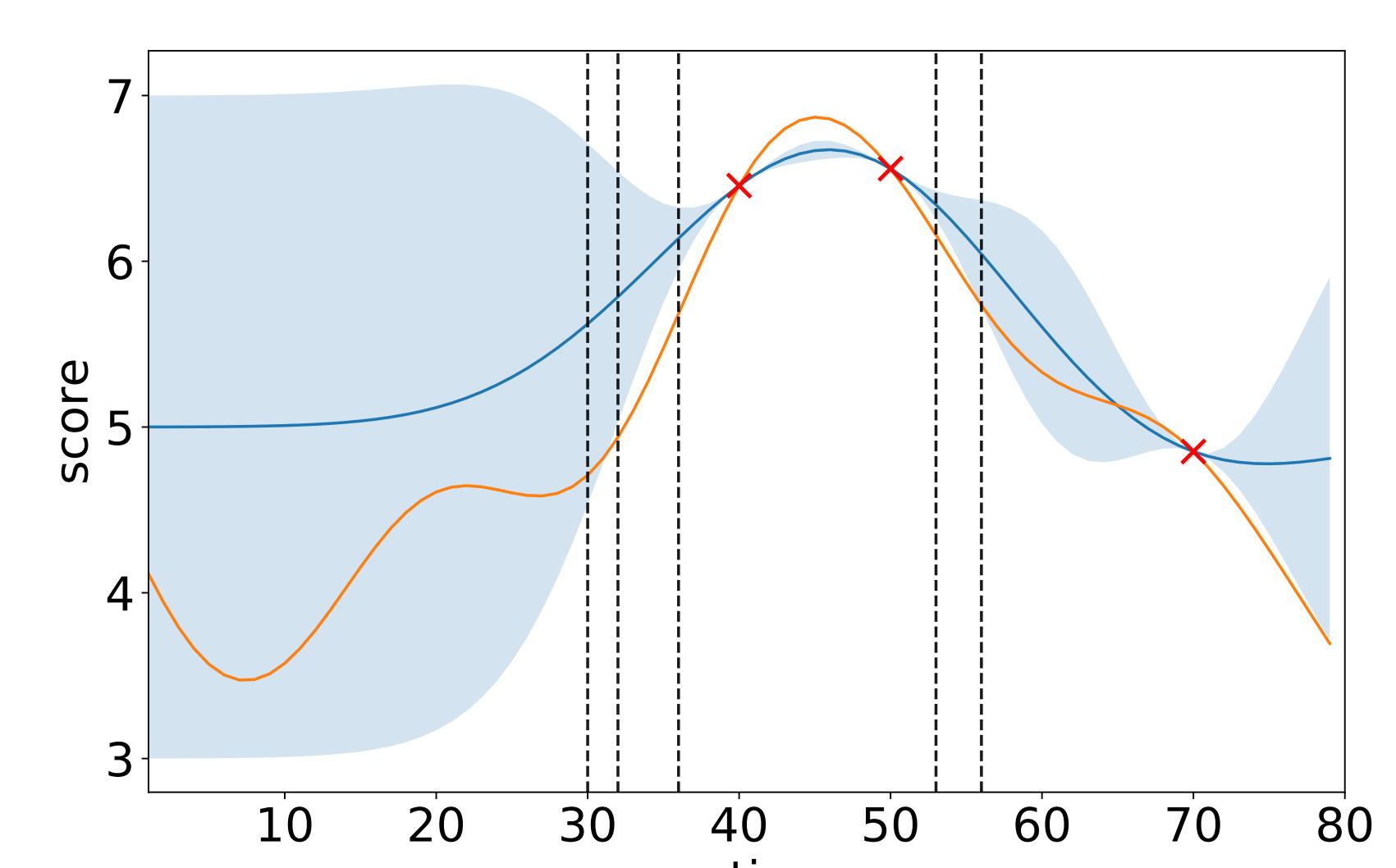
**Mean greedy:** Only choose actions with high expected reward



**Variance greedy:** Choose actions with highly uncertain reward



**Stochastic:** Choose actions at random, without using any information about the reward function



**Entropy Search:** Choose actions that will reduce uncertainty about some global property of the reward function (e.g. location of the maximum, value of the maximum)

## How do people choose 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 value of the maximum as possible on each trial (Max-value entropy search)

### Reward Function Complexity?

- Simple reward function: Generalize to new actions
- Complex reward function: Explore randomly

## Model

### Mixed Strategy:

$$p(a_t = k) \propto \exp\left[ \underbrace{m_t(k)}_{\text{expected value}} + \beta \underbrace{v_t(k)}_{\text{uncertainty}} + \lambda \underbrace{I(\{k, r\}; r^*)}_{\text{mutual information}} \right] / \tau$$

High  $\beta$ : Exploration directed by local uncertainty

High  $\lambda$ : Exploration directed by global uncertainty

High  $\tau$ : Undirected (random) exploration

### Infinite Mixture of Strategies:

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

## Conclusions

Most participants fell into one of four strategies:

**Stochastic:** Relied primarily on random exploration (High  $\tau$ , low  $\beta$  and  $\lambda$ )

**Mixed:** Used a mixture of random, local uncertainty directed, and global uncertainty directed exploration (High  $\beta$ ,  $\lambda$ , and  $\tau$ )

**Directed:** Used a mixture of local uncertainty directed and global uncertainty directed exploration, with less random exploration (High  $\beta$  and  $\lambda$ , low  $\tau$ )

**Greedy:** Didn't explore (low  $\beta$ ,  $\lambda$ , and  $\tau$ )

No strong evidence that strategies depend on goal or function complexity

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

Mean parameters for each group